Proceedings of the 8th International NLPCS Workshop

Special theme:
Human-Machine Interaction in Translation

Copenhagen Business School, 20-21 August, 2011

Edited by

Bernadette Sharp
Michael Zock
Michael Carl
Arnt Lykke Jakobsen

Copenhagen Studies in Language 41
Samfundslitteratur

Proceedings of the 8th International NLPCS Workshop
Special theme: Human-Machine Interaction in Translation

Copenhagen Studies in Language 41

ISSN 0905-09857
ISBN 978-87-593-1615-3

Cover by: SL grafik
Printed in Denmark by Eurographic Danmark A/S
Layout by the editors

Copenhagen Business School
Handelshøjskolen i København

Published by:
Samfundslitteratur
Rosenorns Allé 9
DK-1970 Frederiksberg

Phone: +45 38 15 38 80
Fax: +45 35 35 78 22
slforlagene@samfundslitteratur.dk
www.samfundslitteratur.dk

Editors:
Bernadette Sharp
Michael Zock
Michael Carl
Arnt Lykke Jakobsen

Preface

Welcome to the 8th International *Natural Language Processing and Cognitive Science* workshop, which is part of a series of workshops previously organised in Porto (2004), Miami (2005), Paphos (2006), Funchal (2007), Barcelona (2008), Milan (2009) and Funchal (2010). The aim of this workshop is to bring together researchers and practitioners in Natural Language Processing (NLP) working within the paradigm of Cognitive Science. The overall emphasis of the workshop is on the contribution of cognitive science to language processing in such areas as conceptualisation, representation, meaning construction, human and machine translation processes, ontology building, and text mining. This year, NLPCS was devoted to the special theme of "Human-Machine Interaction in Translation".

NLPCS 2011 received 29 papers from 17 countries. All contributions were reviewed in a double-blind procedure by three members of the programme committee. Based on these reviews, 18 contributions were selected for presentation as full papers, covering the following topics: cognitive processes and evaluation issues in machine translation, syntactic and semantic aspects of natural language processing, sentiment analysis, building of terminological ontologies for knowledge sharing. In addition to the papers four invited keynote lectures by professor R. Mahesh K. Sinha, professor V. N. Shukla, Dr. Michael Zock, and professor Pushpak Bhattacharyya were presented. The texts received are included in the volume.

The editors of these proceedings would like to thank the authors for their contributions and the members of the programme committee for their review work. Special thanks go to the local organisers, Dr. Merete Borch, Dr. Michael Carl, and professor Arnt Lykke Jakobsen.

We look forward to seeing you at NLPCS 2012.

August 2011

The editors

Co-chairs of the workshop:
Michael Carl, Copenhagen Business School, Denmark
Bernadette Sharp, Staffordshire University, U.K.
Michael Zock, CNRS, LIF, Marseille, France
Arnt Lykke Jakobsen, CBS, Copenhagen, Denmark

Workshop Programme Committee
Aretoulaki, M. (Dialog Connection, UK)
Barnden, J. (Birmingham University, UK)
Blanchon, H (IMAG, Grenoble, France)
Carl, M. (Copenhagen Business School, Denmark)
Casacuberta, F. (Universidad Politécnica de Valencia, Spain)
Cristea D. (University A.I. Cuza of Iasi, Romania)
Delmonte, R. (Ca' Foscari, Venezia, Italy)
Endres-Niggemeyer (Fachhochschule Hannover, Germany)
Ferret, O. (CEA, Fontenay-aux-Roses, France)
Fischer, I. (University of Konstanz, Germany)
Hardt, D. (Copenhagen Business School, Denmark)
Jakobsen, A.L. (Copenhagen Business School, Denmark)
Keller, F. (University of Edinburgh, UK)
Kutz, O. (University of Bremen, Germany)
Langlais, P. (University of Montreal, Canada)
Lapalme, G. (University of Montreal, Canada)
Lepage, Y. (Waseda University, Japan)
Macklovitch, E. (Bureau de la traduction, Canada)
Mladenic, D. (J. Stefan Institute, Slovenia)
Murray, W. E. (Boeing Research and Technology)
Neustein, A. (Journal of Speech Technology, USA)
Netter, K. (Consulting GmbH, Saarbrücken, Germany)
Rapp, R. (University of Leeds, UK)
Roche, C. (Université de Savoie, France)
Rosso, P. (Universidad Politécnica de Valencia, Spain)
Schwab, D. (LIG-GETALP, Grenoble, France)
Sedes, F. (Université de Toulouse, France)
Simard, M. (NRC, Gatineau, Québec, Canada)
Thompson, G. (Liverpool University, UK)
Tiedmann, J. (Uppsala University, Sweden)
Tufis, D. (RACAI, Bucharest)
Rayson, P. (Lancaster University, UK)
Sharp, B. (Staffordshire University, UK)
Wandmacher, T. (Systran, Paris, France)
Zock, M. (LIF- CNRS, France)

Table of contents

Invited papers

Man-Machine Integration in Translation Processes: an Indian Scenario

R. Mahesh K. Sinha
Indian Institute of Technology, Kanpur, India
sinharmk@gmail.com

Abstract. Translating natural language text or speech from one language to another is a challenging task. The quality of machine translation is found to be inferior to that of translation produced by a human being. However, machines are good at providing rough translations which can be used by human translators. Thus integrating man with the machine in the translation process is one of the ways of making translation systems practical in real life. There can be a learning loop in this man-machine integration process. This paper is focused on examining these aspects with specific reference to the Indian scenario. The Indian translation scenario is complex with a multiplicity of languages and scripts. As compared to the EU scenario, besides multiplicity of scripts, Indian languages exhibit free word group order; are morphologically rich; have complex usage of predicate verbs; use various distinctive features such as replicative words, onomatopoeic combinations etc. These result in a large number of variations in semantically equivalent utterances or written forms. The translation industry is still in its infancy in terms of preparedness for technology absorption. Machine translation strategies employed are primarily rule-based.

Keywords: machine and human translation, Indian scenario, bootstrapping

1 Introduction

The task of natural language translation from one language to another is attributed to human intelligence. Besides the knowledge of the two languages, it requires an 'understanding' of the source language text. The human translation (HT) process can be visualized as transformation of the mental picture created through understanding the source language into the target language structure that in the translator's judgment is a truthful reproduction of the mental picture to the target language audience. Translation accuracy, comprehensibility and fluency are some of the major concerns in this transformation.

On the other hand, a machine can hold and process a large amount of information that may be logically structured for speedy and relevant information retrieval. However, in case of rule-based MT systems, the limitations arise due to inadequacy of the language grammar and the rules that are usually handcrafted or mined with inad-

equate corpora. A natural language is what the native speakers use for their communication among themselves. To begin with, for a natural language, no grammar formalism exists and subsequently whatever grammar rules get formulated, do not provide complete coverage. In case of corpus based statistical approach, the limitation arises primarily due to the inadequate size of the corpus used in the learning process. The corpus used must be representative, exhaustive and adequate for such systems to succeed. These limitations of machine translation invariably lead to an imperfect translation in most of the situations. Nevertheless, machine translation (MT) provides a good starting point for the human translators giving a rough understanding, generates choices with alternative translations and creates its own resource of example translations in the specific domains and/or documents. Thus the roles of man and machine can be complementary in the translation process, and most of the translation industry worldwide is geared towards facilitating such complementary roles. This is the first level or surface level of man-machine integration (HMI or MMI) in the translation process.

In this paper, I first present the Indian translation scenario and then examine some of the basic issues in HT and MT in the section 3. In section 4, the relevance of HMI in the Indian context is examined and an integration framework is presented.

2 The Indian scenario

2.1 Linguistic scenario

i. A majority of Indian languages have a common origin belonging to the Indo-Aryan family (a sub-family of Indo-European used by 74.24%) or the Dravidian family (used by 23.86%). They are structurally similar to each other with respect to verb-ending (SOV) and relatively free word-order. The Indo-Aryan family consists of north Indian languages, while the Dravidian family consists of south Indian languages, the major languages being Kannada, Malayalam, Tamil and Telugu. All of these languages have undergone intense cross-fertilization over a period of time and have had varied degree of influence. Most of the Indian languages have either been spawned or greatly influenced by Sanskrit. They share approximately 60 % of the lexicons on average. Even though the Dravidian family of languages has evolved independently of Sanskrit, Malayalam and Telugu share about 80% of lexicons from Sanskrit. Hindi-Urdu, also called Hindustani, is greatly influenced by Persian. Other major families of languages are Austro-Asiatic (1.16%) and Tibeto-Burman (0.62%), prevalent in the north east part of India. They have had quite an independent evolution and exhibit distinctive features as compared to the Indo-Aryan and Dravidian families. There are several other lesser known families of languages with even fewer speakers.

ii. According to the 1961 census of India, there are 1652 languages (mother tongues) in India. Many of these languages exist only in oral form. Some of these are on the verge of extinction. There are 22 officially (constitutionally) recognized languages. In addition, there are about 32 languages with more than one million speakers and 122 other living languages, each of which is being used by more than a popula-

tion of 10,000 people. Many of these languages are aspiring to get officially recognized. As a consequence, an interlingual MT methodology is an obvious choice.

iii. The Indian Constitution provides that Hindi in Devanagari script shall be the Official Language of the Union. The Official Language Act also lays down that "both Hindi and English shall compulsorily be used for certain specified purposes such as Resolutions, General Orders, Rules, Notifications, Administrative and other Reports, Press Communiqués; Administrative and other Reports and Official Papers to be laid before a House or the Houses of Parliament; Contracts, Agreements, Licenses, Permits, Tender Notices and Forms of Tender, etc."

(http://india.gov.in/knowindia/official_language.php)

iv. There are ten major scripts in use in India. Roman script is very commonly used by the urban population for writing e-mail and SMS etc in Indian languages. There exists a significant population who know the native language and English but not the native script. Such people prefer to use Roman script for writing the native languages.

v. Text entry and editing in Indian scripts are generally considered cumbersome. Smart user interfaces are needed for man-machine integration.

vi. English is understood by less than 7-10% of the Indian population. However, it continues to be the primary link language of the country and a major resource for new knowledge. Thus English is the major language barrier in the country for knowledge creation and dissemination.

vii. English words, phrases and constructs are very frequently mixed in day to day communication. MT systems have to cater to such mixed language environments.

viii. There is inadequate standardization of terminology for various disciplines in Indian languages. The Indian Commission of Scientific and Technical Terminology (CSTT) has evolved about 6,00,000 terms for Hindi and identified about 25,000 Pan-Indian terms in different fields applicable to Hindi and other Indian languages. Sanskrit has been used as a base for this development. However, this data is found to be inadequate for many applications and domains. As a result, the translators start using non-standard terminology that may be confusing.

ix. Indian scripts are phonetic in nature in the sense that they are written the same way as spoken which is not the case with English and other western languages. Thus the transliteration of named-entities to and from English is error prone. However, transliteration among Indian languages is somewhat straightforward.

x. Indian English is influenced by native language forms and grammar. One very often encounters errors in usage of numbers, narrative forms, interrogative forms and missing articles.

xi. The language divide has significantly contributed to the digital divide. It has also contributed to widening of the social divide. One of the primary reasons for this has been a lack of contents in Indian languages on the web and the corresponding tools.

2.2 Challenges as compared to other scenarios.

Multiplicity of scripts and their phonetic nature: There are 10 major Indian scripts in use. Many a time a text has a mix with Roman script. There are situations where we

see that people are multilingual but know only their own script. Indian scripts are highly phonetic in nature in the sense that the words are written the way they are spoken. Each Indian script has some specific consonants and vowels that ensure purity of transcription. Thus there are variations in transliterations. This directly affects collection of noise free parallel corpora and also monolingual corpora.

Free word order: Indian languages have relatively free word order and a group of words can move to any position in the sentence. As a consequence there is a need for normalization to take care of the equivalences. On the text generation front, certain word orders are preferred to make the generated text more natural.

Rich morphology, Sandhi and Samaas: Indian languages are highly inflectional and rich in morphology. Sandhi and Samaas are concepts borrowed from Sanskrit. Sandhi represents co-joining of words and multiple words may join together to form a single word. Samaas refers to formation of noun phrases with noun-noun or adjective-noun combinations. These have an impact on lexical data base creation, morphological analysis and synthesis, parsing, text generation and corpus variations.

Replication: Replication of words is a peculiar phenomenon encountered in all Indian languages. Words of any part of speech may get replicated and the meaning of the replicated word is different from the individual constituent word. This impacts lexical data base creation, parsing and text generation. Although, the text generation can completely avoid the phenomenon of replication by constructing equivalent form, it loses the inherent naturalness in the target text.

Semi onomatopoeia: Here a word has an addendum of a similar sounding onomatopoeic word. The word group together has a different meaning. The impact is similar to the word replication phenomenon. It is also common practice to use non-sensical words that represent sounds associated with certain actions.

Predicate verbs and other verb forms: Predicate verb phenomenon in Indian language is very complex that leads to multi-word expression formation. These need to be detected at the time of analysis and appropriately composed at the text generation stage. In addition, most of the verb forms have morphologically derived verb forms that represent causativity, transitivity etc.

Localization, honorific and gender markings: There is a difference in the manner in which numbers, time, seasons are used in Indian and European languages. Many of the Indian languages, while paying respect to elders and dignitaries, use plural form of the verb. Similarly the gender of the person is reflected in the verb. These have direct impact on the text generation process.

Divergence influencing Indian English: Indian English speakers often get influenced by their native language and make mistakes in narrations (direct/indirect speech), number, person, determiners, use of articles and modalities. The irony is that many of these mistakes are accepted norms in the community. A translation system has to take these into account.

Mixed language forms: In India it is very common to mix English words in the text. Many a time there may be mixture of three languages. Most of the time the foreign words undergo morphological transformation as per the native language. This makes the situation very complex for analysis, lexical data base creation and translation. *Social, political and cultural issues:* Unlike the European Union, India is a

single sovereign country with a union of multiple states with multiple languages and cultures. Our constitution guarantees equality to all of its citizens in choice of profession and place of work. The complex linguistic scenario poses several social and political issues. The choice of second or third language learning is exercised based on employment opportunities and sometimes by circumstances. This leads to mobility of a certain section of people affecting the local social set up. On one hand, use of an alien language provides a unification platform bringing in cultural affinity, language has always been used as an instrument for political advantage and pursuing 'divide and rule' policy. The voluntary and forced migrations from one linguistic zone to another generate a translation requirement ranging from personal communication to schooling of the children.

As a consequence of the above, the methodologies applicable to European languages need modifications and adaptation. The statistical techniques as employed for European languages are not likely to yield satisfactory results as a corpus collected in general may not have an adequate coverage with the variations due to distinctive features of Indian languages.

There is also a brighter side of the picture. Indian Languages are closer among themselves. There is a rich tradition of unifying grammatical studies and logical deduction, Paninian framework and 'Navya Nyaya' are a few examples. Further, there has been intense cross fertilization for a much longer span of time as compared with EU languages. Thus the interlingua approach works much better for the translation among the languages of the sub-family. Further, a number of techniques and methodologies applicable to a language, are also equally applicable to each language of the sub-family. As a consequence, a methodology dealing with a foreign language to and from any individual member of the Indian language sub-family is applicable to all other members of the family. The need for translation from and to a foreign language (other than English) is primarily in tourism, import/export business and in addressing the national/international security issues. This is besides the general translation requirement of scientific, technical, patent and literary documents.

2.3 Translation Needs in India.

1. The proceedings of the parliament are translated into all the 22 languages by the interpreters and the source language can be any one of the Indian languages. Thus interpreters have to be employed for every pair of languages. Any member of the parliament can raise a question/answer in any regional language, which needs to be circulated to all members in translated form. All the reports of the parliament have to be prepared in English and Hindi.

2. All Government websites have to be provided in English and Hindi. A three language formula (English, Hindi and state language) is promoted by many of the states.

3. All communications within a state take place in the regional language. The communication with the Union is in English and/or Hindi. The medium of instruction for primary and secondary education is in the regional language. For the higher education and most of the professional courses, English is the medium of instruction. Thus there is a language mismatch when one moves to higher education. This is a typical scen-

ario for the rural and middle income group population in each state of the union. However, the urban higher income people send their children to English medium school right from an early stage.

4. The language used in judiciary at the lower court level is mostly in the regional language whereas the language used at the state High courts and the Supreme Court is English. For the fine interpretation of the laws, acts and statutes, English version is considered as authentic and not its translated version. As a consequence, in case of an appeal to the high Court or the Supreme Court, the judgment and proceedings of the lower courts have to be translated into English.

5. Needless to state that all financial and business sectors require their forms, brochures, manuals translated into the regional languages for promotion of their business. It is mandatory for the Government sector industries to provide annual reports in English and Hindi. The pharmaceutical and drug industry have to provide their drug information in all regional languages. Indian railways pass through multiple linguistic and script zones. All information pertaining to travel and transport have to be multilingual and reservation charts and other name lists have to be transliterated.

6. The process, manufacturing and maintenance industries are usually manned by technicians and laborers whose knowledge of English is very poor or know only the regional language. For example, all railway carriage and wagon maintenance staff members have to be trained and any communication gap may result in compromising the safety of the system. This is so for all plant operations and process flow control where human operators are in the loop. Usually the manufacturers of the equipment and machinery do not provide translation into regional languages. For all imported equipment translation of their manuals into regional languages are required. The defence sector has a special requirement of maintaining confidentiality. The aircrafts, warships and other combat machinery are imported from multiple countries, and the manuals have to be translated without delay before being deployed.

7. The India defence and paramilitary services have personnel coming from all linguistic zones of the country and they may be deployed in any other linguistic zone. Their recruitment, evaluation and training have to be provided in regional languages.

8. The police and criminal investigation sector presents a scenario wherein police stations work in the regional language using the regional script. However, the criminals move very quickly from one linguistic zone to another. Thus the criminal records have to be maintained or at least provided on demand in English or another regional language.

9. The competitive examinations held at the national level like those conducted by the Union Public Service Commission, IITs, IIMs and several others, the question papers have to be set in any one of the 22 languages as per choice of the candidate and the answers need to be evaluated in the regional language on demand. Here the translation process has to ensure both high level of accuracy and security.

10. With the high level of penetration of mobile and handheld devices in the country, multilingual communication for information retrieval, advertising, etc. both in textual and speech forms are needed. The rural literacy rate in the country being low, there is more a need for multilingual speech processing and translation. Currently,

only SMS can be received in regional languages with no translation. The speech to speech translation is still a distant reality.

11. The regional language radio and television broadcasts and vernacular newspapers are popular. However, the provision of choice of language is mostly limited to a few channels such as Discovery, Cartoons etc. The subtitling in the entertainment industry is mostly in English or in Hindi. There is a huge translation requirement here which has great commercial potential. The vernacular newspapers are a good source of corpus provided copyright issues are addressed. The national news reported in these newspapers can be used as a source for parallel text mining.

Thus it is seen that the translation requirements in India range from a simple rough translation to a formal translation that requires a high quality translation. The volume of the requirement is huge and requires a huge investment for clearing the backlog. Some of these are mandatory but still cannot be complied with due to practical constraints. The students undergoing technical, professional courses and higher education require translation of text books and other reference materials that are usually available mostly in English. The minimum that they require is a rough translation (in the absence of the translated text-books) that may lead to understand what is being taught. The question-answering through the web and filtering relevant information fall in a similar category. Tourists also have this kind of a requirement. On the other hand all official communications, bulletins, gazettes, manuals, legal documents etc. have to be translated with a high degree of accuracy, The socially relevant topics such as environment, health, agriculture, vocational training, marketing all have varying translation quality requirements.

2.4 MT R & D and translation industry scenario

The work on machine translation in the country started in the early eighties [1,4]. However, even today, the MT researchers and system developers in India are confronted with the problem of coping with limited linguistic resources in terms of corpora. This has led to development of MT systems that are primarily rule-based with some hybridization of examples and limited statistical information. Currently, these systems are being primarily used in the government sector in routine official correspondence, health awareness and limited parliament documents.

The major centres where machine translation research & development is being carried out are IIT Kanpur, IIT Mumbai, IIIT Hyderabad and CDAC Pune. AnglaBharati methodology [6] for translation from English to all Indian languages based on pseudo-interlingua & RBMT and AnuBharati methodology [6] using hybrid EBMT for translation among and from Indian languages have been developed at IIT Kanpur. AnglaBharati technology is being used by some CDAC centres for implementation for different Indian languages. At IIT Mumbai a universal networking language (UNL) based interlingua MT system for English to Hindi [2] has been developed. The group also created a *factored* English-Hindi SMT system with reordering of English sentences [3]. The group at IIIT Hyderabad developed a system called Anusarak [1] which the authors called ‘Machine translation in stages’ for translation among Indian languages. The group also developed a English-Hindi MT system called ‘Shakti’ us-

ing a transfer approach. It had also led a consortium mode project on development of IL-IL multi-part RBMT system named Sampark (http://sampark.org.in). The CDAC Pune group developed a TAG (tree adjunct grammar) rule based English-Hindi MT system called MANTRA (www.cdac.in/html/aai/mantra.asp). The group led a consortium mode project on English-IL multi-engine (RBMT, EBMT,SMT) paradigm based MT system whose performance is dominated only by RBMT system.

As far as research & development on speech-to-speech translation is concerned, the major bottleneck is automatic speech recognition (ASR). Real-time translation with almost instantaneous response is another bottleneck. Indian ASR technology is still very "fragile", and it is sensitive to noise, mismatch in train and test conditions, speaker variability, accents, dialects, etc. There is even a big difference in performance between "read speech" and "spontaneous speech". There is wide geographical variation in speech even in Indian English and Hindi. For example, Hindi as spoken in Delhi is not the same as in Hyderabad. Besides this, in accent there is an influence of mother tongue speech. We do not have good transcribed data for Indian languages with speech collected in natural mode with accent/dialectical variations. Most of the systems developed are lab systems rather than practical systems. ASR would give decent output for translation, if we constrain the task and actually collect speech data for that task and build systems that exploit domain context. The speech-to-speech translation in such limited domain can start only when ASR attains a certain level of maturity.

As we see from the above description, our major hurdle in MT R & D is non availability of appropriate linguistic resources and tools. There is an initiative to have an Indian LDC, however this is a gigantic task involving both manpower and money. High performance tools for POS tagging, morphological analysis, parsing, named entity recognition, verb frames, wordnet are needed. Lack of standardization inhibits sharing of resources among different research groups. Further, there is an acute shortage of trained manpower to work on language technology projects. The universities and colleges are not having appropriate curricula to support this. There are less than 5% of CS and IT graduates who undertake thesis work in this area in the country. Besides this, the IT professionals find other jobs more lucrative. Although there exists very sound traditional linguistic knowledge in terms of grammar formalism and logic, there is a wide gap in exploiting them in computational framework. The computer science and technology professionals are not exposed to this valuable heritage of the country and the scholars possessing this knowledge are not familiar with computational formalism needs. There is very limited funding available for research and most of the funding available (primarily from the Government sector) has the pressure of delivering time bound product oriented technology. There is hardly any participation from private industry.

The Indian translation industry is still in its infancy and is largely dependent on the part-time free-lance translators. While some of them use CAT tools, their usage is limited due to their cost effectiveness and applicability to Indian languages. The industry is also confronted with lack of standardization in usage of terminology, difficulty in dealing with non-Roman scripts and inadequate training. None of them are actually using a machine translation system. This is primarily because the MT system

options available to them have limited performance, and they do not find them attractive or cost effective. Many of them still use paper and pencil for preparing initial draft and for editing. There is also a lack of standardization in the work flow.

3 Issues in HT and MT and integration processes

Whether a text can be translated by humans or machines, or in a HMI process, is driven by several factors, some of which are as under:

Critical vs. casual contents: MT is acceptable for casual translation for personal use as it is fast and involves no intermediary. HT or HMI is invariably used for critical documents such as books, reports, legal, formal communications, etc.

Volume vs. quality: There cannot be quality assurance with MT. In case of a large volume quality translation, a large backlog is a usually observed phenomenon as it involves HT and sometimes with multiple iterations. MT can handle large volumes but may not be acceptable for many applications.

Secure vs. unsecure MT: If one uses an on-line or a server based MT such as Google translate, the user's data is unsecure as it becomes available to the developers for further development and may be available to other competitors. The situation is similar when using crowd sourcing or social networking for obtaining or evaluating translations. The translation industries do not share translation memory or other data to maintain non-disclosure to their clients.

Data driven vs. work-flow driven: MT is primarily data driven, whereas HMI is highly work-flow driven. Due to lack of standardization in HMI work-flow, pipelining or sharing of the task becomes difficult.

Cost effectiveness and User friendliness: The investments in MT systems have to be cost-effective and should reach the break-even point within a short span of time. The degree of ease of operation, training requirements and the translation workbench environment dictate the acceptability of an MT system. Their integration in the work-flow and awareness are important.

HT and MT integration processes: Use of controlled language through pre-editing, automating human pre-editing/post-editing through machine learning using the data at different stages of the translation process, creating machine-aids based on understanding of the human translation process where he/she has to pay more attention are some of the ways in which the HT-MT integration can take place. Each of these when automated provide improvisation to MT engine performance and reduce HT efforts. The entire process can be bootstrapped.

4 HMI in an Indian Scenario

An increase in the translation throughput and improvisation through learning is one of the major outcomes of the HMI process. Another major outcome is production of clean aligned corpora that find application in deriving translation memories and phrase level ready translation pairs. This data is very valuable for effective implementation of SMT and EBMT paradigms. These are general outcomes of the HMI cycle, which is essentially universally applicable to all the language pairs. It is the

available resources (both language & technology) that make the difference in its effectiveness. In fact, given the Indian scenario HMI appears to be the only way to meet the translation demands in India. The following paragraphs summarize, why and how HMI is more suited specifically to the Indian context:

i. Since the concept of spelling in Indian language is somewhat loose, there is a large number of variations in writing the same word and in transliteration of named entities by different translators. For any word alignment task, this becomes sparser. Through HMI preferences and equivalences could be recorded and the machine can be used to normalize this.

ii. Entry of texts in Indian scripts is comparatively a more complex task and sometimes error prone. In HMI, the copy and paste mechanism reduces this effort.

iii. The Indian language texts may not be stored in Unicode / UTF-8 coding form. HMI environment offers a way by which code converters can provide normalization.

iv. Many a time the same text needs to be translated into multiple languages. Since Indian languages are much closer to each other both in structure and lexicons, the text translated into say Hindi may be used to translate into another Indian language by a translator knowing Hindi. HMI offers a way to use such cross lingual information in post-editing.

iv. There is inadequate standardization in usage of terminology. Translators tend to use transliterated terminology or use non-standard substitutes. Through HMI, terminology could be substituted appropriately based on preference and usage. Transliteration variations due to the manner they get pronounced in different regions is also tackled using HMI.

v. In Indian language texts, it is very common to mix English and other language which may appear in un-transliterated form. This usually happens when the meaning of the source language word is not known to the translator. This is a very common phenomenon observed on Google Translate. Interestingly, very often the unknown words get morphologically transformed. As an example, ‘blessed’ in Hindi may be translated as ‘bless kiyaa’. Through HMI, meanings get gathered and lexical database augmented.

vi. There is a shortage of competent manpower in the translation industry. The human pre-editing, MT and post-editing outputs in the HMI cycle can be used both for training and for machine learning. Multi-tier post-editing performed by novice to expert translators provides a mechanism for acquiring skills.

vii. The common errors in Indian English and missing articles handled through HMI cycle provide valuable information for machine learning.

viii. Many a time people know the language but not the script. In all such cases, they tend to use Romanized script. HMI offers a way in which it can be converted to script of choice before pre-editing.

Keeping the Indian scenario in view, a HMI translation process framework [5] is proposed and is shown in Figure 1. The translation system is assumed to be equipped with a centralized MT server or a cloud computing environment. It has a distributed nation-wide network of professional translators and apprentice-translators. Crowdsourcing over this network will provide a broad platform even for training, certification, standardization as well as employment opportunity to educated unemployed. The

different components/modules of the system with generated databases, human interfaces, machine aids and machine learning have been marked.

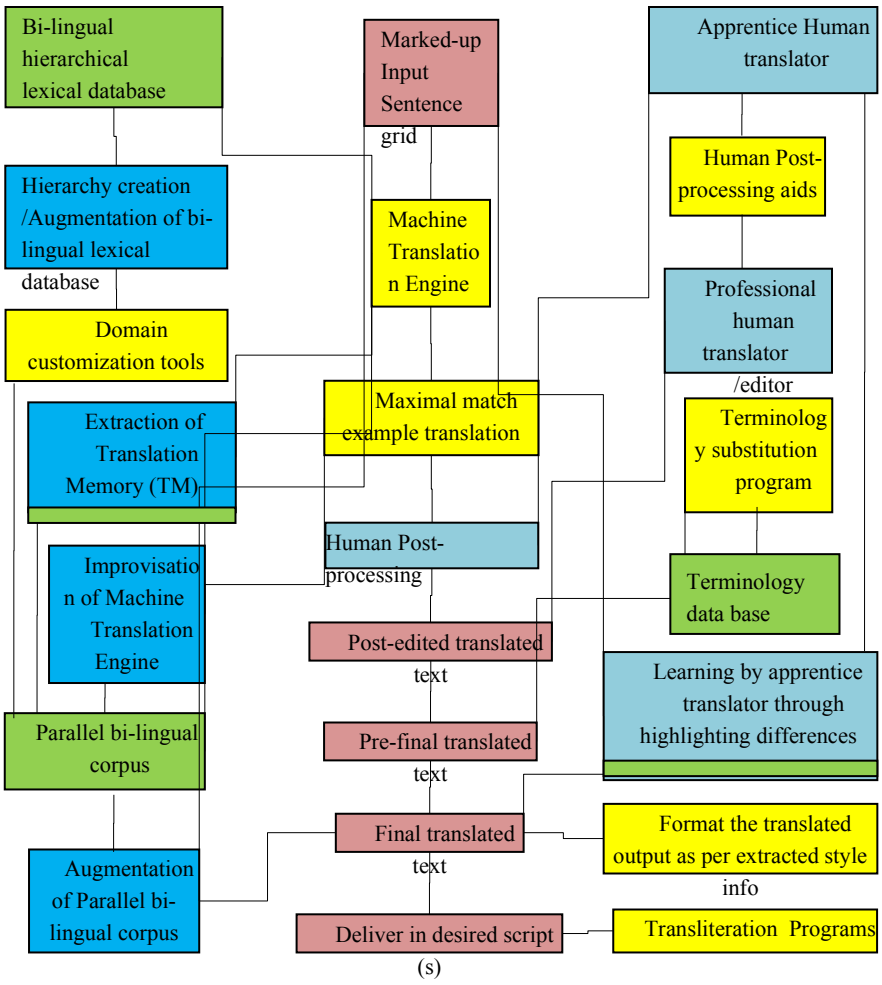


Fig. 1. The overall translation process with man-machine integration

There is a tendency to use the English terminology words as they are, in transliterated form. The mappings of terminology including their transliteration have a lot of regional influence. Keeping these in view, the methodology allows liberty at the regional level. Ultimately, as and when the standardization is formalized, these can get substituted.

The entire process generates a domain specific bilingual corpus. The corpora generated through this process are much more relevant and noise free. This is used for enriching domain-specific lexical database and extracting translation memory. It is also used in building a translation model for use in SMT and for EBMT. The target corpus

generated is used in language modeling in the specific domain. This language model is useful in building tools for automated post-editing. This in turn improves the MT system performance and results in reduction of HT effort in the next HMI cycle.

The entire translation process in this framework is pipelined with participation of varying skills at different stages. The skill needed at the stage of extracting text zones is only that of scanning and image handling with no linguistic expertise needed. The tasks of marking up the extracted text zones, isolating sentences/simplification/pre-editing and identifying text components primarily need the knowledge of the source language. A pool of the marked-up text can be generated and made available for experimentation to different machine translation paradigm. The apprentice translator has to be a bilingual person but need not be a translation expert to begin with. At the post-editing stage, the editor need not be a translator.

5 Conclusions

The machine translation research & development in India is faced with many challenges, and the Indian translation industry is in its infancy. Man-machine integration in the translation process is very much required both to meet the translation demand and to provide impetus to MT research in the country. This is the only way to take the MT systems from lab to users. Keeping the Indian constraints and state of art in view, a multilevel framework for man-machine integration is presented with a bootstrapping mechanism. In order to make this successful, cooperation from different sectors, private as well as Government, and their partnership in revenue modeling will be crucial. The experiences of the European Union community with a similar multilingual scenario are very much relevant in this context. Similarly Indian experiences in dealing with the constraints and exploiting homogeneity within a language family are relevant to the European Union researchers and developers.

References

1. Bharati A., Sangal R., Sharma D. and Kulkarni A.: Machine translation activities in India: A survey, Published in the Proceedings of workshop on survey on Research and Development of Machine Translation in Asian Countries, Thailand, May 13-14, (2002).
2. Dave S., Parikh J. and Bhattacharyya P.: Interlingua Based English Hindi Machine Translation and Language Divergence, Journal of Machine Translation (17), September (2002).
3. Ramanathan A., Choudhary H., Ghosh A. and Bhattacharyya A.: Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT, ACL-IJCNLP 2009, Singapore, August, (2009).
4. Sinha R.M.K.: A Journey from Indian Scripts Processing to Indian Language Processing. IEEE Annals of the History of Computing, Jan-March: 8-31,(2009).
5. Sinha R.M.K.: Indian National Translation Mission: Need for Integrating Human- Machine Translation, (http://www.mt-archive.info/MTS-2009-Sinha-1.pdf), Proceedings MT Summit XII, Aug.26-30, (2009), Ottawa, Canada. `
6. Sinha R.M.K.: An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures, Invited Paper, Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004), November 17-19, (2004), Tata Mc Graw Hill, New Delhi.

Divergence patterns for Urdu to English and English to Urdu Translation.

Shukla, V. (1) and R.M.K. Sinha (2)

(1) CDAC, Noida, India
(2) IIT, Kanpur, India

1 Introduction

Translation divergence is said to occur when two languages differ in their grammars. Thus divergence would occur when a sentence in the source language translates to a sentence in the target language in a very different manner. Divergence for English to Hindi and Hindi to English has been studied to quite an extent. The existing literatures are: [Dave et al, 2001]; [Gupta et al, 2001, 2003] and [Sinha et al, 2005a, 2005b].

Divergence for English to Urdu or Vice versa has not been explored so far barring an exception [Saboor et al, 2010] which has studied the divergence patterns for Urdu to English and not vice versa. Thus here we have tried to identify the various divergence patterns that exist for English to Urdu as well as Urdu to English.

The study is based on the findings of [[Sinha et al, 2005a, 2005b] for Hindi. Hindi and Urdu are structurally very similar. They use similar postpositions, verb morphology as well as complex predicate verb structure [9]. The broad categories of divergence are still based on Dorr classification [Dorr, 1994].

The types of divergences that have been considered are Promotional /Demotional, Structural, Lexical, Categorial, Conflational/ Inflational and Thematic Divergence. There can be many other types of divergences other than those mentioned above and these could occur due to use of Reduplication, Honorific usage of words and omission of a subject in a language.

Divergence Patterns:

A. Conflational and Inflational Divergence

A conflational divergence results when two or more words in one language are translated by one word in another language. Let us illustrate this with a Urdu sentence.

(U)- voh khaana nosh farmaa rahe hein.

The English translation for this Urdu sentence would be:

(E)- They are eating.

In this example, the verb "nosh farmaa rahe" of Urdu sentence is equivalent to one word verb of English (i.e. eat) upon translation. Another example that can be considered is:

(U)- Baraaye Maherbaani aap tashreef le jaayen.
(E)- Please go.

Here the word "please" is being referred to as "baraaye maharbaani" in urdu and the verb "go" is referred as "tashreef le jaayen" and also in the English sentence there is no mapping for the word that has been translated to "aap".

The examples discussed above show Conflational divergence. The opposite case would be when one word in the source language is realized by two or more words in the target language and this is referred to as inflational divergence. We can illustrate this with few examples:

(U)-loo chal rahii hai.
(E)-Hot wind is blowing.

In this example, the word *loo* of Urdu sentence requires two words of English (i.e. hot wind) upon translation. If we consider the case for English to Urdu translation, we find that inflational divergence is very common.

(E)- One moment please!
(U)- Baraaye Maherbaani Kuch Der Intazaar Kijiye

In this example "one moment" has been translated to "Kuch Deyr Intizar Kijiye" and please as "baraaye maherbaani". Another example that would illustrate the inflational divergence is:

(E)- Welcome!
(U)- khush aamdeed

B. Structural Divergence

Structural divergence occurs where an NP argument in one language is realized by a PP adjunct/oblique NP in another language. For example,

(U)-Mere valid lucknow ja rahe hein
(E)-My father is going to Lucknow.

In this example, *lucknow*, the Noun Phrase in the Urdu sentence is converted into PP "*to Lucknow"* in English sentence upon translation. If we consider the case of English to Urdu translation then the following English sentence "He entered the room" which is translated to *" vah kamare mein daakhil huaa", t*he verb '*enter*' in English sentence takes an NP argument '*the room*' whereas its Urdu counterpart *"daakhil honaa"* takes a PP adjunct *kamare mein*.

C. Promotional and Demotional Divergence

As mentioned by [Sinha et al, 2005a] Promotional and demotional divergences or Head-swapping divergences arise where the status (lower or higher) of a syntactic constituent in one language is affected in another language. For instance, when an adverbial element in one language is realized by a verbal element, it constitutes a case of promotional divergence and an opposite case will result in demotional divergence. In demotional divergence the role of the main verb of the source language sentence is demoted upon translation. Some examples of demotional divergence in Urdu-to-English translation are discussed below:

In the examples discussed below, the word realized as a main verb in source language (Urdu) is realized as an Adverbial Modifier or as an adjective in the target language (English). For example,

(U)- sangeet baj rahaa hai
(E)-The music is on.

Here the main verb in Urdu sentences "*baj raha "* is realized as an Adverbial Modifier "*on"* in English sentences upon translation. If we

consider an English sentence “Life goes on” then it is translated to urdu as – “*Zindagi chalti rahti hai”*.

Here “on” is a particle in English sentence and this is realized an auxillary verb “rahti hai*”* in Urdu sentence upon translation.

(U)-Hum thak kar chur ho gaye
(E)-We were dead tired.

In this example, the verb of Urdu sentence *“thak kar chur hona”* is realized as an Adjective “*dead tired”* in English sentence upon translation. If we consider English to Urdu case then the example that can be considered for illustration of demotional divergence is:

(E)- It suffices.

There are two possible translations for the above English sentence:

(U1)- yah kaafi hai.
(U2)- yah zaroorat ke mutaabik hai.

The word “suffice” is realized as the main verb in English but as an adverbial modifier kaafi hai/ zaroorat ke mutaabik hai in Urdu. Here the role of suffice (verb) in English sentence is being demoted in Urdu sentence and realizes as an adverbial modifier.

D. Lexical Divergence

Lexical divergence arises out of the unavailability of an exact translation map for a construction in one language into another language. It also means that the choice taken for the target language word is not the literal translation for an English word. Consider the following English sentences:

(E)- Good luck!
(U)- Allah Ka Fazal Ho

The words used in Urdu translation are not the literal meanings of the English words in the source sentence. Another English sentence that can be considered is:

(E)- It is cold.
(U)- Maahaul Sard Hai

In this translation the word Maahaul in urdu does not have a corresponding map in the English sentence.

(E)- Excuse me!... (to ask for something)
(U)- Maazirat Chaahta Hoon

Here the English sentence verb "excuse" has been translated as a different verb meaning "want"- chaahta hoon. If we consider the following Urdu sentence to be translated to English

(U)-Vah chayan ki niid so rahaa hai
(E)-He is enjoying a sound sleep.

In this example, the verb of Urdu sentence *"so rahaa hai"* is converted into a different verb "*enjoying"* in English sentence upon translation. The verb enjoying as such has no word whose literal meaning exists in the urdu translated sentence.

E. Categorial Divergence

Categorial divergences are located in the sentences where there is a mismatch between parts of speech of the pair of translation languages. It is observed that this is the most common type of divergence that occurs in any pair of language. Since it is concerned with the POS of source language and target language, this type of divergence arises if the lexical category of a word changes during the translation process. Let us illustrate this category of divergence with certain examples. Let us consider the following English sentence:

(E)- She is jealous of me

This sentence can be translated in Urdu in two ways:

(U1)- vah mujhse jalt**i** hai.
(U2)- usako mujhse jalan hai

In English sentence the word, 'jealous' is an adjective and it is realized as a main verb in the first sentence of Urdu, whereas in the second Urdu translation it is being realized as a noun.

(E)- They are waiting.
(U)- ve intezaar kar rahe hein.

Here "waiting" is expressed as a verb in the English sentence whereas in the Urdu translation it is realized as a combination of Noun and verb (*intezaar kar).* It's a very common form of divergence in English to Indian languages [Dave et al, 2001].

Some of the examples mentioned for this category by authors of [Saboor et al, 2010] are:

(U)- vah kitaabi kidaa hai
(E)-He is a bookworm.

In this example, the Adjective *kitaabi kida* of the Urdu sentence is realized as a Noun *bookworm,* in English sentence upon translation. For Urdu into English translation, this divergence also occurs when the subjective complement of SL upon translation is realized as a verb in TL [Gupta et al, 2001;2003].

(U)- murgi ande se rahii hai
(E)-The hen is hatching.

In this case, the Noun ande / *eggs* of Urdu sentence is missing in English sentence and is covered by the verb *hatching*. Another sentence that can be considered in this category is:

(U1)-usko zahar se maar diyaa gayaa
(U2)- usko zahar dekar maar diyaa gayaa

Both the above sentences can be translated to a single English sentence which is:

(E)- He was poisoned to death.

Here, the PP "zahar se" and "zahar dekar"in the Urdu sentence is converted into verb poison in English sentence upon translation.

F. Thematic divergence:

Thematic divergence refers to those divergences that arise from differences in the realization of the argument structure of a verb. Consider the following English sentence.

(E)- Where are you from?

This sentence can have the following three Urdu translations.

(U1)- Aap Ka Taaluq Kahan Se Hai?
(U2)- Aap kahan se Taaluq rakhate hein?
(U3)- Aap kahan se hein?

G. Reduplication:

Urdu, like Hindi and most of the South Asian languages, uses reduplication quite frequently (Abbi 1991). Content words can generally be reduplicated and the effect of the reduplication is to either strengthen/emphasize the original word or to express something like "and those kinds of things". The English counterparts of these constructions do not resort to replicative structure. This distinction may often result into a change in the category of the relevant elements.

Consider an Urdu sentence "*Khudaa zarre- zarre mein basta hai*"- when translated to English this can be simply achieved by the simple English sentence "God is everywhere". Another feature that is exhibited by Indian languages is occurrence of Echo words. The following examples illustrate this:

(U)- *Kya aap kucch thandaa vandaa lenge.*
(E)- Will you take some soft drink?

The echo words generally have no lexical status in the lexicon of the language. However these are used very commonly.

H. Honorific:

Like Hindi in Urdu also honorific features are expressed by several linguistic markers including the use of plural pronoun and plural verbal inflections. This feature is not available in a European language such as English in a similar way. This also causes a type of divergence during the translation process.

(U1) *unake vaalid aaye hein.*
(E)- His father has come.
(U2) *uskaa dost aayaa hai.*
(E)- His friend has come.

In (U1), the subject "valid"/ 'father' is an honorific noun which is reflected by the use of plural inflectional elements on the agreeing elements such as verb and the genitive noun. On the other hand, in (U2), dost 'friend' is a non-honorific noun and no plural inflectional element is used in the sentence.

I. Null subject Divergence:

In Hindi the subject of the sentence can be left implicit, which is not the case in English. Hindi allows dropping of the subject where the subject is obvious [Dave et al, 2001]. Similar situation has been observed for Urdu language as well. This can be illustrated with the following examples:

(E)- Long ago, there lived a king.
(U)- mudaton pahale, ek baadshah tha.

Here there is no explicit mapping for the word "there" of English sentence in the Urdu translation, it is assumed implicitly.

(U)- jaa rahaa hoon.
(E)-Iam going.

The subject "mein/ I" is missing and the presence of this missing subject is reflected in the morphology of the predicate. However the subject needs to be explicitly mentioned in the English sentence.

References:

[Dave et al., 2001] S. Dave, J. Parikh, and P. Bhattacharyya, "Interlingua-based English-Hindi machine translation and language divergence," Machine Translation. vol. 16(4), 2001, pp. 251-304.

[Dorr, 1993] Dorr, Bonnie. *Machine Translation: A View from the Lexicon*. Cambridge, Mass: The MIT Press, 1993.

[Dorr, 1994] B. J. Dorr, "Machine translation divergences: a formal description and proposed solution," ACL , Vol. 20.(4) , 1994, pp. 597-633.

[Gupta et al, 2001] D. Gupta and N. Chatterjee, "Study of divergence for example based English-Hindi machine translation," STRANS 2001. IIT Kanpur, 2001, pp. 132-139.

[Gupta et al, 2003] D. Gupta and N. Chatterjee, "Identification of divergence for English-to-Hindi EBMT," MT Summit-IX, Orleans. LA, 2003, pp. 141-148.

[Gupta , 2005] D. Gupta,"Contributions to English to Hindi machine translation using example-based approach," Ph. D. thesis, Indian Institute of Technology, New Delhi-110016. India, Jan 2005.

[Sinha et al, 2005a] R. M. K. Sinha and A. Thakur, "Translation divergence in English-Hindi MT," EAMT 10th annual conference, Budapest. Hungary, May 2005, pp. 245-254.

[Sinha et al, 2005b] R. M. K.Sinha and A. Thakur, "Divergence patterns in machine translation between Hindi and English," MT Summit X. Phuket. Thailand, Sept 2005, pp.346-353.

[Sinha, 2009] R. Mahesh K. Sinha, Developing English-Urdu Machine Translation via Hindi, Third Workshop on Computational Approaches to Arabic Script-based Languages (CAASL3), MT Summit XII, Aug.26-30, 2009, Ottawa, Canada, pp. 89-95.

[Saboor et al, 2010] Abdus Saboor, Md. Abid Khan, Lexical-Semantic Divergence in Urdu-to-English Example Based Machine Translation, 6th International Conference on Emerging Technologies (ICET), 2010, pp. 316-320.

IndoWordNet and Multilingual Resource Conscious Word Sense Disambiguation

Pushpak Bhattacharyya
Computer Science and Engineering, IIT Bombay, India

Wordnets have become crucial resources for NLP. They are complex structures capturing various kinds of lexico semantic relations among words. The first wordnet in the world was built; for English at Princeton University. This was followed by wordnets of European languages forming the EuroWordnet. At IIT Bombay the first wordnet for Indian languages was constructed for Hindi. This was followed by many other languages including Marathi, Sanskrit, Bangla, Tamil, Telugu, Punjabi, Gujarathi, and North East languages. In the first part of the talk we describe the principles and methodolgies followed in multilingual wordnet construction. We close this part of the discussion with a brief description of the Pan-Indian multilingual dictionary standard that IndoWordnet has given rise to and is the essential resource for multilingual WSD.

Word Sense Disambiguation (WSD) is a fundamental problem in Natural Language Processing (NLP). Amongst various approaches to WSD, it is the supervised machine learning (ML) based approach that is the dominant paradigm today. However, ML based techniques need significant amount of resource in terms of sense annotated corpora which takes time, energy and manpower to create. Not all languages have this resource, and many of the languages cannot afford it.

In the second part of the presentation, we discuss ways of doing WSD under resource constraint. First we describe a novel scoring function and an iterative algorithm based on this function to do WSD. This function separates the influence of the annotated corpus (corpus parameters) from the influence of wordnet (wordnet parameters), in deciding the sense. Next we describe how the corpus of one language can help WSD of another language, i.e., LANGUAGE ADAPTATION. This is presented in three setting of "complete", "some" and "no" annotation. From this we move on to DOMAIN ADAPTATION where the notion of active learning and injection are pursued to do WSD in a domain with

little or no annotated corpora. The extensive evaluation and good accuracy figures lend credence to the viability of our approach which points to the possibility of expanding from one language-domain combination to all language-domain combinations for WSD, i.e., multilingual general domain WSD, a long standing dream of NLP.

The talk is presented in a multilingual setting of Indian languages. There are 22 official languages in India with strong requirements of machine translation and cross lingual search. Our languages of focus in this talk are Hindi and Marathi along with English and the domains of focus are Tourism and Health which are important to India.

The presentation is based on work done with PhD and Masters students and researchers: Dipak Narayan, Nitin, Rajat, Deabsri, Mitesh, Salil, Saurabh, Anup, Sapan and Piyush, and published in fora like ACL, COLING, EMNLP, GWC, ICON and so on.

A semantic map and a lexical compass to help people find the words they are searching for.

Michael Zock

CNRS & LIF (Laboratoire d'Informatique Fondamentale)
université de la Méditerranée (Aix-Marseille II)
UMR 6166, Case 901 - 163 Avenue de Luminy
F-13288 Marseille Cedex 9, FRANCE,
michael.zock@lif.univ-mrs.fr

Abstract. NLP has produced very few applications to support people in their quest to understand or produce language. Yet, such applications are not only badly needed, they are also possible. I will deal here only with one problem, *word finding* (lexical access or word retrieval), showing how electronic dictionaries could be improved to help language producers (speaker, writer) find the words they are looking for. Since finding supposes searching, I suggest to build similar tools to the ones we use elsewhere for this activity, namely, a *map* and a compass. While the *semantic map* defines the territory within which search takes place, the *lexical compass* guides the user, helping her or him to reach the goal (target word). I will discuss how such tools or resources can be built and how they can be used for search.

Keywords: language production, word access, production mode, electronic dictionary, mental lexicon, speaker/writer.

1 Problem: finding the needle in a haystack

One of the most vexing problems in speaking or writing is that one knows a given word, yet one fails to access it when needed. This kind of impairment occurs not only in language production, but also in other activities of everyday life. Being basically a search problem it is likely to occur whenever we look for something that exists in the outside world (objects) or our mind: dates, phone numbers, past events, peoples' names, or, 'you just name it'.

As one can see, I am concerned here with the problem of words, or rather, how to find them in the place where they are stored: the human brain or an external resource, a dictionary. My work being confined to lexical access, I have started to develop with some colleagues a *semantic map* and a *lexical compass* [25] to help language producers find the word they are looking for. More precisely, we try to build an index and a navigational tool, allowing people to access words, regardless of their input, which may be changing and incomplete. Our approach is based on psychological findings concerning the *mental lexicon* [1, 15],— i.e. *representation*, *storage* and *access* of information in the human mind,— observed search strategies and *rational* navigation, the goal being to find quickly and naturally the word one is looking for.

2 The lexicon, a resource

Obviously, a dictionary is a vital component for NLP by and large. Yet, depending on the task (reading, translation) and the material support (hardware, brain) there are different kinds of dictionaries: paper dictionaries, electronic dictionaries, and the mental lexicon. Each one of these resources have their own virtues and shortcomings with respect to coverage, updates, flexibility of input, speed of output, etc.

2.1 The mental lexicon

The *mental lexicon* (henceforth, ML) refers to the knowledge people have with respect to words. Obviously, the very notion is only a metaphor, as, unlike paper- or electonic-dictionaries, which store words in a specific place, there is no special area in the human skull reserved for the storage of 'words'. Actually, producing them involves the solution of many tasks: analysis of the situation (possibly visual input), determination of meaning (*goat* vs. *sheep*), syntactic category (POS), morphology, pronunciation, etc. Put differently, producing a word involves many areas of our brain. Unlike other dictionaries (paper, electronic) which store meaning and forms next to each other (knowledge encapsulation, i.e. local representation), the ML distributes *meaning*, *form* and *sound* across various layers [7, 15], see figure 1. It is precisely this kind of distribution that causes word access problems, but it is also this very same feature that accounts for the incredible flexibility of the human mind.

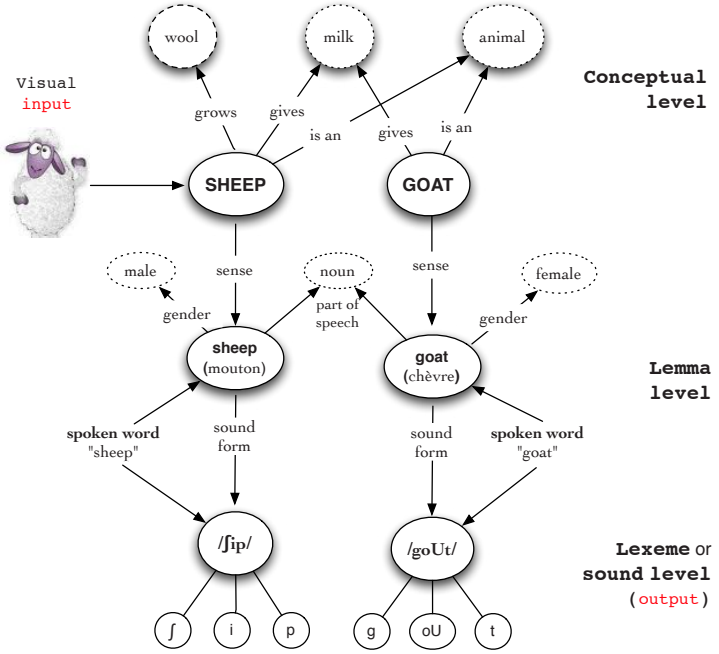


Figure 1: Information distribution, based on Levelt and colleagues' work [15]

Information distribution is supported by many empirical findings like *speech errors* [4, 10, 12][1], studies in *aphasia* [6], experiments on *priming* [16] or the *tip of the tongue phenomenon* [3]. The latter is particular relevant for us, as it shows that people experiencing word-access problems do know many details concerning the target word (meaning, form, related words), a fact that one should draw upon.

There are many other features caracterizing the ML. Let us consider just four of them: (a) high-dimensional connectivity: everything is connected, be it only indirectly; (b) uneven population of the network: some areas are more densely populated than others; (c) nodes are evoked via association: a stimulus (possibly a source-word) triggering one or several outputs (possibly containing the desired target word); (d) links have weights which change dynamically.

Of course, all these features are relevant for our problem. Facts (a) and (c) imply that we have a multi-dimensional, associative network, which allows us to reach anything from anywhere. The driving force pushing us in a given direction are the weights associated to the links, relating two nodes. Fact (b) embodies the notion of *semantic fields*, an idea which is nicely exploited via thesauri, lexical items being grouped by domain or in terms of a more general category: color, time, means of transportation, etc. The dynamic weight change mentionned in point (d) implies that a given stimulus may evoke different words depending on the topic or context. Think of the term 'piano' in the contexts of music (null context) or house-moving.

2.2 Are dictionaries dedicated or a task independant resource?

The question is, do we have one dictionary or several, depending on the task. While many scholars believe that dictionaries are theory neutral with respect to the communication mode (speaking/listening), they are mistaken. Most dictionaries have been designed with the reader in mind. Yet, the readers' needs, input and processing are clearly different from the writers'. *Speaking* is not the inverse of *listening*. The output of a parser (tree) is not a good input for a generator. What holds for sentences is also true for words. Dictionaries for the language producer (speaker/writer) are not just the inverse of dictionaries for the language receiver (listener/reader) [10]. While their end-point and starting-point look alike (meanings/ concepts vs. sounds/words), the way to get there is entirely different in each case. Figure 2 (next page) shows that the differences outweigh the similarities. As one can see, the activated neighbours, i.e. respective search spaces, are quite different. While the word '*safe*' may trigger 'money box, assured, out of danger' when reading, it will probably evoke synonyms like 'vault' or 'strong box' when speaking. In sum, the intersection between the two sets is extremely small. Only one item is shared (safe), the rest of the elements being disregarded, as they are simply not relevant for the other task.

This being so, one may well conclude that in order to get from *forms* to the corresponding *meaning*s (trunk : 'suitcase' or 'nose of an elephant'?) is by no means just the inverse of finding for *meanings* the adequate *forms* ('name for the *hat of a*

[1] For example, the difference between the *meaning-* (conceptual) and *form-* or sound-level is motivated by speech errors like sheep/goat and sheep/sheet, where the first word is the target and the second the actual, mistakingly produced word.

bishop' : tiara, mitre?)[2]. Yet, this seems to be assumed when dictionary builders provide readers and writers with a single resource for both tasks, or when they suggest that creating a dictionary for the speaker is basically just taking a conventional dictionary and endowe it with reverse access [10].

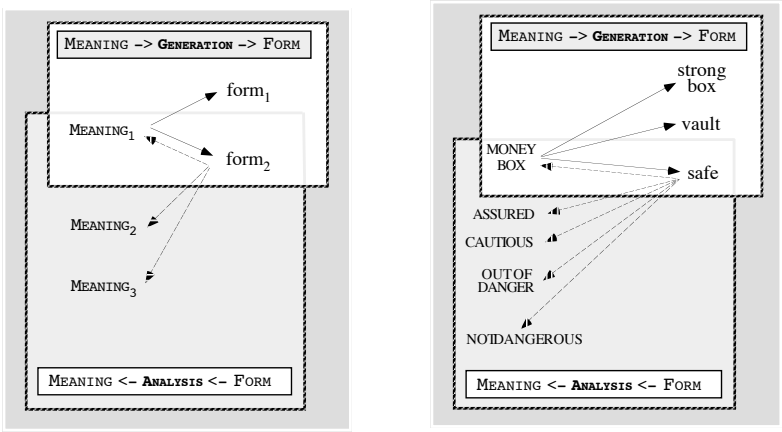


Figure 2: Search spaces varying with the task

I will be concerned here only with the language producer's problem, that is, given some meaning, help him to find the corresponding lexical form (lemma). Obviously, authors (speakers or writers) start from meanings (concepts), but, as we all know, this is by no means enough to guarantee access to the desired form. Words may elude us, and this is generally the moment where we reach for a dictionary. Unfortunately, most of them (in particular, paper dictionaries) are not well suited for the language producer[3]. Little provision is made to allow access based on conceptual input, i.e. the target word's meaning (conceptual search). Yet, with the advent of computers (digital corpora, powerful interfaces allowing flexible display of information) this can be changed.

Still, this is not yet sufficient. In order to build the required functionalities we need to consider the users' varying cognitive states and search strategies. We should also take into account empirical findings from the neurosciences and psycholinguistics. This taken together may allow us to define guidelines for computational lexicographers, helping them to design electronic dictionaries, indexes and nagivational tools suited for the language producer. If our intuitions, observations and

[2] *Ambiguity* (the receiver's problem) seems to be the equivalent to the notion of *choice* (synonymy, paraphrases, etc.), the language producer's problem.

[3] To be fair one must admit that efforts have been made to improve the situation. Actually, there are quite a few *onomasiological* dictionaries. For example, Roget's *Thesaurus* [20], *analogical dictionaries* [2, 19], *Longman's Language Activator* [23] and, of course, all the network-based dictionaries: *WordNet* [11], *MindNet* [18], *HowNet* [8] and *Pathfinder* [22]. There are also *collocation dictionaries* (BBI, OECD), various *Reverse Dictionaries* [9, or, http://onelook.com/reverse-dictionary.shtml] and Rundell's *MEDAL* [21], a hybrid version of a dictionary and a thesaurus, produced with the help of Kilgariff's Sketch Engine [13].

implementations are right, this should yield products in line with the users' goals, their search habits and their organisation of words in the mind, the *mental lexicon* [1].

2.3 A dictionary for the language producer

Obviously, dictionaries for the writer ought to be different from dictionaries for the reader, be it only with regard to the input, the structure, the index and search facilities. There are at least three things that authors know when looking for a specific word: its *meaning* (definition) or at least part of it, its *lexical relations* (hyponymy, synonymy, antonymy, etc.), and the relations it entertains with other words (associations, collocations). Put differently, there are several ways to access a word: via its *meaning* (concepts), via *lexical relations*, via *thesaurus-* or *encyclopedic* relations, i.e. associations (see figure 3). While we will mainly draw on this last possibility, we have presented in [25] the building principles of a lexical matrix which integrates all of them into a single resource.

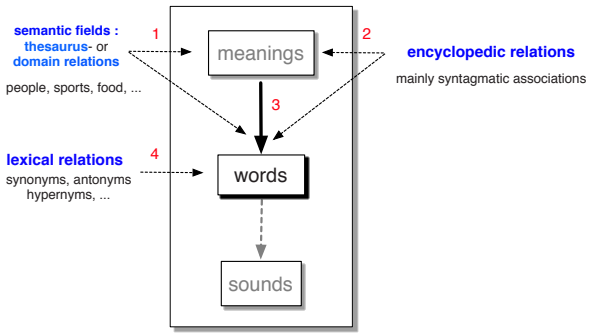


Figure 3: Lexical access via different methods or routes[4]

Yet, people seem to know more than that. Psychologists who studied people being in the tip-of-the-tongue-state [3, 24] have found that their subjects had access not only to meanings (the word's definition), but also to information concerning grammar (gender) and lexical form: *number* of syllables, *beginning*/ *ending* of the eluding word, *part of speech*. While all this information could be used to constrain the search space, —the ideal dictionary being multiply indexed,— we will deal here only with the target word's meaning (definition, bag of words), lexical relations and otherwise

[4] This feature of the ML is very important, as, if one method fails, it allows us to resort to another. It is also worth noting that a thesaurus and an encyclopedia can have an impact at two different levels, the level of ideas (conceptual level) and the level of words (linguistic level). Hence, a thesaurus can be used for conceptualization, i.e. message specification (meaning) as well as for its expression (lexicalization). Indeed, authors often start from a broad concept (ANIMAL), which they gradually narrow down (reptile), before committing to a specific lexical form: alligator/crocodile/caiman. Likewise, *encyclopedic* relations (associations) may be used for message creation, as well as for finding a concrete lexixal form. Concepts evoking (other) concepts (red-fire) and words priming other words (coffee-strong). In both cases the same mechanism is at work, though at different levels and operating on different elements: concepts in one case, words in the other.

semantically related words (associations, collocations in the large sense of the word). Before discussing how such a dictionary could be built and used[5], let us consider a possible search scenario.

3 Dictionaries and searching : a possible scenario

When looking for a word, people tend to start from a close neighbour. Let's admit for the sake of the argument that they cannot think of such a word. The only token coming to their mind being a word they know to be somehow connected to the T_w. Suppose, you were to express the following ideas: *superior dark coffee made from beans from Arabia*, knowing that neither *espresso* nor *cappuccino* are the intended word. While none of this would lead you directly to the desired word, *mocha*, the information at hand, i.e. the word's definition or some of its elements, could certainly be used. In addition, people draw on knowledge concerning the *role* a concept (or word) plays in language and in real world, i.e. the associations it evokes. For example, they may know that they are looking for a *noun* standing for a *beverage* that *people* take under certain circumstances, that the *liquid* has certain properties, etc. In sum, people have in their mind an encyclopedia: all words, concepts or ideas being highly connected. Hence, any one of them has the potential to evoke the others. The likelihood for this to happen depends, of course, on factors such as *frequency* (associative strength), *distance* (direct vs. indirect access), *prominence* (saliency), etc.

Let us see how this could work. Suppose you were looking for the word *mocha* (target word: T_w), yet the only token coming to your mind were *computer* (query- or source word: S_w [6]. Taking this latter as starting point, the system would show all the connected words, for example, *Java*, *Perl*, *Prolog* (programing languages), *mouse*, *printer* (hardware), *Mac*, *PC* (type of machines), etc. querying the user to decide on the direction of search by choosing one of these words. After all, s/he knows best which of them comes closest to the T_w. Having started from the S_w 'computer', and knowing that the T_w is neither some *kind of software* nor a *type of computer*, s/he would probably choose *Java*, which is not only a *programming language*, but also an *island*. Taking this latter as the new starting point s/he might choose *coffee* (since s/he is looking for some kind of *beverage*, possibly made from an ingredient produced in Java, coffee), and finally *mocha*, a type of *beverage* made from these beans. Of course, the word *Java* might just as well trigger *Kawa* which not only rhymes with the S_w, but also evokes *Couawa*, also written *Kawa*, an argotic word of *coffee* in French, or *Kawa Igen*, a javanese volcano.

As one can see (figure 4), this approach allows word access via multiple routes, many ways leading to Rome. Also, while the distance covered in our example is quite

[5] To avoid possible misunderstandings, I would like to stress the fact that we are not building yet another dictionary. We start from an existing one, trying to add one or several indexes in order to ease navigation, as this is precisely what is lacking in most resources.

[6] Note that this, just like many other homonyms, might lead us to a completely different domain: *island* vs. *programming language*. In other words, the link 'homonym' can be considered as a shortcut in the network.

unusual, it is possible to reach the goal quickly. It took us actually very few moves, to find a connection between the world of *computers* to *coffee beans* or the drink made out of them. Of course, *cyber-coffee* fans or people thinking of a particular port city on the Red Sea coast of Yemen (Mocha) might be even quicker in reaching their goal.

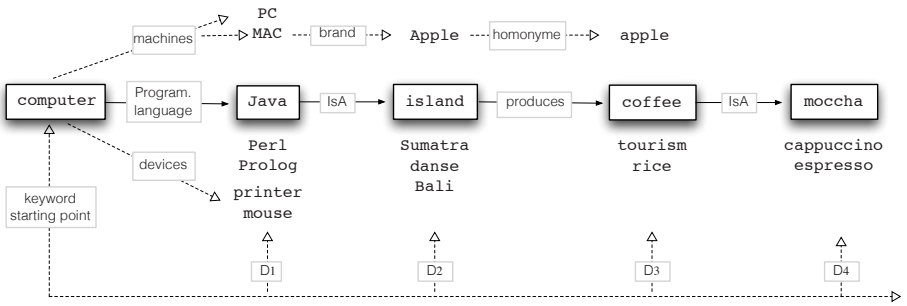


Figure 4: Finding a remote item at the distance of four mouse clicks.

4 Building the resource, i.e the semantic map

The semantic map is basically a lexical graph, all words being connected, and the links or connections being typed and weighted. There are various methods to build such a map. One way is to ask people to get lists of associations [5]. This has been the main strategy of psychologists, who built word association norms [17]. Another way is to use games [14]. Still another approach is corpus-based, by extracting collocations. This is the route we are taking. The following four problems need to be solved.

- **Building a representative corpus**: we need a well-balanced corpus. Since the corpus is supposed to represent the user's world knowledge, this latter must be reflected in the corpus. In other word, the corpus must contain a little bit of everything normal people know concerning the world in general (objects and relations), but also information concerning specific recent events in sports, politics, etc.
- **Indexing**: words have to be indexed. We do this in terms of the associations. To this end we need to discover the trigger words and their collocates. Psychologists have built such lists already decades ago [5, 22]. Similar lists are nowadays freely available on the web. For example, there is the Edinburgh Associative Thesaurus (http://www.eat.rl.ac.uk/) and the compilation of Nelson et al. (http://cyber.acomp.usf.edu/FreeAssociation). There are also resources for German (http://www.schulteimwalde.de/resource.html, http://www.coli.uni-saarland.de/projects/nag/), and Japanese, http://www.valdes.titech.ac.jp/~terry/jwad.html. Next to using the kind of association lists gathered by psychologists, one may try to build this kind of information bottom-up, via corpora and by using a collocation extractor.

- **Ranking**: the weight of the linked words needs to be determined (relative frequency). This is important for ranking the associated words. Ideally, the weight is (re-)computed on the fly to take into account contextual variations. A given word, say Java, is likely to evoke quite different associations depending on the context: *tourism* vs. *programming*.
- **Identification and naming the links**: Associations must not only be identified, they must also be labeled. Qualifying, i.e. typing the links is the hardest task, yet it is vital for navigation. Frequency alone is not only of limited use —(people cannot interpret properly numerical values in a context like this),— it is even misleading: two terms of very similar weight (say, 'mouse' and 'PC') may belong to entirely different categories ('computer device' vs. 'type of computer'), hence choosing one instead of the other may have important consequences concerning the chances of finding or not the desired word.

5 Navigating in the resource by using a lexical compass

Once this resource is built, access is straightforward. The user gives as input a word he believes to be directly or indirectly connected to the T_w[7], say *hospital*, to which the system would answer with all immediate associates ('clinic, sanatorium, doctor', i.e., A_1 - A_4 in figure 5). If the list contains the T_w , search stops, otherwise it continues. The user chooses a word from the list (say, 'doctor', i.e. A_2), or a word being evoked by them (indirect association), and the system will reveal again all directly associated terms: 'surgeon, pediatrician, medic', i.e. $B_1 – B_3$.

Bear in mind that this problem cannot be solved via the well-known algorithm computing the shortest path between two nodes in a network, as this supposes that one knows both points. Yet, in our case we know only the starting point (user query), but not the end point (goal), as if we knew it, there would be no need for search to begin with, we would just display the T_w. Nevertheless, eventhough the user does not know the T_w, he can recognize it (goal) if he sees it in a list [8]. He can do even more. If in response to his query (S_w) we give him a set of words, he will know which one of them is closest related to the T_w. Put differently, search is interactive, the user providing an input (query, or S_w), deciding where to go and when to stop, and the system provides hints, potential T_{ws}, or words indirectly related to it.

As one can see, there are some important differences between a conventional compass and our navigational tool. While the former automatically points to the north, letting the user compute the path between his current location and the desired goal (destination), the latter assumes the user to know the goal or at least its direction. While the user cannot name the goal (he has only passive knowledge), the system cannot guess it. However it can make valuable suggestions. In other words, the

[7] This is, or course a simple case. One could also think of several terms as input.

[8] This kind of passive knowledge is somekow akin to the tip-of-the tongue state. As Brown and McNeill [3] have shown, people being in this state have a lot of information concerning the T_w, to the point that, if this latter is presented to them, they can recognize it without ever making any mistake.

system can give hints concerning potential goals, it is nevertheless the user who decides on the direction to go, as only he knows which suggestion corresponds to the goal or which one of them is the most closely connected.

5.1 Potential interface problems

Since words occur in many different settings or syntactic contexts, every word is likely to have a rich set of connections. Obviously, the greater the number of words associated to a term, and the more numerous the type of links, the more complex the graph will be. This reduces considerably their interface value and their potential to support navigation. There are at least three factors impeding readability:

- *high connectivity* (great number of possibly different kinds of links or associations emanating from each word);
- possible *crossing of links* in the case of indirect association (see $A_2 - B_3$ or $A_3 - B_2$ in figure 5)[9] ;
- *distribution*, i.e. non-adjacency, of conceptually related nodes, that is, nodes activated by the same kind of association, but not being displayed next to each other (see, $B_1 - B_3$), which is quite confusing for the user.

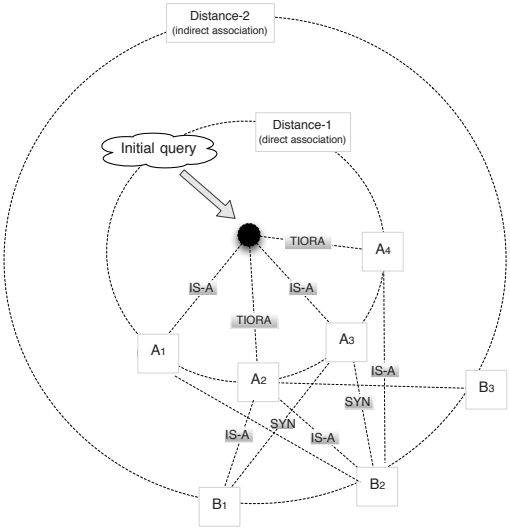


Figure 5: Potential problems with graphs: crossing lines.
`IS-A` (subtype of); `TIORA` (**T**ypically **I**nvolved **O**bject, **R**elation or **A**ctor)

[9] Note, that the crossing of lines can be avoided in the immediate neighbourhood (distance 1, i.e direct associations), but not at the next level, yet, the scope being the entire graph and not only the next adjacent level, i.e. the direct neighbours.

5.2 A possible solution

I believe that there is a fairly straightforward solution to the problem. Here is the underlying rationale. Since all words are connected, we have have a graph, where everything can be reached from anywhere, regardless of the starting point. Yet, our graph can also be seen as a set of trees. Since search could be launched at any point, —any node of the graph could become the query or starting point,— we have as many trees as the graph contains nodes. This input (source word, say 'hospital'), would become the root of the tree, and the associated terms, i.e output or potential target words would be the leaves: 'clinic, sanatorium, doctor, nurse'. If the input and output are linked via different kinds of association ('hospital-clinic vs. 'hospital-ambulance'), we create an intermediate node for each link, giving it the name of the link ('subtype' vs. 'part-of'). Put differently, we create as many nodes as there are different kinds of links emanating from a given node.

In conclusion, rather than displaying all the connected words as a graph or as a huge flat list, we display them in hierarchically organized categorial clusters, factorizing links, so that the nodes present now all words having the same kind of relationship with respect to the S_w. Put differently, we suggest to display by *type* or category (chunks) all words entertaining a specific link with respect to the input, i.e. the S_w (see figure 6). Since the same word may occur at various levels[10] our trees allow for recovery. If one has taken a wrong turn at some point, one may still reach one's goal via a detour.

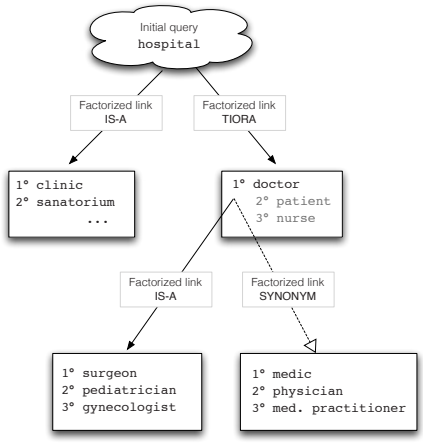


Figure 6: Search-tree with two links, `IS-A` and `TIORA`.
`IS-A` (subtype of); `TIORA` (**T**ypically **I**nvolved **O**bject, **R**elation or **A**ctor)

This kind of presentation seems clearer and less overwhelming for the user than graphs, as it allows for *categorical search*, which is a lot faster than search in a huge bag of words. Of course, this supposes that the user knows to which category a word belongs to, and that the labels, i.e. link names are well chosen. This is crucial, as the

[10] For example, a word like 'coffee' may be connected both to 'beverage' and to 'major export product'.

names must be meaningful, i.e. interpretable by the user, which, may be problematic in our example here. Figure 6 presents visually the rationale outlined here above, except that the S_w and T_w are not 'computer' and 'mocha' but 'hospital' and 'nurse'.

As one can see, the fact that the links are labeled has some very important consequences: (a) While maintaining the power of a highly connected graph (possible cyclic navigation), it has at the interface level the simplicity of a tree: each node points only to data of the same type, i.e. to the same kind of association. (b) With words being presented in clusters, navigation can be accomplished by clicking on the appropriate category. The assumption being that the user generally knows to which category the target word belongs (or at least, he can recognize within which of the listed categories it falls), and that categorical search is in principle faster than search in a huge list of unordered (or, alphabetically ordered) words .

6 Conclusion and future work

I have started this paper with the obvious, though often overlooked fact that language producers contact the lexicon with different expections and knowledge states than the language receivers. Writing is not the inverse of reading. I have then taken a quick look at some of the work done by pschologists studying the ML (structure, functioning). Finally, I have outlined some ongoing work that tries to build on such findings, hence the usage of notions like "assocation, network, weights, etc.".

A dictionary should contain not only a lot of information, it must also reveal it when needed. To this end I suggested to create a semantic map and a lexical compass. The former defines a well structured space for search, whereas the latter provides hints, if not guidance. While the succes of this work will largely depend on the quality of the map and the adequacy of the search methods, it will also depend on the system's capacity to take into account the users' ever changing needs and knowledge states.

Among the issues to be addressed in the near future there is the building of a prototype for a small domain. Yet, there are also several hard problems to be addressed, namely the notion of *links*,— according to my knowledge, there is no complete or satisfying list available at the moment,— and the interpretation of user *queries*. Obviously, a given query will have different 'meanings' depending on the context (moment) of its usage. For example, in the 'mocha' example, the user giving 'island' as key expects specific information concerning the 'island Java' and not just any island, or 'islands in general'. While obviously more work is needed, we do believe that this line of research is worth the effort. It addresses a practical problem, but it goes well beyond language. It deals with human memory, i.e. the way how information is represented, structured and accessed.

References

1. Aitchison, J. (2003). Words in the Mind: an Introduction to the Mental Lexicon. Oxford, Blackwell.

2. Boissière, P. (1862). Dictionnaire analogique de la langue française : répertoire complet des mots par les idées et des idées par les mots. Paris. Auguste Boyer
3. Brown, R. and Mc Neill, D. (1966). The tip of the tongue phenomenon. Journal of Verbal Learning and Verbal Behavior, 5: 325-337
4. Butterworth, B. (1982). Speech errors: Old data in search of new theories. In A. Cutler (Ed.). Slips of the tongue and language production. Amsterdam: Mouton : 73-108
5. Deese, J. The structure of associations in language and thought. Johns Hopkins Press. Baltimore. 1965
6. Dell G., Schwartz M., Martin N., Saffran, E. and D. Gagnon. (1997). *Lexical access in aphasic and nonaphasic speakers*. Psychological Review, 104, 801-838.
7. Dell, G. (1986): A spreading-activation theory of retrieval in sentence production. Psychological Review, 93, 283-321.
8. Dong, Z. and Dong, Q. (2006). HOWNET and the computation of meaning. World Scientific, London.
9. Edmonds, D. (ed.), *The Oxford Reverse Dictionary*, Oxford University Press, Oxford, 1999.
10. Fay, D. and Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. Linguistic Inquiry, 8 : 505-520.
11. Fellbaum, C. (ed.) (1998). WordNet: An Electronic Lexical Database and some of its Applications. MIT Press.
12. Fromkin V. (ed.). Errors in linguistic performance: Slips of the tongue, ear, pen and hand. New York: Academic Press.
13. Kilgarriff, A., Rychly, R., Smrz, P. and Tugwell, D. (2004). The Sketch Engine. In: Proceedings of the 11th Euralex International Congress. Lorient, France : 105-116
14. Lafourcade, M. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In 7th International Symposium on Natural Language Processing, Pattaya, Chonburi, Thailand.
15. Levelt, W., Roelofs, A. and Meyer, A.S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22 : 1-75.
16. Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227-234.
17. Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. http://www.usf.edu/FreeAssociation/
18. Richardson, S., Dolan, W. and Vanderwende, L. (1998). Mindnet: Acquiring and structuring semantic information from text. In: ACL-COLING'98. Montréal: 1098-1102.
19. Robert, P., Rey A. and Rey-Debove, J. (1993). Dictionnaire alphabetique et analogique de la Langue Française. Le Robert, Paris.
20. Roget, P. (1852). Thesaurus of English Words and Phrases. Longman, London.
21. Rundell, M and Fox, G. (eds.) (2002). Macmillan English Dictionary for Advanced Learners (MEDAL). Oxford
22. Schvaneveldt, R. (ed.) (1989). Pathfinder Associative Networks: studies in knowledge organization. Ablex. Norwood, New Jersey, US.
23. Summers, D. (1993). Language Activator: the world's first production dictionary. Longman, London.
24. Vigliocco, G., Antonini, T. and Garrett, M. F. (1997). *Grammatical gender is on the tip of Italian tongues*. Psychological Science, 8, 314-317.
25. Zock, M., Ferret, O. & Schwab, D. (2010) Deliberate word access : an intuition, a roadmap and some preliminary empirical results, In *A. Neustein (éd.) 'International Journal of Speech Technology'*, 13(4):107-117, 2010. Springer Verlag

Machine Translation

Identifying the Translations of Idiomatic Expressions using TransSearch

Stéphane Huet[1] and Philippe Langlais[2]

[1] LIA - Université d'Avignon, Avignon, France
stephane.huet@univ-avignon.fr
[2] DIRO - Université de Montréal, Montréal, Québec, Canada
felipe@iro.umontreal.ca

Abstract. This document presents a case study relating how a user of TransSearch, a translation spotter as well as a bilingual concordancer available over the Web, can use the tool for finding translations of idiomatic expressions. We show that with some care on the queries made to the system, TransSearch can identify a fair number of idiomatic expressions and their translations.

1 Introduction

Idioms are expressions of a given language, whose sense is not predictable from the meanings and arrangement of their elements [8]. For example, an expression like "*to be hand in glove*" meaning "*to have an extremely close relationship*" cannot easily been deduced from what a hand and a glove are. Some expressions are more analyzable than others; for instance, the meaning of the expression "*fights like cat and dog*" might easily be inferred by the senses of "*cat*" and "*dog*". This is not so for the expression "*it rains cats and dogs*". In this work, we are interested in identifying the translation of this second type of expressions.

Idioms — and more generally Multi-Word Expressions (MWEs) — pose significant problems for many applications of natural language processing since they are numerous in most languages and have idiosyncratic meanings that severely disturb deep analysis [11]. The problem of MWEs — and idioms in particular — is especially acute in the case of Machine Translation (MT) where a failure of the system to detect such expressions often leads to unnatural, if not hilarious outputs.

Therefore, one important component of an MT system is its lexicon of MWEs. This is true for rule-based MT systems as well as statistical MT (SMT) ones. Currently, state-of-the-art phrase-based SMT systems rely on models (pairs of phrases) that do not handle MWE specifically. Some authors have been trying to group multi-word expressions before the alignment process [4] or to add a new feature encoding the knowledge that a given phrase pair is a MWE [10, 2]. This last work showed that using manually defined WordNet MWEs could improve MT.

Not only are idioms interesting for improving MT systems, they are as well notably known to pose problems to non-native persons. This is especially true

when a second-language idiom is much different from its translation into the native language. For instance, French speakers might easily catch the English idiom "*play cat and mouse*" because its French translation "*jouer au chat et à la souris*" is literal in this case. On the contrary, they could find hard to understand "*He couldn't say boo to a goose*"[3] because its translation into French "*Il est d'une timidité maladive*" (literaly "*He is sickly shy*") is completely different.

Idiomatic expressions are interesting for professional translators as well. In [6], the authors analyzed the most frequent queries submitted by users to the bilingual concordancer TransSearch. They found that among others things, users frequently queried idiomatic phrasal verb expressions, such as "*looking forward to*". Because they were expecting that the users would query idiomatic expressions, they did not investigate further this aspect of the logfile, but instead concentrated on analyzing the prepositional phrases (some of which being idiomatic) frequently submitted to the system.

In this paper, we study the problem of translating idiomatic expressions from a user perspective. We tried to identify the translations of a number of idioms in the Translation Memory (TM) of the new version of the bilingual concordancer TransSearch. Since many idioms have inflected forms, we show the impact of different strategies for querying the database. For instance, in the (idiomatic) expression "*to keep oneself to oneself*", both the verb "*keep*" and the pronoun "*oneself*" can vary according to conjugation and inflection respectively, and verbatim queries may fail to identify relevant occurrences of the expression.

The remainder of the paper is organized as follows. Section 2 describes TransSearch, the Web application we employed in our experiments. Section 3 presents the data we used and how we submit queries to the TM system to find translations. Section 4 is dedicated to the evaluation of the translations proposed by the system, while Section 5 concludes and explores further perspectives.

2 TransSearch

TransSearch is a bilingual concordancer that allows its users to query large databases of past translations in order to find ready-made solutions to a host of translation problems. Subscribers of the system are mainly professional translators. A recent study of their query logs exhibits that TransSearch is used to answer difficult translation problems [6]. Among the 7.2 million queries submitted to the system over a six-year period, 87% contain at least two words. Among the most frequent submitted queries, several appear to be idiomatic, like "*out of the blue*" or "*in light of*".

2.1 System Features

Made available since 1996 through a Web interface by the Université de Montréal [7], TransSearch has recently been improved to become not only a bilingual concordancer but also a translation finder [1]. Figure 1 which displays the

[3] At the time of writing, *Google Translate* produces the literal translation "*Il ne pouvait pas dire boo à une oie*".

Fig. 1. Result returned by the new TransSearch to the query "*is still in its infancy*". The left column shows likely translations in decreasing order of likelihood, while the main columns shows concordances. The query and the selected translation are shown in color in each of them.

results for the query "*is still in its infancy*" exemplifies the new capabilities of the system. Where a simple bilingual concordancer (as were the previous versions of TransSearch) would only display a list of parallel sentences containing the query in their English part, the new version of TransSearch highlights for each sentence pair the French part associated with the query. Besides, this version displays on the left hand side the whole range of translations (automatically) found in the TM. For the first suggested translation, "*en est encore à ses premiers balbutiements*", three of the sentence pairs containing a variant of this translation (see the merging process described in Section 2.2) are displayed in context.

With respect to an ordinary bilingual concordancer, where the identification of translations in sentences is left to the user, we believe the new version of TransSearch dramatically improves usability, by displaying a general view of the TM content for a given query.

The previous query example has shown that the system is able to find results for queries with several words. The user can also submit more advanced queries to search discontinuous expressions. For example, Figure 2 displays the results for the query "*make .. hair stand on end*". The '..' operator enables the user to indicate the system that occurrences of 2 words in the query (here "*make*" and "*hair*") can be up to 5 words apart inside a sentence. Another operator '...' allows for searches without constraining the distance between two words. From a linguistic perspective, these two operators are useful since they enable the user to spot expressions where words may be separated by a few words, such as nominal groups in the examples of Figure 2.

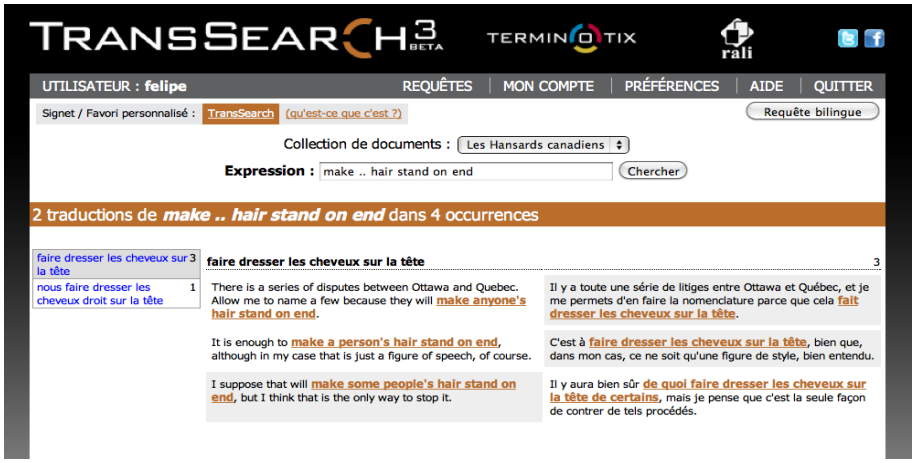


Fig. 2. Result returned by TRANSSEARCH to the query "*make .. hair stand on end*".

Besides, another advanced type of query is available in TRANSSEARCH: morphological expansions. The system considers all the morphological derivations of the terms associated with the '+' symbol, when retrieving sentence pairs. Figure 3 shows the results for the query "*take+ no for an answer*". In this example, the interface displays expressions containing different inflected forms of the verb "*take*". This last operator is specially useful for morphologically rich languages like French or Spanish and allows the user to spot translations without taking care of their possible inflections.

By default, TRANSSEARCH searches for the given expression regardless of languages (French or English). In some cases however, it is necessary to specify the language, for instance in order to distinguish between the French and English words "*tape*" ("*to hit*" in French). Using the same mechanism, it is also possible to look up occurrences of a specific translation of a given query by filling at the same time the French and English fields of the query form. For example, a user can check that "*les dés sont pipés*" is a correct translation of "*the dice are loaded*" by looking at the same time at these two expressions into the TM sentence pairs.

2.2 Processing Steps

In order to suggest several translations for a given query, TRANSSEARCH performs several processing steps that we briefly describe hereafter. Many current computer-assisted translation tools mainly rely on sentence-level matching to exploit their translation memory. TRANSSEARCH operates at a finer-grained level using word alignment techniques, which are commonly used in SMT. The term translation spotting, coined by [13] and relabeled here as *transpotting*, is defined as the task of identifying the target language word-tokens that correspond

Fig. 3. Result returned by TransSearch to the query "*take+ no for an answer*".

to a given source language query in a pair of sentences known to be mutual translations; it is a core step in the new version of TransSearch.

We call *transpot* the target word-tokens automatically associated with a query in a given pair of sentences. For instance in Figure 1, "*en est encore à ses premiers balbutiements*" and "*soit encore tout nouveau*" are 2 out of 14 distinct transpots displayed to the user for the query "*is still in its infancy*".

The method used to transpot queries in the retrieved sentence pairs is described in details elsewhere. In a nutshell, our transpotting algorithm uses statistical word-alignment models and enforces that the transpots identified are sequences of contiguous words. As mentioned in [12], contiguous tokens in the source language sentence tend to be aligned to contiguous tokens in the target language. This statement is confirmed by the good experimental results presented in the study of [1].

Queries that occur frequently in the TM receive numerous translations using the transpotting methods described above, some being clearly wrong, some others being redundant (morphological variations of the same translation). We estimate that a user will focus on the 10 first translations presented, so we want to provide as many correct and diversified translations as possible at the top of the result page. Therefore, two postprocessing steps were introduced inside the TransSearch engine. The first one filters out bad transpots using supervised learning. A classifier was trained on a corpus where transpots were manually labeled as "good" or "bad", using features such as the ratio of grammatical words inside the hypothesized transpots. Once transpots have been filtered out, the second step merges those which are different inflectional forms of the same sequence of canonical words. For instance, "*au nom du*" and "*au nom des*" will be considered as similar, since "*du*" and "*des*" are contractions of "*de + le*" and "*de + les*" respectively, where "*le*" and "*les*" are definite articles. Furthermore,

as it was noticed that translations that differ only by a few grammatical words or punctuation marks, like "*de la part de*" and "*part de*" are often redundant for the user, those are combined as well. At the end of this second post-processing step, only the most frequent transpot of each merged set is displayed on the left hand side of the user interface (see Fig. 1 to 3). These transpots are shown as a list sorted in the decreasing order of their transpotting frequency.

3 Methodology

3.1 Resources

Translation Memory The largest TM used in TransSearch comes from the Canadian Hansards, a collection of the official proceedings of the Canadian Parliament. For our experiments, we used an in-house sentence aligner [5] to align 8.3 million French-English sentence pairs extracted from the 1986-2007 period of the Hansards. This bitext was indexed with Lucene[4] to form our TM.

Idiom Lexicon Classifying an expression as idiomatic or not is not an easy task. Therefore, we resorted to the phrase book [9] written by Jean-Bernard Piat, a translation teacher as well as a translator. This book oriented towards general public market provides a list of 1,467 idiomatic expressions in both languages (French and English) categorized by subjects (e.g. human body).

According to the author, the expressions were chosen because they are frequently used. A minority of these expressions are expressed in an informal language (e.g. "*to be well-upholstered*"). He also mentioned that it happens sometimes that an idiomatic expression in one language (e.g. "*to burn the midnight oil*") is not idiomatic in the other language (e.g. "*travailler tard dans la nuit*").

Examples of entries in this book are reported in Table 1. A few entries have several equivalent translations such as "*make your flesh creep*" and "*give you goose pimples*" for "*donner la chaire de poule*". Globally, there are on average 1.17 English translations and 1.01 French translations per entry.

All expressions but seven, are used in the context of a sentence. According to the author, using expressions in a context makes them easier to understand and to use for the readers. The lexicon contains a high proportion of verbal phrases (around four out of five of the available entries) that are used in their inflected form, like "*He took to his heels*" for the phrase "*to take one's heels*". Other entries are fixed expressions such as "*When there's a will, there's a way*" or "*Hands off!*".

3.2 Preprocessing

In order to take into account contextualization that makes lexicon entries too specific, the used lexicon was manually annotated by the first author of this paper.

[4] http://lucene.apache.org

Table 1. Excerpt of the entries we considered in our experiment. R stands for the reference translation, G stands for the translation made by Google Translate (provided as a proxy to literal translation). Words in parenthesis have been manually marked as contextual words that are not part of the idiomatic expression.

French	English
Il est agile comme un singe	R *He's as nimble as a goat* G *He is agile as a monkey*
Elle était sur son trente et un	R *She was dressed to kill* R *She was all dressed up* G *She was on her thirty-one*
(Je vais d'abord) me rincer la dalle — familiar —	R *(I'm going to) wet my whistle (first)* G *First I'll rinse my slab*
(Il aime) rouler des mécaniques — familiar —	R *(He likes) flexing his muscles* R *(He likes) playing the tough guy* G He loves rolling mechanical
J'ai vu trente-six chandelles	R *I saw stars* G *I saw thirty-six candles*

All words judged as extra-information with respect to the idiomatic expression were annotated as such in the lexicon. Those are the words in parenthesis in the examples of Table 1. They are typically modal verbs (e.g. "*can*", "*must*"), semi-modal verbs (e.g. "*am going to*", "*are likely to*"), catenative verbs (e.g. "*want to*", "*keep*"), adverbs (e.g. "*only*", "*finally*"), adverbial phrases (e.g. "*in Italy*", "*when he heard the news*") or noun phrases (e.g. "*this poet*", "*his latest book*"). Finally, at least one word was classified as extra-information for 486 out of 1,467 entries.

3.3 Queries to the Translation Memory

In order to test the ability of TransSearch to find translations for idioms, three types of queries were submitted to the system: queries built from either the English side or the French side of the entry, and bilingual queries where both sides where searched for at the same time. As mentioned in Section 3.1, a few entries have more than one English or French reference translations. For these entries, we collected results found from all the equivalent translations. Since TransSearch user interface doest not allow users to write an "or" operator between several equivalent translations, we had to simulate the behavior of this operator by submitting independently translations and then by merging results retrieved by TransSearch.

Table 2 shows the number of lexicon entries found in the TM, using bilingual (column 2), English (column 3) or French queries (column 4) and considering various ways of querying the system. As expected, building verbatim queries from the lexicon leads to retrieve information inside the TM for a small number

Table 2. Number of the lexicon entries found inside the translation memory using several types of query.

Query types	bilingual	English	French
verbatim queries	37	136	248
EN: *I have no axe to grind*			
FR: *Je ne prêche pas pour ma paroisse*			
+ manual removal of extra words	91	302	410
EN: *I have .. axe to grind*			
FR: *Je .. prêche .. pour ma paroisse*			
+ removal of extra pronouns	106	381	509
EN: *have .. axe to grind*			
FR: *prêche .. pour ma paroisse*			
+ verb lemmatization	210	624	650
EN: *have+ .. axe to grind*			
FR: *prêcher+ .. pour ma paroisse*			
+ pronoun and determiner lemmatization	238	700	705
EN: *have+ .. axe to grind*			
FR: *prêcher+ .. pour sa+ paroisse*			

of expressions only (line 1). After taking into account the manual preprocessing step introduced in Section 3.2, that is, after removing extra words, nearly three times as many queries have at least one hit in the TM (line 2). Still, at best, a user could retrieve no more than 410 (French) expressions from the 1,467 ones by simply querying them verbatim or removing extra words.

An inspection of the submitted queries revealed that many of them correspond to flexible idioms, that is, idiomatic expressions that can vary from one occurrence to another. In order to capture those variations and to increase therefore the number of hits in the TM, we wrote rules that abstract away some of those variations. For this, we used a mix of linguistic information as well as the operators we described earlier. We resisted to the temptation of adjusting this process for each query and instead applied some rules in a systematic way, given a set of linguistic markers semi-automatically annotated in the lexicon.

The performed processing steps for the entry [“*I have no axe to grind*”, “*Je ne prêche pas pour ma paroisse*”] are illustrated in Table 2. A set of rules deleted personal pronouns at the beginning of an expression (see line 3); a list of pronouns to be removed has been collected for this in each language. Then, lemmatized verbs were replaced by the corresponding lemma and auxiliary verbs were removed (see line 4); we used for this an in-house lemmatization resource available for both languages. Last, we also considered lemmatizing pronouns and determiners within an expression (see line 5).

It should be noted that we chose to modify entries using a set of limited rules in order to avoid over-abstracting idiomatic expressions. For instance, we noticed that the indefinite pronoun "*it*" in English usually occurs in fixed expressions and cannot be replaced by another personal pronoun. Therefore, we kept this pronoun verbatim in the queries made.

We observe in Table 2 the dramatic increase of the number of hits in the TM according to the level of abstraction of the query. At best, the rewriting rules we applied allow TransSearch to return sentence pairs for 700 English entries and for 705 French entries, i.e. roughly half of the lexicon. Each set of rules increases the number of queries with at least one hit. Surprisingly, verb lemmatization led to a higher improvement of the coverage for English queries than for French ones. This shows that, on the contrary to what we expected first, this process is also relevant for weakly inflected languages.

This experiment also shows that in order to get the best of the system, users should use the linguistic operators at their disposal. Since we know that most queries made by real users of the application do not use those operators it could mean one of two things. When users submit a query to the system without getting any answer, they might simply abandon the search for a translation or on the contrary, they might figure out a way to process the query in order to find a match in the TM. Inspecting the log-files of the application exhibits evidences that both strategies happen in practice. This means that automatically processing the query of a user is an interesting prospect to consider.

Another interesting outcome of the experiment we conducted is that the Hansards indexed by TransSearch are rather good for identifying the idiomatic expressions we considered.

4 Evaluation

We have measured the quantity of idiomatic expressions we could find by querying the Hansards indexed by TransSearch. We now turn to the evaluation of how good the application is for spotting the translations of the retrieved expressions. Once again, it should be noted that in most bilingual concordancers we know of, this part is left to the user.

4.1 Objective Evaluation

For the French and English queries obtained after applying our rewriting rules, TransSearch was able to retrieve on average respectively 36.1 and 31.7 sentence pairs from the TM. Among this material, the transpoting algorithm identified respectively 12.5 French and 14.9 English (different) translations (shown to the user on the left of the navigator). Since a manual analysis of all the suggested translations would be a tedious task, an evaluation was performed thanks to the sanctioned translations belonging to the idiom lexicon described in Section 3. As shown in Table 2 (last line), a query and its sanctioned translation are found simultaneously in the sentence pairs returned by the system for 238 lexicon entries.

Table 3. Recall (%) measured using the lexicon sanctioned by the translation memory as a reference.

k	1	2	3	5	10	all
English queries	41.6	56.3	59.2	65.1	69.3	74.8
French queries	41.6	49.6	54.6	62.6	69.3	76.5

Therefore we restrained our objective evaluation to those 238 queries. Table 3 provides the proportion of those queries where the k-first translations displayed by TransSearch contain (at least) one of the reference translations sanctioned by the lexicon.[5]

The recall of 75% measured when all the translations returned by the system are considered demonstrates that the embedded transpotting algorithm has the ability to find translations in the retrieved sentence pairs. The result of 41,6% obtained when considering the first translation returned by the system (that is, the most frequent one) is not bad either, especially since the reference we used is rather incomplete. For instance, our lexicon contains the translation "*être dans un état second*" for the idiom "*to be in a daze*", while TransSearch displays this translation after "*est nébuleux*", which is as well a good translation of the English idiom. Similarly, TransSearch returns no less than 34 different translations[6] of the query "*be+ around the corner*", most of which being perfectly legitimate translations, while our reference contains only one.

4.2 Manual Evaluation

The objective evaluation revealed the great potential of TransSearch for identifying the translation of idiomatic expressions, but also showed that a manual evaluation was required in order to account for the sparseness of our bilingual lexicon. Therefore, we conducted a manual evaluation involving 5 bilingual annotators that were presented with lists of identified translations among 100 randomly chosen French queries. They were asked to indicate in those lists those translations that they found correct, partially correct or wrong. No specific guidelines were given to explain these labels. At the time of writing, 50 queries were judged by 3 annotators and the 50 other by 2.

Globally, the quality appreciated by the annotators turned out to be variable, some annotators tending to classify more easily translations as correct. This translated into a low value of 0.25 obtained when computing the Fleiss inter-annotator agreement [3]. Figure 4 illustrates some cases of divergence.

The results of this evaluation are reported in Table 4. Since a given query can be rated differently by several judges, we credited divergent annotations equally.

[5] In order to account for inflectional variations, we compared lemmatized translations.

[6] The 10 most frequent ones are: *est à nos portes*, *arrive à grand pas*, *était imminent*, *nous attend*, *me guette*, *est sur le point*, *s'annonce*, *est en vue*, *sommes au bord de*, and *survenir*.

appeler un chat un chat	J1	J2	J5
▷ we should call it what it is	correct	correct	correct
▷ we can say the d word and the m word	correct	wrong	partial
▷ calling manure a rose doesn't change the smell	correct	wrong	partial

manger à tous les râteliers	J1	J2	J5
▷ slurps at everyone 's trough	correct	correct	correct
▷ double - dipper	partial	correct	partial
▷ them pot lickers and accusing them of being at the trough and pork barrelling	wrong	partial	wrong

Fig. 4. Examples of annotations of some French idiomatic queries.

For instance, if a translation is judged correct by one annotator, and wrong by another one, a credit of 0.5 will be given to each label.

For all but 7 queries, TransSearch is able to identify a translation classified as correct by at least one annotator. For these queries, the average rank of the first correct translation is 1.4, which indicates that relevant translations can usually be found among the two first displayed by TransSearch. Also, on average, we observe that only 36% of the translations proposed to the user are labeled as wrong.

Table 4. Average percentage of translations judged correct, partially correct or wrong per query on a sample of 100 English queries randomly selected. *avr* stands for the average number of translations produced per query, while *rank* indicates the average rank of the first translation labeled as correct by at least one annotator.

correct	partial	wrong	*avr*	*rank*
42%	22%	36%	13.4	1.4

5 Conclusion

In this work, we have studied the problem of identifying translations of idiomatic expressions in both English and French, using a brand new version of the bilingual concordancer TransSearch. We showed that a user that would query the system verbatim would often fail to find a match in the TM and that some cleverness is required in order to get good use of the system, such as resorting to the morphological (+) and the proximity (..) operators available in the query language recognized by the system. We automatized the querying process and showed that a rough half of the idiomatic expressions queried to the system finally got a match in the TM, while a high proportion of the translations returned by the system are correct.

Acknowledgments

This work was funded by an NSERC grant in collaboration with Terminotix.[7] We are indebted to Sandy Dincky, Fabienne Venant and Neil Stewart who kindly participated to the annotation task.

References

1. Julien Bourdaillet, Stéphane Huet, Philippe Langlais, and Guy Lapalme. TransSearch: from a bilingual concordancer to a translation finder. *Machine Translation Journal*, 24(3–4):241–271, 2010.
2. Marine Carpuat and Mona Diab. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proceedings of NAACL-HLT*, pages 242–245, Los Angeles, CA, USA, 2010.
3. Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Pai. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York, NY, USA, 3rd edition, 2003.
4. Patrik Lambert and Rafael Banchs. Data inferred multi-word expressions for statistical machine translation. In *Proceedings of MT Summit*, pages 396–403, Phuket, Thailand, 2005.
5. Philippe Langlais. A system to align complex bilingual corpora. Technical report, CTT, KTH, Stockholm, Sweden, 1997.
6. Elliott Macklovitch, Guy Lapalme, and Fabrizio Gotti. TransSearch: What are translators looking for? In *Proceedings of AMTA*, pages 412–419, Waikiki, Hawaii, USA, 2008.
7. Elliott Macklovitch, Michel Simard, and Philippe Langlais. TransSearch: A free translation memory on the World Wide Web. In *Proceedings of LREC*, pages 1201–1208, Athens, Greece, 2000.
8. Tom McArthur, editor. *The Oxford Companion to the English Language*. Oxford University Press, 1992.
9. Jean-Bernard Piat. *It's raining cats and dogs et autres expressions idiomatiques anglaises*. Librio. J'ai lu, 2008.
10. Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the ACL-IJCNLP Workshop on Multiword Expressions*, pages 47–54, Suntec, Singapore, 2009.
11. Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15, Mexico City, Mexico, 2002. Springer.
12. Michel Simard. Translation spotting for translation memories. In *Proceedings of the HLT-NAACL Workshop on Building and using parallel texts: data driven machine translation and beyond*, volume 3, pages 65–72, Edmonton, Canada, 2003.
13. Jean Véronis and Philippe Langlais. *Evaluation of Parallel Text Alignment Systems — The Arcade Project.*, chapter 19, pages 369–388. Kluwer Academic Publisher, 2000.

[7] `www.terminotix.com`

Addressing SMT Data Sparseness when Translating into Morphologically-Rich Languages

Alexandru Ceauşu[1] and Dan Tufiş[2]

[1] Centre for Next Generation Localisation, Dublin City University
[2] Research Institute for Artificial Intelligence, Romanian Academy
aceausu@computing.dcu.ie; tufis@racai.ro

Abstract. The phrase-based translation approach has overcome several drawbacks of the word-based translation methods and proved to significantly improve the quality of translated output. However, they show less improvement on translating between languages with very different syntax and morphology, especially when the translation direction is from a language with limited word order and morphological variations to a highly inflected language. We describe an experiment that uses morpho-syntactic descriptions to translate and generate morphological information in factored machine translation. We show that from English to a morphologically rich language this setting has better performance than the baseline phrase-based system, when only a small parallel corpus is available. Also, we show that it scales well to a large parallel corpus when additional target monolingual corpus is available.

Keywords: statistical machine translation, morphologically-rich languages

1 Introduction

The phrase-based translation approach has overcome several drawbacks of the word-based translation methods and proved to significantly improve the quality of translated output. However, it shows less improvement on translating between languages with very different syntax and morphology, especially when the translation direction is from a language with limited word order and morphological variations to a highly inflected language. Tree-based models were introduced to handle long range reordering – what is believed to be the most difficult part to model in in statistical machine translation (SMT). The rich morphology of a highly inflected language permits a flexible word order, thus shifting the focus from long-range reordering to the selection of a morphological variant. Translating the correct surface form realization of a word is dependent not only on the source word-form, but it also depends on additional morpho-syntactic information.

Morphologically rich languages have a large number of surface forms in the lexicon to compensate for a flexible word order. The large number of word-forms can make very difficult to establish translation equivalents classes between the lexicons.

Both Transfer and Interlingua MT employ a generation step to produce the surface form, from a given context and a dictionary form of the word. In order to allow the same type of flexibility in using the morpho-syntactic information in translation, factored translation models [1] provide the possibility to integrate the linguistic information into the phrase-based translation model.

Most of the SMT approaches that have as target a morphologically rich language employ factored translation models. Our approach is similar to several other factored machine translation experiments such as adding the morphological features as factors [2], adding supertags on source language [3], and mapping syntax to morphology [4]. Our results are comparable with the ones reported in papers describing Arabic-English SMT experiments. For large amounts of training data, applying only a minimal segmentation in the Arabic part of the corpus yields better results than the baseline; however, when only a limited amount of training data is available, better results are achieved with part-of-speech tags and complex morphological analysis [5]. The importance of the generation model is highlighted in [6] through its usage in a hybrid (rule based and SMT) Arabic-English system.

2 Morpho-syntactic Description Codes

In highly inflectional languages, encoding the morpho-syntactic properties of the word-forms requires a large set of description codes. The Multext European project in co-operation with EAGLES Lexical Specification Group developed a set of recommendations [7] for the languages in Western Europe. Starting with these specifications, the Multext-East Copernicus project further developed them so as to account for the specificity of six other languages from Central and Eastern Europe – Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene [8]. The size of the tag-set greatly differs among languages: from a tag-set of around 100 tags in English to more than 2000 tags in Slovene.

Data sparseness in tagging highly inflectional languages with large tag-sets and scarce training resources is a problem that cannot be addressed using only common tagging techniques. Tiered tagging [9] is a two-stage technique addressing the issue of data-sparseness. It uses an intermediary tag-set of a smaller size on the basis of which a common POS tagging technique can be used. In a second phase, it replaces the tags from the small tag-set with tags from the fully-specified morpho-syntactic tag-set (MSD tag-set) also taking into consideration the context. Tiered tagging relies on the assumption that the values of a part of the attributes in a MSD tag (the determinant attributes) and the word-form are sufficient to infer the rest of the attribute values. Based on this assumption, in the second phase of tiered tagging, the original MSD tag-set is recovered using a lexicon and a set of hand-written rules. The original idea of tiered tagging has been extended in [10], so that the second phase is replaced with a maximum entropy-based MSD recovery. In this approach, the rules for POS to MSD conversion are automatically learnt from the corpus. Therefore, even the POS labels assigned to unknown words can be converted into MSD tags. If an MSD-lexicon is

available, replacing the POS label for the known words by the appropriate MSD tags is almost 100% accurate.

3 Data Preparation

3.1 SEE-ERA.net Corpus

This corpus has been compiled within a SEE-ERA.net project [11] and it is based on the much larger JRC-Acquis multilingual corpus [12]. 1200 documents of high quality sentence alignment were extracted from the JRC-Acquis corpus. The documents have translations in all the languages of the project (Bulgarian, Greek, English, Slovene and Romanian) plus Czech, English, French, and German. The SEE-ERA.net corpus has morpho-syntactic description codes for Bulgarian, Czech, Greek, English, Slovene and Romanian. The aligned documents have 60,389 translation units of approximately 1.4 million tokens per language. For the experiments in 4.1 we used the English-Bulgarian, English-Greek, English-Slovene and English-Romanian parts of the corpus. A sample of an aligned translation unit, with English, Romanian and Slovenian parts, is presented in Example (1).

```
<tu id="60389"><seg lang="en"><s id="32005L0004.n.26.1.en"><w
lemma="do" ana="Vmps">Done</w><w lemma="at" ana="Sp">at</w><w
lemma="Brussels" ana="Np">Brussels</w><c>,</c><w lemma="19"
ana="Mc">19</w><w lemma="January" ana="Ncns">January</w><w
lemma="2005" ana="Mc">2005</w><c>.</c></s></seg></tu>

<tu id="60389"><seg lang="ro"><s id="32005L0004.n.26.1.ro"><w
lemma="adopta" ana="Vmp--sf">Adoptat•</w><w lemma="la"
ana="Spsa">la</w><w lemma="Bruxelles"
ana="Np">Bruxelles</w><c>,</c><w lemma="19" ana="Mc">19</w><w
lemma="ianuarie" ana="Ncms-n">ianuarie</w><w lemma="2005"
ana="Mc">2005</w><c>.</c></s></seg></tu>                              (1)

<tu id="60389"><seg lang="sl"><s id="32005L0004.n.25.1.sl"><w
lemma="v" ana="Sl">V</w><w lemma="Bruselj"
ana="Npmsl">Bruslju</w><c>,</c><w lemma="19." ana="Mdo">19.</w><w
lemma="januar" ana="Ncmsg">januarja</w><w lemma="2005"
ana="Mdm">2005</w></s></seg></tu>
```

3.2 STAR Corpus

The STAR bilingual parallel corpus (Romanian-English) was developed during the research project STAR[1]. The parallel part of the corpus mainly contains juridical

[1] http://www.racai.ro/star/

documents, but it also includes journalistic type data. STAR also has a Romanian balanced monolingual corpus containing a large range of documents, from literary texts to news and scientific documents. The content of the STAR corpus is sourced from several other corpora:

- the DGT (Directorate-General for Translation) Translation Memory corpus, a juridical corpus based on the Acquis Communautaire [12];
- EMEA (European Medicines Agency), a corpus with medical content from the Opus Corpus [13];
- SE Times (Southeast European Times corpus), a journalistic corpus from the Opus Corpus [13];
- NAACL news, the English-Romanian journalistic corpus used for NAACL 2005 word alignment shared task [14];
- Romanian balanced monolingual corpus (20 million tokens).

The data cleaning stage for this corpus includes understated processing steps like: deleting the data that is duplicated, removing lines of text that are in other languages, removing lines or tokens of more than a specified character length, etc. Cleaning Romanian data collected from the web (NAACL, SE Times) was a real challenge. Besides spelling errors, there are three specific types of text distortions occurring in Romanian texts: (i) missing diacritical characters, (ii) different encoding codes for the same diacritical characters and (iii) different orthographic systems. When ignored, they have a negative impact on the quality of translation and language models and thus, on the translation results. For details on the process of diacritics restoration see [15].

Table 1. The contribution of each sub-corpus to the STAR parallel corpus

Corpus	Tokens (millions)		Sentence pairs (thousands)
	English	Romanian	
DGT Translation Memory	12.5	12	621
EMEA (Opus Corpus)	10	11	698
SE Times (Opus Corpus)	4.4	4.7	166
NAACL news	0.8	0.7	39
Raw total	27.7	28,4	1,525
Cleaned total	27.3	27,7	1,495

After Romanian text normalization, in order to create the EN-RO bitext, the processing stages of sentence splitting and tokenization had to be adapted to the respective languages. Sentence splitting and tokenization have shared resources like abbreviations, segmentation rules, token merging rules, etc. In the final stage of data

preparation – the bitext cleaning – we removed the sentence pairs that are too short or too long and the sentence pairs of a source to target ratio of more than 1/9. Table 1 shows the contribution of each sub-corpus to the STAR parallel corpus and the amount of the remaining data after cleaning.

The corpus was tokenised, POS-tagged and received morpho-syntactic annotation using our publicly available web services [16]. Each token is composed of four factors: (i) the word-form, (ii) lemma disambiguated with its lexical category, (iii) the POS-tag from the reduced tag-set, and (iv) the MSD. Table 2 provides an example for the annotations available in the STAR corpus.

Table 2. Example of annotated sentence pair

English	Romanian
Grounds / ground^Nc / NNS / Ncnp	Motive / motiv^Nc / NPN / Ncfp-n
of / of^Sp / PREP / Sp	de / de^Sp / S / Spsa
non-recognition / recognition^Nc / NN / Ncns	refuz / refuz^Nc / NSN / Ncms-n al / al^Ts / TS / Tsms recunoaşterii / recunoaştere^Nc / NSOY / Ncfsoy
for / for^Sp / PREP / Sp	hotărârilor_judecătoreşti /
judgments / judgment^Nc / NNS / Ncnp	hotărâre_judecătorească^Nc / NSRN / Ncfsrn
relating / relate^Vm / PPRE / Vmpp	în / în^Sp / S / Spsa
to / to^Sp / PREP/ Sp	materia / materie^Nc / NSRY / Ncfsry
parental_responsibility /	răspunderii_părinteşti /
parental_responsibility^Nc / NN / Ncns	răspundere_părintească^Nc / NSOY / Ncfsoy

4 Factored Translation with Morpho-Syntactic Description Codes

Factored translation models extend the phrase-based translation by taking into account, not only the surface form of the phrase, but also, additional information like the dictionary form (lemma), the part-of-speech tag or the morphological specification. It also provides, on the target side, the possibility to add a generation step. All these new features accommodates well in the log-linear model employed by the decoder:

$$P(e|f) = exp \sum_{i=1}^{n} \lambda_i h_i(e, f) \quad (1)$$

where $h_i(e,f)$ is a function associated with the pair e, f and λ_i is the weight of the function.

Factored translation offers great possibilities on modeling translation: (i) there can be several translation steps; (ii) the fluency of the output can be checked on different levels with several language models; (iii) long-range word reordering can be achieved with more than one reordering model; and (iv) on the target side, there can be different generation steps.

To improve the translation into morphologically-rich languages, the multitude of options provided by the factored translation can help validate the following assumptions:

- Aligning and translating *lemma* could significantly reduce the number of translation equivalency classes, especially for languages with rich morphology;
- *Part of speech affinities*. In general, the translated words tend to keep their part of speech and when this is not the case, the part-of-speech chosen is not random;
- The *re-ordering* of the target sentence words can be improved if language models over POS or MSD tags are used.

4.1 Multilingual Setting

Based on the SEE-ERA.net corpus, we tested, using the MOSES factored framework [17], several configurations of translation, generation and reordering steps. The language pairs tested were English-Greek, English-Bulgarian, English-Slovene and English-Romanian. After cleaning, we split the corpus into training, development and test sets resulting in almost 57,000 sentence pairs for training, 500 for the development test and 1000 for testing. The 4-gram word-form language models and the 5-gram POS or MSD language models were built only using the training data sets. Considering current practices for training SMT systems, our training corpus is very small, but, as we will show, the additional linguistic information, made available in the pre-processing phase, compensates for the scarcity in raw data.

In order to test the improvement of the factored model over the phrase-based approach, we built strong baseline systems for each language pair. The baseline systems were trained using word alignment on lemmas and they employ an additional lexicalised reordering model. The default distance reordering model operates in a window of tokens and provides a reordering cost given the difference between source and target positions. We choose to use a better reordering model for the baseline system. The lexicalised reordering model has a probability assigned for the position change (monotone, swap or discontinuous) of a target phrase given the source phrase.

The baseline phrase-based translation systems and the different factored configurations had their parameters tuned on the development set using MERT [18].

We found that translating lemmas and morpho-syntactic descriptors and then generating, accordingly, the word-forms achieved better results than the baseline phrase-based translation model. Table 3 presents BLEU scores [19] for some of the factored configurations tested for the English-Romanian part of the SEE-ERA.net corpus. The BLEU scores for the English-Romanian translation direction are consistent with the scores for the other language pairs in the SEE-ERA.net corpus, although some of the configurations could not be tested because the intermediary tag-set (POS tag-set) was not available for the Bulgarian, Greek and Slovene parts.

The first row in Table 3 is the baseline system. It has a translation model trained on word-forms (column 2), no generation model (column 3), a word-form language model (column 4) and a lexicalised reordering model trained on word-forms (column 5).

While the use of linguistically informed language models (wordform +MSD/POS) and translation and generations models (lemma+MSD/POS) ensured improvements over the baseline, we noticed a significant drop in performance when the system used the word-form or MSD reordering models instead of the distance model. One possible explanation for the drop in performance for configurations 5 and 6 is that the lexicalized reordering model is made redundant when using a language model over MSD (or POS tags).

Table 3. Different factored configurations and their BLEU scores for the English-Romanian part of the SEE-ERA.net corpus

Config	Translation model	Generation model	Language model	Reordering model	BLEU score
1	word-form	-	word-form	word-form	51.76
2	lemma	lemma -> word-form	word-form	distance	51.79
3	lemma POS	lemma -> POS lemma,POS -> word-form	POS word-form	distance	52.31
4	lemma MSD	lemma -> MSD lemma,MSD -> word-form	MSD word-form	distance	52.76
5	lemma MSD	lemma -> MSD lemma,MSD -> word-form	MSD word-form	word-form	46.39
6	lemma MSD	lemma -> MSD lemma,MSD -> word-form	MSD word-form	MSD	45.77

As a side-note, on this particular corpus, the high BLEU scores might be explained as a consequence of the nature of the corpus – juridical texts have limited vocabulary, with long sequences of words repeated across the entire corpus. Although the scores are higher than the ones reported on news test corpora, the differences in absolute BLEU points can still offer good indices on the performance of the different factored configurations.

One particular configuration of factored translation (configuration 4 in Table 3) has provided better results than others. The proposed configuration (see Fig. 1) can be summarized as: (i) translate lemmas, (ii) generate all possible word forms and associated morpho-syntactic descriptions corresponding to a given lemma, (iii) translate the associated morpho-syntactic descriptions and (iv) generate the target surface forms given the lemma and the morpho-syntactic description. In this configuration, the decoder uses two language models: one for the word-forms and another one for the morpho-syntactic descriptions.

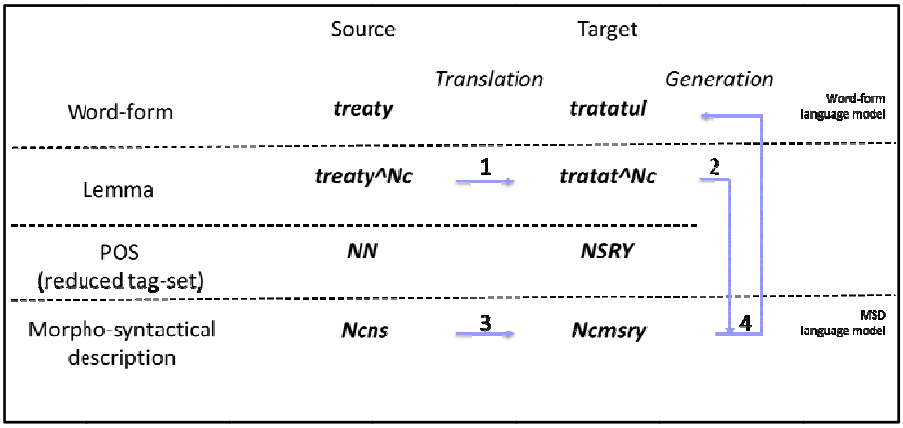


Fig. 1. Factored translation configuration with generation steps using lemma and MSD translation steps

We tested the systems using the BLEU score and we observed (see Table 4) improvements in accuracy ranging between 1% for Romanian and 2% for Slovene. Better handling of long-distance dependencies based on the MSD language model, a robust lemma translation equivalents table and a more precise selection of morphological variants are all possible explanations for the improvement in translation accuracy.

Table 4. BLEU scores for English-Bulgarian, English-Greek, English-Romanian and English-Slovene parts of the SEE-ERA.net corpus

Direction	Baseline	Factored
English-Bulgarian	38.94	39.60
English-Greek	42.22	43.07
English-Romanian	51.76	52.76
English-Slovene	40.73	42.68

The difference in BLEU scores between English-Romanian systems and the other systems are inherent to better lexical resources used for the tokenization and tagging of the English and Romanian texts. The idiomatic expressions and terminology tokenization were correlated between the English and the Romanian parts of the corpus.

4.2 English-Romanian Factored Translation

Using the STAR corpus (1.5 million sentence pairs) we tested if the factored configuration maintains its improvement over the baseline when a larger amount of training data is available. Similar to the experiments with the SEE-ERA.net corpus, we built a strong baseline system (lemma word alignment and lexicalised reordering model) that scored 53.82 BLEU points on 1000 sentences test-set (see Table 5).

For the bigger corpus, the English to Romanian factorized system achieves a BLEU score of 53.41, showing no improvement over the baseline system, on par with the results reported on other experiments with factored translation models [2], [20].

Table 5. English-Romanian factored translation on the STAR corpus

MT System	*BLEU score*
Phrase-based	53.82
Factored only parallel data	53.41
Factored plus monolingual data	54.52

Leveraging the fact that the generation step only deals with the target language, we used the STAR monolingual corpus in addition to the Romanian part of the parallel data to build the generation models and to train the MSD language model. The size of the statistical generation table trained on the 1.5 million sentences of the parallel data has almost 270,000 entries of the format shown in Example 2 (the format of an entry in the generation table is: <a b p(a|b) p(b|a)>):

```
complot^Nc/Ncms-n    complot   1.0000000   1.0000000
prăjit^Af/Afpfsrn    prăjită   1.0000000   0.9583333
absurd^Af/Afpms-n    absurd    1.0000000   0.3076923
încărca^Vm/Vmip1s    încarc    1.0000000   1.0000000
punctare^Nc/Ncfsoy   punctării 1.0000000   1.0000000
```
(2)

The size of the generation table that was trained on the additional Romanian monolingual data has almost 620,000 entries. Not all the additional entries are valid Romanian words. The monolingual corpus was built from texts of different domains (newspaper articles, old Romanian literature, contemporary literature, scientific articles, etc.) with different types of diacritical representations. We estimate that cleaning up the generation table would further improve the BLEU score. In the new configuration, the larger training data for the generation model and the MSD language model produced an increase of 0.7 absolute BLEU points over the baseline system (see Table 5).

4.3 Analysis of the Results

Although the factorized model (with additional monolingual data) has a marginal increase in BLEU score (at the cost of lower translation speed), we estimate that the actual improvements are higher from a human evaluation point of view. We observed that the factorized model frequently produces translations of better word order and more accurate morphological variant selection over the baseline model.

In order to assess in how many cases the translation system chooses the correct morphological variant, we investigated a difficult case of morphological attributes translation: the agreement of the words in a noun-phrases that include a conjunction. The baseline system, in this particular case of the test set, has a correct agreement in 61 of the 81 (75%) noun phrases that include a conjunction. The factored system with

a generation model trained on more monolingual data has the correct agreement in 75 of the cases (92%).

In Table 6 we present a case of noun phrase agreement in which the baseline system misses the correct morphological variant. The noun *prelucrarea* (processing) is (wrongly) a definite form while the noun *export* (export) is (correctly) an indefinite form.

Table 6. Example of noun phrase agreement in the English-Romanian phrase-based and factored translation systems

Token	*English*	*Romanian phrase-based*	*Romanian factored*
1	representative	piețe	piețe
2	markets	reprezentative	reprezentative
3	for	pentru	pentru
4	processing	prelucrarea (def)	prelucrare (indef)
5	and	și	și
6	export	export (indef)	export (indef)

5 Conclusion and Further Research

The paper presented two scenarios in which factored machine translation for morphologically-rich languages can show improvements in performance over the baseline phrase-based translation: (i) when there is very little amount of parallel data available and (ii) for a larger parallel corpus, when an additional, target-side, monolingual corpus with automatic annotations is available. The experiments described in this paper showed that an additional generation step on the target-side can prove as useful for statistical machine translation as it is for rule-based approaches to MT.

One major research priority in SMT is to overcome the scarcity of parallel data for less-resource language pairs. As future research, we are considering extending the factored experiment with comparable parallel data. The comparable data is available through the ACCURAT project. The aim of the ACCURAT project is to research methods and techniques to overcome one of the central problems of machine translation (MT) – the lack of linguistic resources for under-resourced areas of machine translation. The main goal is to find, analyze and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for less-resourced languages and narrow domains.

Acknowledgments. This work has been supported by the ACCURAT project (http://www.accurat-project.eu/) funded by the European Community's Seventh Framework Program (FP7/2007-2013) under the Grant Agreement n° 248347. It has also been partially supported by the Romanian Ministry of Education and Research through the STAR project (no. 742/19.01.2009).

References

1. Koehn, P., Hoang, H.: Factored Translation Models. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868–876, Prague, (2007)
2. Avramidis E., Koehn, P.: Enriching morphologically poor languages for statistical machine translation. In: Proceedings of ACL-08/HLT, pp. 763–770, Columbus, Ohio (2008)
3. Haque, R., Naskar, S.K., Ma Y., Way, A.: Using Supertags as Source Language Context in SMT. In: Proceedings of the 13th Annual Meeting of the European Association for Machine Translation, pp. 234-241, Barcelona (2009)
4. Yeniterzi, R., Oflazer, K.: Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 454–464, Uppsala (2010)
5. Habash, N., Sadat, F.: Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proc. of NAACL*, New York. (2006)
6. Habash, N., Dorr, B., Monz, C.: Challenges in Building an Arabic-English GHMT System with SMT Components. In *Proceedings of AMTA'06*, Cambridge, MA, USA (2006)
7. Monachini, M., Calzolari, N. (Eds.): EAGLES Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora A Common Proposal and Applications to European Languages http://www.ilc.pi.cnr.it/ EAGLES96/morphsyn/ morphsyn.html (1996)
8. Erjavec, T., Monachini, M. (Eds.): Specifications and Notation for Lexicon Encoding. Deliverable D1.1 F. Multext-East Project COP-106. http://nl.ijs.si/ME/CD/docs/ mte-d11f/ (1997)
9. Tufiş, D.: Tiered Tagging and Combined Classifiers. In: F. Jelinek, E. Nth (eds) *Text, Speech and Dialogue* LNCS vol. 1692, pp. 28-33 Springer-Verlag Berlin Heidelberg, (1999)
10. Alexandru Ceauşu: Maximum Entropy Tiered Tagging, Janneke Huitink & Sophia Katrenko (eds), *Proceedings of the Eleventh ESSLLI Student Session*, ESSLLI 2006, pp. 173-179 (2006)
11. Dan Tufiş, Svetla Koeva, Tomaž Erjavec, Maria Gavrilidou, and Cvetana Krstev: Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In Marko Tadić, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.) *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages* (FASSBL 2008), pp. 145-152, Dubrovnik, Croatia, September 25-28 (2008)
12. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th LREC Conference, Genoa, pp.2142-2147 (2006)
13. Tiedemann, J.: News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pp. 237-248, John Benjamins, Amsterdam/Philadelphia (2009)
14. Martin, J., Mihalcea, R., Pedersen T. (eds.): Proceedings of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond", Ann Arbor, Michigan, Association for Computational Linguistics (2005)

15. Tufiş, D., Ceauşu, A.: DIAC+: A Professional Diacritics Recovering System, in *Proceedings of LREC 2008* (Language Resources and Evaluation Conference), May 26 - June 1, Marakkech, Morocco. ELRA - European Language Resources Association (2008)
16. Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D.: RACAI's Linguistic Web Services, in *Proceedings of LREC 2008* (Language Resources and Evaluation Conference), May 26 - June 1, Marakkech, Morocco. ELRA - European Language Resources Association (2008)
17. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, demonstration session, Prague, (2007)
18. Och, F. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, pp 160-16, Association for Computational Linguistics (2003).
19. Papineni, K., Roukos, S., Ward, T., Zhu W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 311-318 (2002)
20. Birch, A., Osborne, M., Koehn, P.: CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Prague (2007)

Venetan to English Machine Translation: Issues and Possible Solutions

Suhel Jaber[1], Sara Tonelli[2], and Rodolfo Delmonte[1]

[1] Università Ca' Foscari, Venezia, Italy
[2] Fondazione Bruno Kessler, Trento, Italy

Abstract. In this paper we describe a prototype of a Venetan to English translation system developed under the *Stilven* project financed by the Regional Authorities of Veneto Region in Italy. The general approach is a statistical one with some preprocessing operations both at training and translation time (ortographic normalization and POS tagging to make use of factored models) which are needed especially to overcome two main problems: the scarcity of Venetan resources (our Venetan-English corpus is made up of only 13,000 sentences, amounting to 128,000 Venetan tokens) and the diasystemic nature of Venetan, which really represents an ensemble of varieties rather than a single dialect. We will present in detail the problems related to Venetan, our ideas to solve them, their implementation and the results obtained so far.

Keywords: Machine translation, less-resourced languages, language varieties

1 Introduction

Stilven[3] [7] is a project approved in December 2007 which started its activities in February of the following year. The task was creating a computational infrastructure for the analysis and translation of Venetan language (see for example [18]). Venetan is a dialect nowadays but was the official language of the Veneto Republic for as long as 8 centuries, up to the end of the XIXth century, when the Republic became part of newborn Italian nation. Since then, Venetan has been slowly abandoned in favour of Italian. Nowadays, depending on the region, Italian speakers can usually master a dialect and the main language. In particular, Venetan speakers show a much wider usage of dialect - their original language - in most working places, in the family and in social life.

Venetan proficiency by local speakers has been lately assessed as reaching 75% of the population in the Veneto region. Furthermore, more than 5 million speakers are scattered around the world, since in the past two centuries a large part of the population emigrated from Italy to other countries. Also, a small community of Venetan speakers is very active on the Internet, contributing to the diffusion of this language through several web-sites including a version of Wikipedia in Venetan (`http://vec.wikipedia.org/wiki/Vèneto`).

[3] `http://project.cgm.unive.it/stilven_en.html`

Venetan dialect is now considered a *diasystem*, where speakers use their own variety and manage to understand each other. Venetan is nowadays a spoken dialect, which has developed a number of varieties. In [6][4] , seven example dialogues are reported, each corresponding to a variety spoken in a Venetan city (i.e. Venice, Vicenza, Rovigo, Padua, Treviso, Belluno and Verona).

As to similarities, all varieties apart from Venetian use subject clitic inversion in questions. As an example of syntactic differences, we mention that Belunese is the only variety to allow verb fronting before question word: 'Féu che' (Do what), and clitic subject for weather verbs, 'Piòvelo' (It rains). Lexical differences are many and constitute the main distinguishing element: for example 'céo' (boy), is only used by Trevisan, while 'sani' (see you) is only used by Belunese. As to the remaining differences, they are all understood by the majority of Venetan people.

2 *Stilven* Project Objectives and Activities

Very much like what has been done with METIS [5], our system aims at translating free text input by taking advantage of a combination of statistical, pattern-matching and rule-based methods. The following goals and premises were defined for the project:

- use simple NLP tools and resources,
- use bilingual hand-made dictionaries,
- use Italian as intermediate language,
- use translation units at sentence boundaries,
- use different tagsets for source language (SL) and target language (TL).

Moreover, a translation system that has to cope with varieties has two main problems to solve:

- lexicon extension including all specialized items present in one variety and not in the others;
- grammatical flexibility that must properly process sentences with different structural organization according to each variety.

Syntactic peculiarities will be discussed in the next sections, whereas the problem of accounting for lexical varieties has been tackled by implementing a number of different lexica which refer at the same time to the four main varieties, to Italian and to English.

3 Linguistic Resources and Orthographic Normalization

In parallel to the implementation of the *Stilven* system, several linguistic resources were created in order to support the development of NLP applications

[4] Available at `http://www.linguaveneta.it/sussidiario.html`

for Venetan. First, we collected as much text as possible from the web and from people collaborating on a voluntary basis. Texts collected were then homogenized as to the orthography. They are organized into 7 different genres and include children stories, the translation of a book of American history, the translation of 'The Little Prince' by Antoine de Saint-Exupéry, the translation of a series of political newspaper articles, the translation of famous quotes taken from the LOGOS website (`www.logos.it`), the translation of a manual of the Venetan orthography rules [8] and a small set of especially built sentences directed to grammatical issues. As a whole, we collected texts for 200,000 tokens.

Also, frequency lists were compiled based on these texts. The lists were then the basis for the wordform lexicon of Venetan, which has been compiled on the basis of the Italian one available in our laboratory, thus comprising in each entry the corresponding Italian wordform and lemma. Semantic and syntactic properties of the Venetan wordform would then be derived directly from the Italian fully specified subcategorized lexicon.

We then normalized a big translation lexicon (52,000 entries) containing lemmas of Venetan paired with Italian and English. Moreover, we used parallel English-Italian texts to derive multiwords that could then be matched with those present in the Venetan-English parallel texts. From these materials we managed to collect a small dictionary of 200 multiwords which include very frequent function multiwords, like adverbial and prepositional locutions.

Normalization is a common issue to many languages in the world such as Arabic, Chinese and Japanese, which share the same problem of orthographic variation. Normalization is needed to allow the wordform to be checked against a lexicon where standardized orthography has been used. In our case, lexemes are produced in the lexicon with an official orthography according to the GVU (Unified Venetan Writing) obeying rules formulated some years ago by linguists [8] and published in the website of Veneto Region[5].

To make a comparison with Arabic, we see that orthographic variations may arise for a number of reasons, the first of which is certainly the dialectal variation. Then there is the objective problem of rendering some phonemes into a romanized valid corresponding character. As a result, an Arabic name may have hundreds if not thousands different variants in its romanized version. Coming back to Venetan, the problem is not so acute and the solution that can be adopted is the one that is also applied to other languages, that is an orthographic rule-based approach. In other words, due to the small number of variants it is not fit to use a lexicalized approach where all variants are stored after being automatically and then manually validated, for instance on the basis of their frequency of occurrence on the web. It will then be sufficient to list all cases of orthographic variations occurring in Venetan and then to formulate a corresponding set of rules. These rules coincide with what has been done for Arabic, for instance. In particular, consider the following rule for the recognition of some typical characters. As may be seen, the starting point is the corresponding phoneme, and on

[5] `http://win.elgalepin.org/gvu/index.html`

the right hand side there is a list of possible graphemes. Note that the mapping is one-to-many.

Example 1.
/dz, ts/ → d dh t z th

/k/ → k q c ch

/j/ → j g dj

The other remarkable orthographic problem concerns the need to use word stress on E and O to differentiate open vs. closed phoneme. The difference is crucial to characterize minimal pairs which otherwise would not be disambiguated, as for example in *béco* (goat) vs. *bèco* (beak), *péxo* (weight) vs. *pèxo* (worse), *bóte* (keg) vs. *bòte* (strikes), *fóla* (crowd) vs. *fòla* (lie).

So here again the problem lies in the lack of native speakers' awareness of the need to introduce such diacritics because they do not hear the ambiguity. Normalizing in this case is more complex because the meaning changes according to the type of accent chosen.

4 The Tagger

The first tool we worked on was the tagger of Venetan based on a semi-automatically annotated corpus of 128,000 words.

To increase the entries of the training corpus, we decided to decompose all idiomatic expressions and all locutions which amounted to some 2,000 entries. We also intended to use the 52,000 lexical entries that we collected as described in Section 3. So we added an article in front of all nouns and adjectives. Then, we composed pseudo sentences by joining nouns and adjectives to infinitival verbs (available in the lexicon) and adverbs. In this way, we collected another additional 85,000 entries which increased the size of the training corpus to almost 200,000 tokens.

4.1 The tagset

One of the interesting aspects of this work was the tagset we eventually came up with after a number of dubious cases, on the basis of our previous work on Italian. The POS with the corresponding meaning are reported in Table 1.

The most interesting cases regard the subdivision of cliticized verbs into three subcategories, namely VCL (verb cliticized, not inflected), VCLI (verb cliticized, inflected) and VPRON (verb inflected with cliticized subject pronoun). The reason for this subdivision is due to the need to separate VPRON from VCLI. While this is not needed in Italian and other Romance languages, Venetan requires another class of cliticized verbs, because it allows subject pronouns in questions. Here, the peculiarity is not only constituted by the amalgam in a

POS	Extended POS	POS	Extended POS
abbr	abbreviations, acronyms	num	number
ag	adjectives	par	parenthetical - punctuation “" ( ) -
art	articles - definite e indefinite	pk	complementizer “che"/that
avv	adverbials	prep	preposition
clit	clitic generic	part	preposition amalgamated with article
clitg	clitic “ghe” / there, you	pavv	prep./adverb “con su sora soto” /
nt	noun temporal		with, on, over, down
clits	clitic “se” / reflexive,impersonal	poss	possessive
ccom	conjunction “come" / like	prog	progressive periphrastic “drio”
cong	conjunctions “or, and"	pron	pronoun personal
congf	conjunction sentential	q	quantifier
cosu	conjunction subordinate	rel	relative pronoun “che" /
neg	negation		that, which, who, whom
date	number date	relob	relative pronoun oblique “cui” / whose
deit	deictic pronoun	relin	relative pronoun indefinite
dim	demostrative adjective	sect	sector number followed by fullstop
np	noun proper geographic		or parenthesis
dot	punctuation	v	verb inflected
fw	foreign word, also non-words	vav	verb “ver” / to have auxiliary and lexical
in	intensifier	vcl	verb cliticized non inflected
ind	indefinite quantifiers	vcli	verb cliticized inflected
int	interrogative pronouns	vd	verb gerundive
intj	interjections	vi	verb infinitival
n	noun common	vprog	verb progressive “star" / to stay
nh	noun proper human, appellation, social role	vpron	verb inflected with
punt	punctuation, : ;		cliticized subject pronoun
punto	punctuation sentence end . ? !		

Table 1. POS Tagset and explanation

question, but it is the clitic form that is very special. Final vowel is usually ‘-o’ for ‘-to’ (you) modifying the normal ‘-ti’ ending with ‘-i’. The use of ‘-ti’ is present and is determined by a phonological rule: the presence of a nasal ‘-n’ in the verb ending, as in ‘gonti’ (have you), ‘fonti’ (make you), ‘sonti’ (are you). We counted 647 cases of VPRON, 341 cases of VCLI and 1214 cases of VCL. It is important to note that these forms are very productive in conversations.

4.2 Tagger comparison

Given the tagged training corpus containing 218,864 tokens, we decided to compare the performance of two supervised taggers: the *Brill's Tagger* [4] in the Python implementation included in the NLTK suite [1][6], and the *HunPos Tagger* [9] [7], an open source reimplementation of the well-known *TnT tagger* [3],

[6] Available at `http://www.nltk.org/`
[7] Available at `http://code.google.com/p/hunpos/`

based on HMM. In this way, we compare for the first time the behaviour of Brill's transformation-based approach and of HMM-based statistical approach on Venetan documents. A similar comparison was performed for example by [10] for Bangla, by [16] for Dutch and by [2] for English.

Brill's tagger relies on a *transformation-based approach*, which combines a rule-based approach and statistical methods. In short, it picks the most likely tag based on a training corpus and then applies a certain set of rules to see whether the tag should be changed to anything else. Then, it saves any new rules that it has learnt in the process, for future use. In this way, Brill's tagger tries to transform an initial bad tagging into a better one in an iterative fashion.

As for stochastic taggers based on HMM (Hidden Markov Models), the training set is used to compute a statistical model that, given a word sequence, chooses the tag sequence with maximum probability.

The tagger implementation we use, called *HunPos*, is based on second-order Markov model, and the output probability is based on the previous tag in addition to the current tag. It also includes a suffix guessing algorithm to deal with unknown words.

As reported by [2], the performance of the HunPos tagger on WSJ data, measured by its error-rate, proved to be much better than that of the Brill's tagger if a small training-set is used. For large training-sets of 100,000 sentences the performances seem to be about the same, with the former tagger edging out ahead on the larger tagsets and the Brill's tagger edging ahead on the small tagset. However, the advantage of Brill's tagger is that it is easier for the user to manually correct the automatically induced knowledge of the tagger.

While the documents used for training are a collection of texts coming from the different sources reported in Section 1, the test set includes 371 sentences (10,493 tokens) translated from scientific articles in the domain of biology. In this way, we make the test more challenging because we introduce also a domain shift.

Since our main goal is to understand which of the two taggers performs better in order to integrate it into the machine translation system in a future step, we focus our evaluation on the tokens which got a different annotation from the two taggers. Results are reported in Table 2.

N. of tokens with different annotations	2052
N. of correct labels assigned by HunPos	1517
N. of correct labels assigned by Brill	365
N. of wrong annotations by both taggers	170

Table 2. Evaluation of taggers performance

Our evaluation confirms the results obtained for English [2]: also for Venetan the *HunPos* tagger performs remarkably better than the *Brill's tagger*. In particular, 74% of the diverging tags are correctly labeled by *HunPos* while only

18% of them are correct assignments by the *Brill's tagger*. The latter assigns the N label, which is the most frequent one in the training set, to unknown cases, while in *HunPos* the guesser seems to work particularly well, detecting also many foreign words, proper names, etc. Most of the cases in which both taggers fail concerns the classification of clitics, which are often homograph of articles, prepositions and pronouns, and can occur in different positions inside the sentence. Therefore, their recognition is one of the main challenging tasks of Venetan tagging.

We also perform a standard ten-fold cross-validation in order to assess the overall performance of *HunPos* on in-domain data. The final accuracy amounts to 90%, which is below the performance of state-of-the-art taggers for other languages, but is still a promising result given that the training corpus was quite small and it was enriched with automatically-generated pseudo sentences.

5 Venetan to English Translation

Problems related to Venetan translation into English and viceversa are very close to those encountered when translating from/into Italian. The most interesting types of problems include subject clitic doubling, amalgams (prepositions + article; verb + enclitic), proper nouns preceded by articles and subjects adjoined as enclitics in interrogative sentences.

To implement our Venetan-English machine translation system we have decided to use a statistical approach [12]. Unlike rule-based approaches, statistical machine translation allows for the automatic induction of a phrase dictionary based upon sentence-aligned corpora: given these corpora, available algorithms like the one implemented in the GIZA++ package [15] are able to infer probabilities of alignments between source and target single words or phrases (where the term *phrase* indicates merely a sequence of words and has no linguistic connotation whatsoever) and build a so-called *translation model*.

The probabilities contained in these phrase-tables for each entry are not the only factor affecting the overall probability of a potential translation over another. There is also a measure of how natural a potential target sentence is, approximated via n-grams probabilities extracted from a monolingual corpus in the target language: the result of this procedure constitutes the so called *language model*. Finally, a *reordering model* accounts for word displacement phenomena. These probabilities contribute to the overall probability of a certain target sentence being the translation of a given source one according to automatically computed weights. The toolkit we have used to implement our system is the Moses open source toolkit [14].

5.1 Reasons for the choice of a statistical approach

Going for a statistical approach to machine translation allows for the possibility to automatically learn the bilingual dictionary by training the translation model. Aside from the fact that coding all the rules by hand in a rule-based approach

would prove much longer a task than automatically inferring a translation model, the real problem is that to actually code all the possible rules, we should make available a sufficiently large corpus of the source language which should undergo analysis first, and in the case of Venetan, only thinking of morphologically analyzing the plethora of irregular words turns out to be an extremely complicate matter. Venetan allows for the cliticization of subject pronouns, a feature rare in Romance languages. The problem arises when the stem to which the subject pronoun is attached, i.e. the verb, undergoes changes which have not been thoroughly studied yet and for which there does not appear to be a constant rule. Here is an example[8]:

1-vuto 2-magnar 3-con 4-mi 5-?
1-do you want 2-to eat 3-with 4-me 5-?

In the Venetan, sentence the token 'vuto' should be morphologically analyzed as verb stem 'vu-' and subject pronoun clitic '-to', as opposed to its non contracted form 'ti vol'. Now coding generalized rules to transform 'vu-' into 'vol' (or vice versa) proved to be a hard task because other verbs behave in a different way, for instance 'gheto' for 'ti ga ' in the following sentence:

1-'sa 2-gheto 3-da dir 4-?
1-what 2-do you have 3-to say 4-?

By looking at these examples it becomes clear that given our current knowledge of Venetan, the best way to deal with such phenomena is that of a direct mapping between the full-fledged verb form and its contracted stem version via ad hoc rules (i.e. a dictionary lookup pass), but again, that defeats the purpose of manually coding the dictionaries. So we decided it would be easier to just rely on alignment algorithms and go with the statistical approach. Note that the English translation of both examples above is what our system currently outputs, which is indeed the correct translation.

5.2 Issues encountered in using the statistical approach for an under-resourced language

Statistical machine translation, being inherently a data-driven approach, works well when there is lots of data. Given that our corpus is so small (i.e. 13,000 parallel sentences with 128,000 words), we have encountered some problems. As far as parameter tuning is concerned, for example, what we have found out is that the language model weight has to be lowered from the default provided by the toolkit to optimize our results. Even in cases where correct alignments are actually inferred during the translation model training phase, there still can arise problems. Namely, the probabilities assigned to correct alignments are not generally very conclusive, and therefore the decisive discriminant for the choice

[8] In all examples, spacing and numbering render graphically segmentation of the source sentence into source phrases and selection of best target phrase for each source one as performed by the decoding algorithm during translation

of a translation over another turns out to be the language model; this can hurt translation quality in some cases, an example of which is reported below.

1-clifford 2-xe ndà 3-via
1-go 2-went 3-away

The source sentence contains a proper noun ('clifford'), the frequency of which is probably too low in the Europarl corpus [11] used to the train the language model with the toolkit by [17]. If the language model can actually make a difference, it will steer the decision of translation towards the token most frequently found in that position, and since 'clifford' does not appear often enough, it will consider even bad sequences of known tokens a better choice. In formal terms, there certainly existed a trigram such as '<null> <null> go' in the language model, where <null> in the n-gram world is the null token that is inserted n-1 times at the beginning of a sentence to evaluate the probability for that very sentence of beginning with some other non-null token.By lowering the language model weight, the proper noun (which indeed appeared several times in the training corpora used for the translation model) finally makes it to the target sentence, but the consequences of diminishing the language model's impact are reflected in crunchier overall translations, as the following one. Note that the correct English translation would be 'clifford went away'.

1-clifford 2-xe 3-ndà 4-via .
1-clifford 2-is 3-to go to 4-away .

Increasing the frequency of proper nouns like 'clifford' in our n-grams by collecting ad-hoc data is not a proper solution, as it would create a language model that is less representative of the reality we are trying to model.

5.3 Unfactored vs. factored models

Another advantage of state-of-the-art statistical methods nowadays is that linguistic knowledge can easily be integrated to enhance translation quality. More specifically, factored translation models [13] allow for the representation of a token in the sentence-aligned corpora as a vector of factors and for the learning of mappings between phrases of one (or more) particular source token factors to phrases of one (or more) particular target token factors, that is to say a training phase in which the alignment occurs between phrases of (sub)vectors of factors. In order to have a preliminary idea of the impact of factored translation models on Venetan, we did some initial tests with manually or semi-automatically annotated tags. Once the pipeline is consolidated, we will also integrate the *HunPos* tagger presented in Section 4.

The main reason for resorting to factored translation models in our case has been the possibility of learning alignments between phrases of source vectors made up of a wordform plus its relevant tag and phrases of target wordforms. In other words, our setup implements a single translation step in which the input factors are surface form and part of speech in Venetan, and the output factor

is surface form in English. In this way we have been able to disambiguate the plethora of Venetan homographs, as in the sentence below, where '|' is the factor separator. The correct English translation would be 'His/her father is good'.

1-so|poss pare|nh 2-xe|vex 3-bon|ag 4-.|punt
1-his father make 2-is 3-good 4-.

This is still not perfect (the word "make" has been wrongly added) but surely better than the unfactored version below.

1-so 2-pare 3-xe 4-bon .
1-i 2-think 3-it is 4-good .

The homograph 'pare' actually means both 'father' and 'seems', and the unfactored version favours the verb meaning, whereas the source sentence clearly means 'his father is good'. The factored model gets the meaning part better but shows evidence that the word alignment algorithm failed in segmenting phrases in the target language corpus at translation model training time.

5.4 Other shortcomings of the statistical approach for under-resourced languages

Even if a given vector appears in several sentences of the source language corpus, if it is not repeatedly translated into the same word in the aligned target corpus sentences, the algorithm cannot infer an alignment that spans only that single source vector. So what it really does is to singly align only the words that appear more often, and consider the other intervening words in the sentence as one big phrase that gets aligned in its entirety to what remains of the target sentence. Therefore, if a sentence containing one of those problematic words is inputted for decoding, it can be correctly translated only if that word is part of a sequence that is identical to the one to which it belonged in the training set, otherwise the problematic word will simply pass onto the output untranslated. Below you can see examples of this behaviour. Note that the English output is correct.

1-se|cosu 2-no|neg no|neg podaria|vsup nar|vi pi|q coi|part pàtini|n
1-if 2-i couldn't skate anymore

This is a sentence taken directly from the original corpus. A correct segmentation here should have separated 'se|cosu no|neg' from the rest of the source sentence and should have mapped its translation to 'then', which is what indeed appears in the original target sentence instead, rather than 'if'. Yet 'if' is a correct translation of 'se|cosu' by itself, and since this source-target pair has been noticed by the learning algorithm many times in the sentences of the original sentence-aligned corpora, a potential mapping for it has been established with a degree of confidence high enough to authorize its separate translation at decoding time. The rest of the sentence, on the other hand, is translated as a single phrase, and exactly the way it was found in the target corpus sentence during

training time. To stress this point, we show how in the example below, the vector 'pàtini|n' is not translated anymore just because we interrupt the long phrase found in the original corpus with a ',|punt'.

1-se|cosu	no|neg	2-,|punt	3-no|neg	4-podaria|vsup	5-nar|vi	6-pi|q	7-coi|part	8-pàtini|n
1-if not		2-,	3-i	4-could	5-go	6-more	7-with	8-pàtini

Finally, we would like to point out that there are no reordering problems with our translations.

6 Conclusions and future work

In this paper we have presented some issues related to the development of a Venetan to English machine translation system. Since Venetan is a diasystem, many challenges have to be tackled while creating NLP tools, from the poor resources available to orthographic normalization.

While detailing the applications implemented so far, we have suggested some solutions to the above mentioned problems. We have further described the tools under development, including a PoS tagger and a statistical machine translation system. Since we do not have large enough Venetan-English corpora to overcome wrong or missing alignment problems, and resources are generally scarce for Venetan, we will try to exploit Italian also. We will not use Italian as a pivot language within the statistical system as that would not really solve the problem, since Italian-Venetan language resources are as scarce as English-Venetan ones and therefore using the pivot language would only worsen translation results. What we will do instead is to transform an Italian text contained in a large En-Ita parallel corpus into Venetan text with a rule-based approach. The idea is that manually coding rules to translate from Italian into Venetan is still less expensive than coding rules to transform Venetan into English, as Venetan is in fact a dialect of Italian, with which it shares a lot of grammatical rules and of lexical entries. In addition, dialects lack native words for a number of lexical domains, like for instance, bureaucratic domain, scientific domain etc, which is where we will look into next.

Finally, it must be said that we have carried out the pre-processing of most of our language resources in a semi-automatic way with the support of human translators, annotators, editors, etc., and that we have used the whole amount of the resulting parallel corpus for training our system. The lack of a numerical evaluation of its performance stems from the problems we have encountered in automatically normalizing, categorising, and tagging existing (out of domain) Venetan texts, such as those contained in the Venetan version of Wikipedia, which we would like to use for the evaluation of our translation system. This is another issue we are working on.

References

1. Bird, S., Loper, E.: NLTK: The Natural Language Toolkit. In: Proceedings of the ACL demonstration session. pp. 214–217. Barcelona, Spain (2004)
2. Block, S.: A Comparison of three part-of-speech taggers. Master's thesis, Uppsala Universitet (2009)
3. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000. Seattle, WA (2000)
4. Brill, E.: Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. In: Proceedings of the Workshop on Natural Language Processing Using very Large Corpora. Boston, MA, USA (1997)
5. Carl, M., Melero, M., Badia, T., Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., Yannoutsou, O.: METIS-II: Low resource machine translation. Machine Translation 22(1-2) (2008)
6. Cortelazzo, M.: Noi Veneti - Viaggi nella storia e nella cultura veneta... Regione Veneto (2001)
7. Delmonte, R., Bristot, A., Tonelli, S., Pianta, E.: English / Veneto Resource Poor Machine Translation with STILVEN. In: BULAG 33 - International Symposium on Data Sense Mining, Machine Translation and Controlled Languages, and their application to emergencies and safety critical domains (2009)
8. Giunta Regionale del Veneto: Manuale di Grafia Veneta (1995), available online at http://www.veneto.org/gvu/
9. Halácsy, P., Kornai, A., Oravecz, C.: HunPos - an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions. pp. 209–212. Prague, Czech Republic (2007)
10. Hasan, F., UzZaman, N., Khan, M.: Comparison of different POS Tagging Techniques (n-gram, HMM and Brill's tagger) for Bangla. Advances and Innovations in Systems, Computing Sciences and Software Engineering pp. 121–126 (2007)
11. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of MT Summit (2005)
12. Koehn, P.: Statistical Machine Translations. Cambridge University Press (2010)
13. Koehn, P., Hoang, H.: Factored Translation Models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Prague, Czech Republic (2007)
14. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic (2007)
15. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19–51 (2003)
16. Stehouwer, J.H.: Comparing a TBL Tagger with an HMM Tagger: Time Efficiency, Accuracy, Unknown Words (2006), internship report
17. Stolcke, A.: SRLIM - An Extensible Language Modeling Toolkit. In: Proceedings of the International Conference on Spoken Language Processing. Denver, USA (2002)
18. Tonelli, S., Pianta, E., Delmonte, R., Brunelli, M.: VenPro: A Morphological Analyzer for Venetan. In: Proceedings of the 7^{th} International Conference on Language Resources and Evaluation. Valletta, Malta (2010)

Looking for the best Evaluation Method for Interlingua-based Spoken Language Translation in the medical Domain

Marianne Starlander and Paula Estrella

University of Geneva, ETI-TIM-ISSCO, 40 Bd du Pont d'Arve, 1211 Genève 4
Marianne.Starlander@unige.ch
University of Córdoba, FaMAF, Haya de la Torre s/n, Ciudad Universitaria, 5000 Córdoba
pestrella@famaf.unc.edu.ar

Abstract. This paper focuses on the quality of rule-based machine translations collected using our open-source limited-domain medical spoken language translator (SLT) tested at the Dallas Children's hospital. Our aim is to find the best suited metrics for our Interlingua rule based machine translation (RBMT) system. We applied both human metrics and a set of well known automatic metrics (BLEU, WER and TER) to a corpus of translations produced by our system during a controlled experiment. We also compared the scores obtained for both type of evaluation with those obtained on translations produced by the well known statistical machine translation (SMT) system GoogleTranslate[1] in order to have a point of comparison. Our aim is to find the best suited metric for our type of Interlingua RBMT SLT system.

Keywords. Key words: Machine translation evaluation, Spoken language translation, Automatic metrics

1 Introduction

MedSLT is a medium-vocabulary open-source speech translation system intended to support medical diagnosis dialogues between a physician and a patient who do not share a common language [1]. The translation module is rule-based in order to provide a more predictable translation, prioritizing precision over recall given the safety-critical nature of the task. This implies that we prefer to produce no translation at all instead of bad translations that would account for recall, but that could entail communication errors between the physician and the patient, potentially leading to diagnosis errors.

More specifically, the translation is interlingua-based making the system multilingual, translating a dozen of language combinations to/from {ENG, FRE, JPN, SPA,

[1] http://translate.google.com/

ARA, CAT} [2]. This Interlingua architecture helps us add new languages more easily but at the cost of reaching a common representation for several languages by focusing on the meaning of the sentences to produce the corresponding common semantic representation. Therefore, the resulting translations are less literal than those produced by other models such as example-based or statistical models. This approach avoids problems of divergences and discrepancies that would inevitably arise between the large varieties of language families handled.

In order to assess our system, we have carried out a set of evaluations using different methodologies in the quest for the most appropriate one for our system [3-4]. We have applied a set of state-of-the-art metrics, including human-based and automatic ones. However, as it has often been mentioned in the MT literature [5], automatic metrics based on computing the similarity of an output against one or more references (like BLEU, WER, and most of the commonly used metrics) seem to be less suited for rule-based machine translation (RBMT) systems, given that they tend to reward translations that are more literal and close to a given reference. Thus, our Interlingua architecture seems at first sight incompatible with this type of automatic metrics for the evaluation of its translation quality. In this paper we want to further study the suitability of these automatic metrics compared to tailor-made human metrics and we will therefore apply a set of metrics to a corpus gathered during an experiment, where we tested the medical spoken translation system in a controlled environment, very close to that of real use. The results of running a series of tests on the RBMT data are then compared to those obtained using statistical translations produced by GoogleTranslate (GT)[6]. We chose to use GT as a baseline, to have a point of comparison with a statistical machine translation (SMT) system, but we are aware that the comparison is slightly unfair since GT has not been particularly trained for this task. However, many bilingual resources exist on the web in the domain of medical diagnosis, so this choice is less unfair than if we had chosen a SMT system trained, for example, on the Europarl corpus [7]. Some tests have been conducted using automatically generated data to build a SMT system equivalent to MedSLT for English-French and English-Japanese, but not for English-Spanish since the resulting translation does not outperform the RBMT [8].

The rest of the paper is organised as follows: In section 2 we give more background about MedSLT explaining in more details why our system is Interlingua and RBMT based, and what this choice implies on the resulting translations. In section 3, 4 and 5 we describe the experiment conducted using different ways of evaluating MT. In section 6 we study the correlation between our human metrics and the chosen automatic metrics. Finally in section 7, we conclude on whether these metrics can be useful for evaluating RBMT in a spoken language translation context.

2 System description

In this study we are using the bidirectional English-Spanish version of MedSLT that was used during tests conducted at the Dallas Children's Hospital in 2008 [3]. This system enables an English speaking physician to communicate with his Spanish

speaking patient during a medical examination. Both speech recognition and translation are rule-based. The speech recognition (SR) component uses the Nuance 8.5 platform [9], equipped with grammar-based language models. The experiments carried out in [8] have shown that for a safety critical task such as MedSLT, statistical SR does not give better results than the RBSR, although it adds robustness to a hybrid system.

The workflow for the bidirectional version of the system is as follows [10]: the physician presses the SR button and speaks into the microphone; he can then check the back-translation of his utterance. If the physician accepts the produced string, it gets synthesised to the patient. The patient clicks on the SR button to speak a direct answer to the question; alternatively, to produce an answer he can make use of the help window that displays a set of potential answers that are covered by the system. The patient checks the produced back-translation and launches the synthesis if the sentence is correct.

In either case, the back-translation is the result of the entire processing of the input by the system. This means that at run-time, the recogniser produces a source language semantic representation which is first translated by one set of rules into an Interlingua form. Then a second set of rules simultaneously translates the representation back into: the source language (that is, a back-translation, so that the user can check if the system has correctly understood and translated the spoken sentence) and into a target language representation. Then, a target-language Regulus grammar compiled into generation form turns this representation into one or more possible surface strings.

An Interlingua-based architecture has been chosen to avoid having to multiply the number of Interlingua to target language translation rules. Instead it keeps a unique translation for all utterances that have received the same Interlingua representation [11], which is almost flat, as you can see for the following example sentence in Figure 1.

Source: Do you have a sore throat?
Interlingua: [[body_part,throat], [prep,in_loc], [pronoun,you],
[state,have_symptom], [symptom,pain], [tense,present], [utterance_type,ynq],
[voice,active]]
Backtranslation : Do you experience a pain in the throat?
Target : ¿ Le duele la garganta ?

Fig. 1. Example of Interlingua representation

The resulting representation is an unordered list of semantic elements. The attributes are derived from the canonical English form, removing most of the grammatical information. The advantage of such a representation is its simplicity. It enables us to easily write translation rules that are expressive enough to convey the nuances in the concepts used in specific domains. We always try to keep the most idiomatic translation. For example, we chose to translate the source sentence “Do you have a sore throat” by “Le duele la garganta” (closer to “Does your throat hurt”) instead of the more literal “Tiene un dolor de garganta.” As a consequence, the translations pro-

duced are clearly freer and more coherent than translation produced by direct linguistic MT or statistical translation as you can see in **Table 1** below.

The use of this Interlingua avoids surface divergences in order to keep only the meaning of the sentences. As a consequence certain losses in style sometimes occur but most of the time the lost information is not important for the purpose of SLT in the medical domain.

Source sentences	**Target sentences by GoogleTranslate**
Did the doctor do a strep test?	Dijo el médico haga una prueba para estreptococo
Did the doctor run a strep test?	Dijo el médico realizar una prueba de estreptococos
Did they do a strep test?	Hicieron una prueba de estreptococos

Table 1. Source sentences collected during the tests all producing the same translation "le han realizado la prueba rápida por estreptococo?" with our system.

As mentioned before, the main goal is to achieve coherence and reliability so that a physician can communicate efficiently and without danger with his patient, which explains why the output of the system is often more idiomatic and thus freely translated.

3 Evaluation of RBMT vs. SMT output

Given the non-literal nature of our translations, we believe that the classical automatic metrics that are based on the resemblance of a MT with one or more references would give low scores on our RBMT system and higher scores for the SMT system that produces more literal and similarly long translations as the original. Hence, purely reference oriented metrics as BLEU [12] and WER should prove less suited for our system. In evaluation campaigns such as WMT09 [13], there is no real comparison between RBMT and SMT. This is why we specifically want to compare the results obtained for RBMT and SMT outputs using the same metrics. Our main objective remains to find more appropriate metrics for RBMT and especially Interlingua based RBMT similar to our system.

According to [14], BLEU shows a favorable bias towards SMT, so we would like to verify this claim. In recent studies some new "less literal" metrics have emerged, such as translation edit rate (TER) [15-16] and METEOR [17]. In [16], TER is described as reaching a higher correlation with human judgments because it assigns lower costs to phrasal shifts than BLEU, which implies that it might be better suited for RBMT than the classical n-gram metrics. However, our corpus is quite different from classical written MT, because our sentences are very short and often syntactically quite remote from a literal translation as mentioned in section 2, we have thus decided to run both types of metrics on our test corpus. As a comparison point, we are going to study the relation between our tailor-made human metrics and more classical

human metrics, and will also analyse their correlation with the chosen automatic metrics, namely BLEU, WER and TER.

We will now explain the experimental framework by describing the data collection. Then we will give a detailed description of the human and automatic metrics applied.

4 Data collection

The data we are using in this experiment has been collected during a test-phase in 2008, where our aim was mainly to compare two versions of the system [3]. We had organised a data collection with English speaking physicians and Spanish speaking standardised patients at the Dallas Children's Hospital. The aim of the task was to determine whether the patient suffered from a bacterial infection (strep throat) or not. Eight physicians and 16 patients participated. The patients were acted by native-Spanish in-house interpreters of the Dallas Children's Hospital. We asked the patients to simulate viral sore throat or strep throat symptoms, described in eight different fixed scenarios. None of the participants had used the system before. Our test corpus for this study consists of 222 English to Spanish translated diagnosis questions from our Dallas data collection.

5 Human evaluation

In our research, we wanted to focus on the end usage of the produced translations and get away from linguistic issues. In our particular case, what is most important is that the message comes across and this is why the scale chosen focuses on the meaning, in a specific context of use: communication between a doctor and his patient while asking diagnosis questions. As suggested in [18], we are aiming at a metric directly related to the final use of the produced translation rather than using the classical metrics that are commonly applied to evaluate the degree of adequacy and fluency of a translation.

5.1 Scale description

Our scale is focused on evaluating if the produced translations are useful for our task or if they could be dangerous. Therefore, this evaluation scale tried to leave purely linguistic aspects on the side, that is, instead of judging the syntactic or linguistic aspects of the translations, the evaluator's task consisted on indicating whether the message from a patient was correctly sent to the doctor. For this purpose, the 4-point scale chosen relates the meaning of a sentence to its potential to create misunderstandings or false communication between a doctor and his patient. The scale is described as follows:

- CCOR (4): The translation is completely correct. All the meaning from the source is present in the target sentence.

- MEAN (3): The translation is not completely correct. The meaning is slightly different but it represents no danger of miscommunication between doctor and patient.
- NONS (2): This translation doesn't make any sense, it is gibberish. This translation is not correct in the target language.
- DANG (1): This translation is incorrect and the meaning in the target and source are very different. It is a false sense, dangerous for communication between doctor and patient.

In the evaluation form sent to the judges we included the description of the scale and we provided them with the following examples, by way of tutorial on how to proceed with the evaluation:

Source	**Target**	**Score**
Do you experience pain?	Le duele ? (*Does it hurt)*	MEAN
Do you have a headache	Tiene tos ayer *(Do you have a cough yesterday)*	NONS
Are you having fever?	¿El dolor está aliviado cuando tiene fiebre? *(The pain is decreasing when you have fever)*	DANG

Table 2. Evaluation examples for annotators

As mentioned before, this scale is clearly focused on meaning and you could thus wonder why some trace of grammar and style remains present in the category CCOR: this is only to reflect the difference between sentences that are clearly correct in all aspect and sentences that are slightly different but have most of the meaning present. One of the typical examples for the MEAN category is the following sentence, where the meaning is similar although the sentences are syntactically distant: "*do you experience pain*" vs. "*does it hurt*" for the Spanish sentence "*le duele*".

The order can also appear as surprising since a nonsense sentence (NONS) receives a higher score (2) than a DANG sentences (1). This can simply be explained by the fact that in the context of a medical dialog, a nonsense sentence, that clearly appears as such is more easily recognised and rejected than a sentence that "looks" correct but the meaning is in fact totally different (for example: false negative sentence). This kind of sentences could produce serious diagnosis errors. The main aim of this scale is to encourage the evaluators to forget about linguistic differences and focus on the meaning. But as we will see in section 4.2, where we compare the results of the evaluation by translators and non-translators, we noticed that this is clearly difficult for translators as they continue to rate more severely than non-linguists. While conducting previous studies we also noticed the impact of the attitude towards machine translation and technology in general on the severity of the evaluation [19]. We had at that time already noticed how difficult it is for "classical" translators to take a certain distance with grammar and style issues in order to focus solely on the meaning, compared to the results of non-translators on the same task.

We thus asked two groups to evaluate the output of our system. The group of translators is composed by a sub-set of the Spanish language Interpreters of the Dallas Children's Hospital who had participated in the data collection and by a number of

professional English-Spanish translators. The second group is composed by non-translators, with a pro-technology background, since most of them happen to be Spanish speaking computer scientists. We asked each group to evaluate a set of 222 sentences translated by our system and by GoogleTranslate applying our human metric.

We will first present the results for our human metrics and then we will pass on to the automatic metrics before studying the correlation between them.

5.2 Results

Table 3 below shows that the average for both types of systems is quite close. The scores are only slightly higher for the RBMT system when it is evaluated by non-translators. The difference between non-translators and translators is clearer in favour of the RBMT. But as a whole the difference between the two systems is not significant if we consider only the averages.

	RBMT	SMT
Translators	3.40	3.43
Non-translators	3.62	3.46
All	3.51	3.44

Table 3. Average using our scale (4=highest, 1=lowest).

In order to get a better idea of the actual quality of each system in Table 4 we show the percentage of each category of the scale, in a majority wins perspective rather than by calculating the average score as in Table 3.

Cat	RBMT Trans.	SMT Trans.	RBMT Non-trans	SMT Non-trans
1=DANG	4.5%	3.2%	3.2%	2.7%
2=NONS	2.3%	6.3%	0.0%	2.3%
3=MEAN	25.2%	20.7%	16.7%	15.3%
4=CCOR	68.0%	65.8%	76.1%	70.7%
No Agreement	N/A	4.1%	4.05%	9.0%

Table 4. Translation quality by category, by majority wins

In this table we can see that in fact our RBMT obtains better results, again especially with our group of non-translators since they evaluated 76.1% of sentences produced as totally correct, compared to only 70.7% for the SMT system. When using human metrics the problem of agreement between judges always arises, so we decided to calculate the inter-rater agreement using the *AgreeStat Excel VBA program* [20].

Kappa estimate	RBMT	SMT
Translators	0.1758	0.3698
Non-translators	0.0973	0.2591

Table 5. Kappa estimate for our 4-point scale

Table 5 shows that our Kappa estimate is particularly low for the RBMT system. This can quite simply be explained by the fact that this scale is more difficult to apply

consistently especially without previous training on how to interpret the scale. However, since these Kappa estimates remain quite difficult to interpret, we decided to follow [21]' and to calculate the percentage of total agreement between judges, that is the number of times all three judges agreed on the choice of our 4-point scale. As you can see in Table 6, the overall percentage of both categories of evaluators is very low especially for our RMBT system. For the RBMT it is interesting to note the difference between the translators and the non-translators; the latter group is more coherent, while we get the reversed trend for the SMT. It is very interesting to see that the non-translators get a much higher agreement for the RBMT than the translators.

	RBMT	SMT
All 6 evaluators	18.5%	27.9%
Translators (3)	33.8%	49.5%
Non-translators (3)	41.9%	46.4%

Table 6. Agreement between evaluators

The question that arises at this point is if our tailor-made metric has removed all fluency and linguistic differences in quality, leaving us with two quite different output sets that get almost equal results. In order to further study this observation we decided to conduct an extra study using a more classical human metric, namely a ranking evaluation.

The second human evaluation task clearly shows that the output by our RBMT is preferred in 61.1% cases to the output produced by the SMT (34.5%). The Kappa estimate for this task is of 0.5564 which is much higher than the results obtained for the 4-point scale displayed in Table 5. The reason for this is probably that the ranking scale is easier to apply and gives less variation possibilities.

We will now explore the possibility of using automatic metrics in order to finally achieve objective MT evaluation suitable for RBMT.

6 Automatic metrics

As mentioned in section 3, we chose to evaluate standard classic automatic metrics such as Word Error Rate (WER) and BLEU [12] compared to newer metrics like the Translation Edit Rate (TER) [15] computes the number of edits needed to change the output so that it semantically corresponds with a correct translation. Although another potentially suitable automatic metric is METEOR [17], we have not run it in this experiment because we are lacking the Spanish language resources needed by this metric; this is clearly one disadvantage of this metric preventing its wide use in evaluation.

6.1 Resource description

Since the above cited automatic metrics are very dependent on the reference, we have run the tests with three different reference sets for our 222 source sentences: (1)

three human translations provided by the interpreters of the Dallas Children's Hospital themselves and completed by translations produced by professional English-Spanish translators, (2) a set of translation used as corpus reference for our system and (3) a mix of the two first sets of reference translations in order to provide both more literal human translations and translations that we as developer aimed at in our Interlingua perspective.We are well aware that the corpus is quite small but this is due to the cost of creating such a pool of human references.

We will now analyse the results obtained for the automatic metrics before studying their respective correlation to the human metrics described in the previous section.

6.2 Results

As you can see in Table 7, the average obtained for all sentences are quite similar for both types of systems when we use human translations only (columns 4 and 5) and with an equal number of translations from translators and the developer's corpus (columns 6 and 7). There only appears a clear difference in favour of the RBMT for all metrics if you use as only reference the developer's corpus (columns 2 and 3). The latter result is coherent with our second human evaluation task.

Metrics	RBMT-ref_dev	SMT -ref_dev	RBMT-ref_trans	SMT-ref_trans	RBMT-ref_all	SMT-ref_all
BLEU	0.84	0.17	0.35	0.33	0.35	0.33
WER	0.12	0.80	0.59	0.67	0.55	0.58
TER	0.10	0.67	0.53	0.65	0.65	0.66

Table 7. Result for the automatic evaluation

The RBMT corpus has served as reference before in a similar evaluation task using BLEU [22], but we think it is fairer to use a mix of the two types of references (columns 6 and 7). These results point out the importance of the choice of references, since they are totally different according to the translation references used. In order to have a better grasp of why the results are so close in columns 4-6, we decided to check the results by applying the metrics at the sentence level.

Source	Target	Bleu4	Hum.	TER	WER	bleu2
Are you coughing?	¿Tiene tos?	0	4.0	0.0	0.0	1
Do you have a cough?	¿Tiene tos?	0	4.0	0.0	0.0	1
Do you have a fever?	¿Tiene fiebre?	0	4.0	60.0	75.0	0
Have you vomited?	¿Ha vomitado?	0	4.0	1	1	1

Table 8. Sentences level evaluation sample

This analysis makes immediately clear that the overall scores cannot be used as such, as you can see in the sample provided in Table 8. To illustrate this, Table 8 shows the scores for Bleu (4-grams), human evaluation, TER, WER and Bleu (2-grams) for sentences 19, 48, 50 and 145 of our corpus. As we mentioned in our system description, our sentences are very short and actually 20% of sentences (46/222)

are shorter than 4 words for RBMT and 14% for SMT (32/222), which explains how the scores for these sentences using the classic BLEU based on 4-grams does not suit well for our test corpus. For almost 10% of our sentences, we get a score of 0 while our human evaluators rated with a 4.

Although they carry the same content, our translation is often quite distant from the original syntax, as shown in the following example. Our translation for "*Do you have a rash*" is "*Tiene una erupción cutánea*" but all human references contain a more regional variation as "*Tiene sarpullido*" or the more familiar variation "*Tiene un picor*" or "*Tiene urticaria*". Another example of this kind is our translation for "*What are you allergic to?*" which is "*Qué le da alergias?*". This solution has been adopted in order to avoid ambiguities in the gender (e.g. *alérgica/alérgico*) that our reference translators did not take into account: "*A qué es alérgico?*". In those cases, only a semantic metric, rich in synonyms and regionalisms could detect that these sentences are equivalent, even if on the n-gram side there is almost no resemblance. These two observations explain how the BLEU score in Table 8 are artificially drawn down for our system.

In order to find the fairest metric for our task, we calculate the correlation with the human evaluation on a sentence basis and added scores for BLEU2 and BLEU3.

Correlation type	RBMT	SMT
bleu vs H	0.127	0.264
bleu-3 vs H	0.205	0.290
bleu-2 vs H	**0.331**	0.223
Ter vs H	**-0.304**	-0.208
wer vs H	**-0.487**	-0.262

Table 9. Correlation between automatic and human metrics on segment level

Table 9 shows that the highest correlation occurs for BLEU-2 and WER and not for TER compared to our initial hypothesis. It is interesting to note that the correlations for the SMT are much lower than those for RBMT and that BLEU (3-gram and 4-gram) correlates better with humans in the case of the SMT. Finally, these results show that TER is not behaving so differently from the classic n-gram metrics and that a possible set of metrics to apply in future evaluations could be our human metrics plus BLEU-2 and WER.

7 Conclusion

The aim of this paper was to find the best suited metrics to evaluate the output of our RBMT system. One of our findings is that the automatic metrics we used did not show a bias in favour of SMT; in fact, correlations are lower for the SMT and also the automatic scores, depending on the set of references used. However, they did not prove adequate either. The results obtained in section 5 prove that, given the nature of our corpus, we still need to find a better metric. It turns out that in our context, the

classical BLEU based on 4-grams is not suited at all and should be replaced by BLEU based on bi-grams. It would have been interesting to apply other metrics, such as METEOR, in order to explore other aspects of our translations but, as mentioned before, we need the necessary resources for the language under evaluation, Spanish in this case. According to the study described in [13], the best correlation with human evaluation is acheived with UPC [23], which is a combination of many metrics commonly used but the interesting idea is that the authors aim at not only assess one facet of MT quality, which is in most cases the lexical ressemblance but to try to englobe syntactic and semantic aspects. This is the direction we need to take, because what we aim at is a metric that assesses the quality of MT through the semantic equivalence to the reference translation, which is what the authors of [24] propose to do using recognition of textual entailment (RTE). This kind of metric that really includes semantics in its assessment of quality would probably obtain better results on our RBMT system. Ideally, we should use metrics calculating the resemblances of the output not on a surface level but more deeply on the semantic representation inspired from [22] but the drawback of such a metric would be that it can not be generalized to other systems' output.

8 References

1. Bouillon, P., Rayner, M., Chatzichrisafis, N., Hockey, B.A., Santaholma, M., Starlander, M., Isahara, H., Kanzaki, K., Nakao, Y.: A generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation. In: Tenth Conference of the European Association of Machine Translation, pp.50-58. Budapest, Hungary (2005).
2. Bouillon, P., Halimi, S., Nakao, Y., Kanzaki, K., Isahara, H., Tsourakis, N., Starlander, M., Hockey, B.A., Rayner, M.: Developing Non-European Translation Pairs in a Medium-Vocabulary Medical Speech Translation System. In: 6th International Conference on Language Resources and Evaluation, pp. 1741-1748. Marrakech, Morocco (2008).
3. Starlander, M., Bouillon, P., Flores, G., Rayner, M., Tsourakis, N.: Comparing two different bidirectional versions of the limited domain medical spoken language translator MedSLT. In: 12th annual conference of the European Association for Machine Translation, pp. 174-179. Hamburg, Germany (2008).
4. Starlander, M., Estrella, P.: Relating recognition and translation quality with usability of two different versions of MedSLT. In: Machine Translation Summit XII, pp.324-331. Ottawa, Ontario, Canada (2009).
5. Callison-Burch, C., Osborne, M.: Re-evaluating the role of BLEU in machine translation research. In: 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 249-256. Trento, Italy (2006).
6. Google translate site, http://translate.google.com/
7. European Parliament Proceedings Parallel Corpus site, http://www.statmt.org/europarl
8. Rayner, M., Estrella, P., Bouillon, P.: Bootstrapping A Statistical Speech Translator From A Rule-Based One. In: Second Workshop on Free/Open-Source Rule-Based Machine Translation, pp.21-28. Barcelona, Spain (2011).
9. Nuance Communications: Nuance Grammar Developer's guide, version 8.5, Menlo Park, CA, USA (2003).
10. Bouillon, P., Flores, G., Starlander, M., Chatzichrisafis, N., Santaholma, M., Tsourakis, N., Rayner, M., Hockey, B.A.: A Bidirectional Grammar-Based Medical Speech Transla-

tor. In: Workshop on Grammar-based approaches to spoken language processing. ACL 2007, pp. 41-48. Prague, Czech Republic (2007).
11. Bouillon P., Rayner M., Novellas Vall, B., Starlander, M., Santaholma, M., Nakao, Y., Chatzichrisafis, N.: Une grammaire partagée multi-tâche pour le traitement de la parole : application aux langues romanes. In: TAL (Traitement Automatique des Langues), vol. 47, no. 3, pp. 155-173. Hermes and Lavoisier, Paris, France (2007).
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318. Philadelphia, USA (2002).
13. Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J.: Findings of the 2009 Workshop on Statistical Machine Translation. In: Fourth Workshop on Statistical Machine Translation, pp.1-28. Athens, Greece (2009).
14. Hartley, A., Popescu-Belis, A.: Évaluation des systèmes de traduction automatique. In: Chaudiron, S. (ed.) Évaluation des systèmes de traitement de l'information, pp. 311-335. Hermès, Paris, France (2004).
15. Snover, M., Dorr, B. J., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: 7th Conference of the Association for Machine Translation in the Americas, pp. 223-231. Cambridge, Massachusetts, USA (2006).
16. Snover, M., Madnani, N., Dorr, B. J., Schwartz, R.: Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In: Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics, pp. 259-268. Athens, Greece (2009).
17. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL-2005: Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65-72. University of Michigan, Ann Arbor, 2005.
18. Boitet, C., Bey, Y., Tomokio, M., Cao, W., Blanchon, H.: IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations. In: International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation, pp. 23-30. Kyoto, Japan (2006).
19. Rayner, M., Bouillon, P., Chatzichrisafis N., Santaholma, M., Starlander, M.: MedSLT: A Limited-Domain Unidirectional Grammar-Based Medical Speech Translator. In: First International Workshop on Medical Speech Translation, HLT-NAACL, pp. 44-47. Omnipress Inc. New-York, USA (2006).
20. AgreeStat Excel VBA program site , http://www.agreestat.com/agreestat.html
21. Hamon, O., Fügen, C., Mostefa, D., Arranz, V., Kolss, M., Waibel, A., Choukri, K.: End-to-end evaluation in simultaneous translation. In: 12th Conference of the European Chapter of the ACL, pp. 345-353. Athens, Greece (2009).
22. Rayner, M., Estrella, P., Bouillon, P., Halimi, S.: Using Artificial Data to Compare the Difficulty of Using Statistical Machine Translation in Different Language-Pairs. In: Machine Translation Summit XII, pp. 300-307. Ottawa, Ontario, Canada, (2009).
23. Giménez, J., Màrquez, L.: The UPC Participation at the Metrics MATR Challenge 2008. In: Metrics MATR Workshop at AMTA'08 Machine Translation, Waikiki, Hawai'I (2008).
24. Padó, S., Cer, D., Galley, M., Jurafsky, D., Manning, C.D.: Measuring machine translation quality as semantic equivalence: A metric based on entailment features. In: Machine Translation 23, 2-3, September, pp. 181-193. Kluwer Academic Publishers Hingham, MA, USA (2009).

Machine Translation Evaluation for Croatian-English and English-Croatian Language Pairs

Marija Brkić[1], Sanja Seljan[2], and Maja Matetić[1]

[1] University of Rijeka, Department of Informatics, Croatia
mbrkic@uniri.hr; maja.matetic@ri.t-com.hr
[2] University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences, Croatia
sseljan@ffzg.hr

Abstract. This paper presents a bidirectional machine translation evaluation study for the Croatian-English and English-Croatian language pairs. Translations from Croatian into English have been obtained in four different domains from four online machine translation services, i.e. Google Translate, Stars21, Translation Guide and InterTran. These translations have been evaluated by three different automatic accuracy metrics, i.e. F-measure, BLEU and NIST, as well as by human evaluators. Evaluations are based on a single reference per sentence. In the reverse translation direction, Google Translate output has been analyzed in the same manner. System level correlation between F-measure, BLEU, NIST and human assessments is given and the significance of the results is discussed.

Keywords: online MT (machine translation), manual evaluation, automatic evaluation, F-measure, BLEU, NIST

1 Introduction

A large-scale experiment which measures how strongly 26 automatic metrics correlate with human assessments of translation quality for five European languages is presented in [1]. The aim of this work is to evaluate the online available machine translation (MT) services for the Croatian-English language pair and vice versa, and to see how well the selected automatic evaluation metrics, which are unforgiving for morphological errors, correlate with human assessments.

Evaluation methods can be manual or automatic. Nevertheless, both categories are extremely subjective [2]. The quality of automatic measures can only be determined by comparison to human assessments [3]. Human assessments are considered gold standard for evaluation. However, they are expensive with

respect to time and money [4]. Automatic metrics have many advantages compared to human assessments. Besides the fact that they are useful for comparing the performance of different systems on a common translation task, they are extremely useful during system development because they are fast and have low-cost [5]. The correlation between two metrics is usually computed using the Pearson correlation coefficient.

The automatic evaluation scores guide the development of the MT system based on concrete performance improvements [5]. The system is tested multiple times on one distinct set of test sentences, either for adjusting parameter settings or for examining the effects of system-design changes [6]. An example that follows illustrates this process in the field of statistical MT (SMT). The basic idea behind phrase-based SMT is to segment source sentences into phrases, translate each phrase and compose target sentences from these phrase translations. In other words, there are three components that contribute to producing the best possible translation—the phrase translation table, the reordering sub-model and the language sub-model [7]. If only lower quality bilingual data is available, the system needs to rely more on the monolingual target language sub-model, meaning that sometimes a sub-model needs to be given more weight [8]. The automatic evaluation scores guide the setting of weights.

The second section of this paper highlights desirable properties of any automatic evaluation metric and focuses on the three metrics most heavily used in MT community. In the third section a detailed description of the conducted study is given and the results are presented. Section four discusses the obtained results. The findings and directions for future work are summarized in the conclusion.

2 MT Evaluation

MT evaluation should be able to determine semantic equivalence or similarity between sentences, which makes it a hard problem. This is evident from the fact that any number of different translators translates the very same sentence differently. Besides determining semantic equivalence, desirable properties of an evaluation metric are that it is tunable, meaningful, consistent, correct, reliable, general, and has low cost [7]. A metric is tunable if system performance can be optimized towards it. If it gives intuitive interpretation of translation quality, it is meaningful. A consistent metric gives the same results by repeated usage, i.e. inter-annotator agreement. If better systems are ranked higher, then the metric is also correct [7]. MT systems that score similarly also perform similarly if the metric is reliable. Furthermore, a metric should be as sensitive as possible to differences in MT quality between different systems, and between different versions of the same system. Finally, if a metric is applicable to different MT tasks in a wide range of domains and scenarios, it has appropriate generalization power [5]. Turian et al. add reliability on shorter texts as another desirable

property. However, MT evaluation metrics are usually less reliable on shorter translations. As the most important criterion, they point out the ability to rank the systems the same way human evaluators would rank them.

2.1 Automatic Evaluation

An automatic evaluation metric should besides the quality of translation satisfy some extra requirements. It should be fast, easily integrated into the existing workflow, customizable, and its memory requirements must not go beyond memory available on the machines under consideration [7]. Before scoring with an automatic evaluation metric, the translated text and the reference translations, i.e. one or more human translations of the same sentence, are conditioned to improve the efficiency of the algorithm, e.g. case information is removed, numerical information is kept together as single words, punctuation is tokenized into separate words, and adjacent non-ASCII words are concatenated into single words [9]. All automatic evaluation metrics use one or more reference translations. These reference translations are used for comparison with the MT output or hypothesis translations. Automatic metric is considered better if it has higher degree of correlation with human assessments [7]. There are numerous automatic metrics, such as Word Error Rate (WER) [6], Position-independent Word Error Rate (PER) [4], Translation Edit Rate (TER) [3], F-measure [10], Bilingual Evaluation Understudy (BLEU) [11], NIST [9], ROUGE [12], and METEOR [5]. One of the first automatic evaluation methods applied to SMT, WER, which is borrowed from speech recognition, is based on the Levenshtein distance [7]. WER, TER and PER are error measures, while the rest of the metrics fall into the category of accuracy measures [4]. The metrics differ in the way they measure similarity. However, the hypothesis translation which is closer to reference translation is ranked better by all of the metrics [2].

The rest of this section describes the three selected fully-automatic accuracy evaluation metrics, F-measure, BLEU and NIST. Due to the limited scope of this work, the remaining metrics have been excluded from further study, and will be dealt with in the future. METEOR has been excluded due to the lack of necessary language tools.

F-measure. The measures that are typically used in evaluation are precision, recall and F-measure (1). Precision is the percentage of generated words that are actually correct. Recall stands for the percentage of words that are generated and that are actually found in the reference translation. F-measure is the harmonic mean of recall and precision [7]. It is also known as the F1-measure because precision and recall are evenly weighted.

$$F\text{-}measure = \frac{precision \times recall}{(precision + recall)/2} \quad . \tag{1}$$

BLEU. BLEU ranks MT output according to a weighted average of the number of n-gram overlaps with the reference translations [2]. It is based on the modified unigram precision, which relies on the notion that a reference word should be considered exhausted after a matching hypothesis word is identified. The total count of each hypothesis word is clipped by the maximum number of times the word appears in any of the reference translations. These clipped counts are added and divided by the total number of hypothesis words. Modified n-gram precision is computed by analogy. Modified unigram precision accounts for adequacy, while modified n-gram precision accounts for fluency. A modified precision score, p_n, for the entire corpus, is calculated as in (2).

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} count_{clip}(n\text{-}gram)}{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} count(n\text{-}gram)} \quad . \tag{2}$$

The clipped n-gram counts for all the sentences are added and divided by the number of hypothesis n-grams in the test corpus. N-gram matches are, therefore, computed at the sentence level, but the modified n-gram precision is the fraction of n-grams matched in the entire corpus. A weighted linear average of the modified precisions enables combining of the modified precisions for various n-gram sizes. However, the modified unigram precision is much larger than the modified bigram precision, etc. In order to take this exponential decay into account, a weighted average of the logarithm of the modified precisions is calculated. The brevity penalty (3) is a multiplicative factor which penalizes hypotheses shorter than their reference translations. It is computed over the entire corpus to allow some freedom at the sentence level [11]. Main critiques directed towards this metric are that it ignores the relative relevance of words, it does not address the overall grammatical coherence, the actual BLEU scores are meaningless, and human BLEU scores are barely higher than that of an MT system, although the translations are of much higher quality [7]. Furthermore, BLEU is quite unintuitive and relies upon a large number of sentences in order to correlate with human assessments [3].

$$BP = \begin{cases} 1 & \text{if } outputLength > referenceLength \\ e^{(1-\frac{referenceLength}{outputLength})} & \text{if } outputLength \leq referenceLength \end{cases} \quad . \tag{3}$$

ReferenceLength is the test corpus effective reference length (sum of the best match lengths which are the closest reference translation lengths; if there are two lengths equally close, the shorter one is taken), and *outputLength* is the total length of the hypothesis translation corpus. Finally, the BLEU metric is defined as in (4) [9].

$$BLEU = \exp\left\{ \sum_{n=1}^{N} w_n \log p_n - \max\left(0, \frac{L^*_{ref}}{L_{sys}} - 1\right) \right\} \quad . \tag{4}$$

NIST. Since IBM showed a strong correlation between BLEU scores and human assessments of translation quality, DARPA commissioned NIST to develop an MT evaluation facility based on the IBM work. Since BLEU uses a geometric mean of co-occurrences over N, the score is equally sensitive to proportional differences in co-occurrence for all N. This might lead to counterproductive variance due to low co-occurrences for the larger values of N. This problem is overcome by using an arithmetic mean of n-gram counts. Furthermore, n-grams that are more informative, i.e. that occur less frequently, deserve more weight (5) [9]. The formula for calculating NIST score is given in (6), where the ratio used in minimization stands for the number of words in the translation being scored and the average number of words in a reference translation, averaged over all reference translations. Factor β is chosen to make the brevity penalty factor 0.5 when the number of words in the system output is two-thirds of the average number of words in the reference translation, and when N equals to 5. A change in the brevity penalty factor is made to minimize the impact of small variations in the length of a translation [9].

$$Info(w_1 \dots w_n) = \log_2 \left(\frac{\#w_1 \dots w_{n-1}}{\#w_1 \dots w_n} \right) \quad . \tag{5}$$

$$NIST = \sum_{n=1}^{N} \left\{ \frac{\sum_{co\text{-}occur} Info(w_1 \dots w_n)}{\sum_{output}(1)} \right\} \times \exp \left\{ \beta \log^2 \left[min \left(\frac{L_{sys}}{\overline{L}_{ref}}, 1 \right) \right] \right\} \quad . \tag{6}$$

3 Experimental Study

The study has been divided into two parts. In the first part translations from Croatian into English have been obtained from four online MT services, i.e. Google Translate, Stars21, Translation Guide and InterTran. Google Translate currently supports translation between 57 languages. Croatian has been supported since 2008. Stars21 for the Croatian-English language pair offers services from Google Translate, InterTran or TranStar, which is again powered by Google Translate. In spite of that, the reported results for Google Translate and TranStar are not exactly the same, which can most probably be explained by different pre- or postprocessing techniques used by these services. The service offered by Translation Guide is powered by InterTran, which, on the other hand, is powered by WordTran and NeuroTran. WordTran consists of word-by-word and phrase-by-phrase translations. NeuroTran is a rule-based system which takes care of inflections and word-order. Although they are powered by the same engine, the translations obtained from Translation Guide and InterTran differ somewhat. To put it more precisely, Translation Guide has much higher percentage of untranslated words, especially words with diacritics. This can again be explained by different pre- or postprocessing techniques.

Source texts are short excerpts from four different domains, i.e. city, law, football and monitors. These excerpts contain 9, 9, 7 and 9 sentences, respectively.

The obtained translations have been evaluated by three different automatic metrics, i.e. F-measure, BLEU and NIST, as well as by 48 translators on a 1-5 scale according to two criteria, fluency and adequacy. The evaluators have scored the MT output according to the fluency criterion without seeing the source sentences or reference translations. Next, they have scored the output according to the adequacy criterion with regard to the source sentences. Since the two criteria are usually related, we have taken the average of the two.

According to human assessments, TranStar proves to be the best system with the score of 4.66 when the score is averaged over all domains, and Translation Guide the worst with the score of 1.17. Google Translate performs slightly worse with the score of 4.62, and beats TranStar only in the football domain, which is also the domain with the highest score achieved.

The same study has been conducted in the reverse direction, but has included only the popular Google Translate system. Excerpts from all four domains contain 9 sentences each. Altogether 50 human assessments have been collected and the obtained score averaged over all four domains is 4.29. The best scored domain is the city domain, and the worst scored is the football domain.

Prior to running automatic evaluation, lowercasing and tokenization have been done. All of the calculations are based on a single reference per sentence. The results are presented in the subsequent subsections.

3.1 F-measure

F-measure ranges from 0 to 1. Individual F-measure scores obtained for translations from Croatian into English and vice versa are presented in Table 1. Overall F-measure scores obtained for translations from Croatian into English are presented in Table 2. The overall F-measure score obtained for Google Translate system for the English-Croatian language pair is 0.7224.

Table 1. F-measure obtained for four different MT systems in four different domains.

		Domain			
	System	City	Law	Football	Monitors
CRO → EN	Google Translate	**0.6945**	0.6301	**0.7999**	**0.8873**
	TranStar	**0.6945**	**0.6475**	0.7857	**0.8873**
	Translation Guide	0.1957	0.1458	0.2281	0.2166
	InterTran	0.3754	0.4082	0.4109	0.3343
EN → CRO	Google Translate	0.8030	0.7599	0.6373	0.6781

Table 2. Overall scores obtained for four different MT systems for the Croatian-English language pair.

CRO → EN	F-measure	BLEU	NIST
Google Translate	0.7348	**0.5383**	7.2234
TranStar	**0.7376**	0.5337	**7.2596**
Translation Guide	0.1907	0.0551	2.4969
InterTran	0.3863	0.0873	3.5205

3.2 BLEU

Possible BLEU scores range from 0 to 1. Individual BLEU scores obtained for translations from Croatian into English and vice versa are presented in Table 3. Overall BLEU scores obtained for translations from Croatian into English are presented in Table 2. The overall BLEU score obtained for Google Translate system for the English-Croatian language pair is 0.5836.

Table 3. BLEU scores obtained for four different MT systems in four different domains.

		Domain			
	System	City	Law	Football	Monitors
CRO → EN	Google Translate	**0.5050**	0.3719	**0.5941**	**0.7796**
	TranStar	**0.5050**	**0.3957**	0.5402	**0.7796**
	Translation Guide	0.0576	0.0405	0.0708	0.0530
	InterTran	0.1144	0.0814	0.1003	0.0545
EN → CRO	Google Translate	0.7102	0.5577	0.5290	0.5311

3.3 NIST

A NIST score of 0 means that the hypothesis and the reference have no n-grams in common. Higher positive scores suggest better translations. Individual NIST scores obtained for translations from Croatian into English and vice versa are presented in Table 4. Overall NIST scores obtained for translations from Croatian into English are presented in Table 2. The overall NIST score obtained for Google Translate system for the English-Croatian language pair is 7.2016.

Table 4. NIST scores obtained for four different MT systems in four different domains.

		Domain			
	System	City	Law	Football	Monitors
CRO → EN	Google Translate	**5.5714**	4.8921	**6.0526**	**6.7125**
	TranStar	**5.5714**	**5.0639**	5.9828	**6.7125**
	Translation Guide	2.0486	1.7176	2.0723	2.3891
	InterTran	3.1390	2.8658	3.1540	2.6141
EN → CRO	Google Translate	6.2356	6.0487	5.0035	5.2658

4 Discussion

4.1 Croatian-to-English Translation Task

According to all automatic measures, Google Translate and TranStar are best suited for translating technical manuals, i.e. monitors, from Croatian into English, and worst suited for translating legal documents. Surprisingly, human evaluators find Google best suited for translating in the domain of football, and TranStar best suited for translating legal documents for translation from Croatian into English. Both are worst suited for literary descriptions, i.e. city information. All of the metrics, as well as human evaluators almost completely agree on the system rankings. The only disagreement shows BLEU which gives the highest rank to Google Translate instead of TranStar (Fig. 1). System level correlation between F-measure, BLEU, NIST and human assessments for the Croatian-to-English translation task is given in Fig. 2. A correlation of 1 means that there is a positive linear relationship between the two variables, a correlation of -1 means that there is a perfect negative linear relationship between them, and a correlation of 0 means that there is no linear relationship between them. The correlation between F-measure and BLEU is high for Google Translate and TranStar, somewhat lower for Translation Guide, and the lowest for InterTran. The same applies to the correlation between F-measure and NIST. The correlation between BLEU and NIST is extremely high for all but the worst ranking system Translation Guide. The correlation between automatic metrics and human assessments is much lower. The strongest correlation between all three automatic metrics and human assessments shows Google Translate. The least strong correlation between BLEU and human assessments shows, surprisingly, TranStar. The correlation between automatic metrics and human assessments averaged over all the systems for the Croatian-to-English translation task is shown in Table 5. The significance of the correlations has been tested through a two-tailed test at the 0.05 significance level with two degrees of freedom. The results in Table 5 are not statistically significant. The correlation between F-measure and BLEU, as well as F-measure and NIST, is statistically significant

for Google Translate and TranStar, while the correlation between BLEU and NIST is significant for all but InterTran. None of the automatic metrics significantly correlates with human assessments.

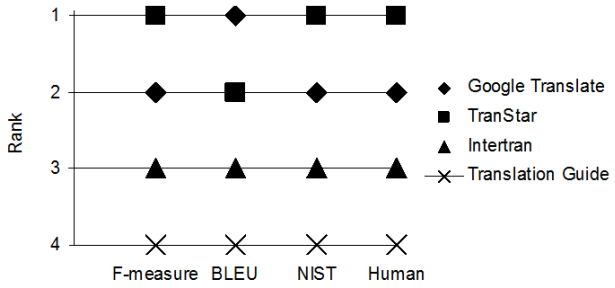


Fig. 1. Ranking of four systems in the Croatian-to-English translation task (Google Translate, TranStar, Translation Guide and InterTran) according to three automatic metrics (BLEU, NIST and F-measure) and human assessments.

Table 5. Correlation between automatic metrics and human assessments averaged over all systems for the Croatian-to-English translation task.

	F-measure	BLEU	NIST
BLEU	0.8490		
NIST	0.8665	0.8441	
Human	0.3182	0.3040	0.2135

4.2 English-to-Croatian Translation Task

The correlation between automatic metrics and human assessments for translations from English into Croatian is given in Table 6. According to the two-tailed significance test at level 0.05 with two degrees of freedom, only the correlation between NIST and F-measure is statistically significant. The correlation for each criterion separately, i.e. fluency and accuracy, is shown in Fig. 3. Lastly, the scores for each metric for both translation directions are shown in Fig. 4. We observe that there is a negative linear relationship between the two translation directions, however, this relationship is not statistically significant.

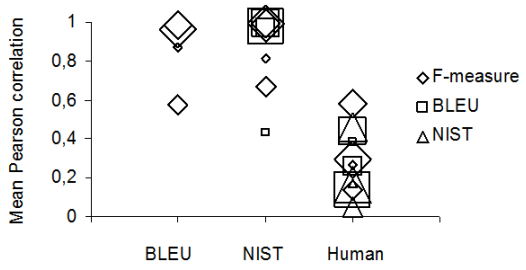


Fig. 2. System-level correlation between F-measure, BLEU, NIST and human assessments in the Croatian-to-English translation task. The size of a shape for a system depends on the majority-agreement-ranking of four systems, where the best system, i.e. TranStar, has the biggest shape.

Table 6. Correlation between automatic metrics and human assessments averaged over all systems for the English-to-Croatian translation task.

	F-measure	BLEU	NIST
BLEU	0.8272		
NIST	0.9933	0.7712	
Human	0.7437	0.6766	0.6809

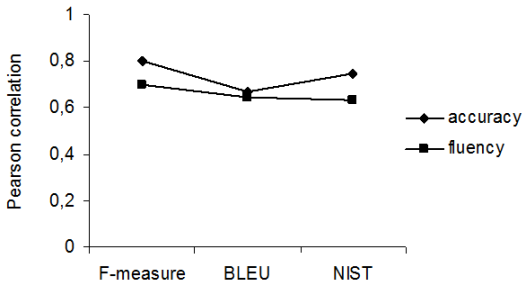


Fig. 3. System-level correlation between automatic metrics and human accuracy and fluency assessments in the English-to-Croatian translation task.

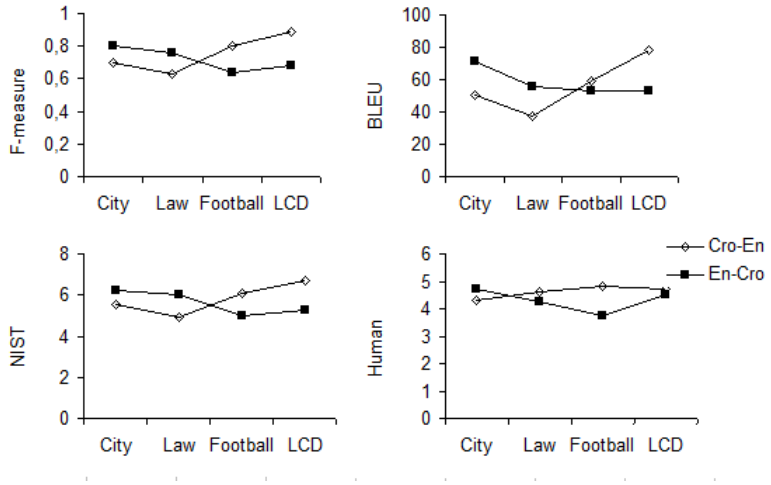


Fig. 4. Google Translate correlation between two translation directions according to three automatic metrics and human assessments.

5 Conclusion

In the first part of the study translations from Croatian into English have been obtained from four MT systems, i.e. Google Translate, TranStar, Translation Guide and InterTran in four different domains, and scored by the three fully-automatic accuracy evaluation metrics. All of the metrics, as well as human evaluators, almost completely agree on the rankings of the systems. The correlation between F-measure and BLEU, as well as F-measure and NIST, is statistically significant for Google Translate and TranStar, while the correlation between BLEU and NIST is significant for all but InterTran. In the reverse translation direction only Google Translate output has been evaluated and the correlation between F-measure and NIST proved to be statistically significant. However, none of the automatic metrics significantly correlates with human assessments. This might be due to the size of the test set, which is limited due to a lack of human evaluators and to the time-consuming nature of manual evaluation task. F-measure has the highest correlation with human assessments for the two highest ranking systems in the Croatian-to-English translation task, as well as for the reverse translation direction. For the worse two systems BLEU correlates with human assessments the best. Although not statistically significant, we observe that there is a negative linear relationship between the two translation directions.

We conclude that adding more reference translations might improve reliability of the automatic metrics, since a hypothesis that might be perfectly correct

is scored badly if it differs a lot from a reference translation. In our future work we will investigate the correlation between the remaining metrics in the field of MT evaluation.

References

1. Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.: Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 17–53. Association for Computational Linguistics, Uppsala, Sweden (2010)
2. Jurafsky, D., Martin, J., Kehler, A.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall (2009)
3. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231. Cambridge, Massachusetts (2006)
4. Leusch, G., Ueffing, N., Ney, H.: A Novel String-to-string Distance Measure with Applications to Machine Translation Evaluation. In: Proceedings of MT Summit IX, pp. 240–247. New Orleans, Louisiana (2003)
5. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Ann Arbor, Michigan (2005)
6. Nießen, S., Och, F., Leusch, G., Ney, H.: An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, pp. 39–45. Athens, Greece (2000)
7. Koehn, P., Corporation: Statistical Machine Translation, Cambridge University Press (2010)
8. Mauser, A., Hasan, S., Ney, H. Automatic Evaluation Measures for Statistical Machine Translation System Optimization. In: International Conference on Language Resources and Evaluation, pp. 3089–3092. Marrakech, Morroco (2008)
9. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In: Proceedings of the 2nd International Conference on Human Language Technology Research, pp. 138–145. Morgan Kaufmann Publishers Inc., San Francisco, California (2002)
10. Turian, J., Shen, L., Melamed, I.: Evaluation of Machine Translation and Its Evaluation. In: Proceedings of the MT Summit IX, pp. 386–393. New Orleans (2003)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania (2002)
12. Lin, C., Och, F.: Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 605–612. Association for Computational Linguistics, Barcelona, Spain (2004)

Translation and the human factor

Knowledge of Provenance and its Effects on Translation Performance in an Integrated TM/MT Environment

Carlos Teixeira

Intercultural Studies Group, Universitat Rovira i Virgili,
Avda. Catalunya 35 – 43002 Tarragona, Spain
carlostx@linguanativa.com.br

Abstract. The integration of machine translation (MT) and translation-memory (TM) systems in professional translation settings has turned pre-translation + post-editing into an attractive alternative in terms of productivity for all parties involved in the translation process. In some cases, source files are pre-translated using a combination of customised MT and TM before reaching the translators, who then become reviewers, or post-editors. But how does this actually affect productivity and how do translators feel when performing this new activity? In order to look for answers to those questions, I ran a pilot experiment comparing two different environments. The basic difference between the two is the availability of information on the provenance of the suggested translation for a particular segment (whether it comes from MT, TM, and at which match percentage). Data were collected using screen recording, keystroke logging and post-performance interviews.

Keywords: translation technology, translation memory, machine translation, process research, speed, productivity, performance, provenance, trust.

1 Introduction

Until recently, machine translation (MT) and translation memories (TM) were seen as totally different approaches to using technology in translation. While the first approach was largely restricted to end users interested in grasping the general idea of a text written in a language they could not understand (usually while browsing the Internet), the second was addressed to professionals in the translation industry, such as translators, translation agencies or translation departments in large companies.

However, this scenario has been changing at a rapid pace in the last few years, mainly due to quality improvements and the general availability of statistical machine-translation systems, based on large amounts of human-produced bilingual text. This has allowed MT to be progressively integrated into TM tools in professional translation environments, bringing new possibilities as well as new challenges.

The potential productivity gains derived from this integration of machine translation (MT) and translation memories (TM) are calling for new work methods in

the translation market. As an example, some translation agencies pre-translate their source files using a combination of TM and customised MT before sending them out to translators, who then become reviewers, or post-editors. In this scenario, translators review each segment without knowing its provenance, i.e. whether it came from a translation memory (and at which match percentage) or from a machine-translation engine. Could this missing information have an impact on the way translators perform their tasks, compared to a more traditional environment, where translators would know where each translation suggestion comes from? In other words, how does the 'knowledge of provenance' of translation suggestions affect translators' behaviour in environments that integrate TM and MT?

2 Previous Research in the Field

None of the published studies on translation technology that we are aware of seems to take into account this specific aspect that distinguishes translation memory systems from machine translation systems: TM systems show translators the 'provenance' and the 'quality' of the translation suggestions coming from the memory, whereas MT systems display the 'best translation suggestion possible' without any indication of its origin or degree of confidence. It is our assumption that this missing distinction might be one of the reasons for discrepant results in some studies that compare translation speed when (post-)editing MT and TM suggestions.

As an example, [2] compares the performance of TM vs. MT when translators work in a 'traditional' TM system, i.e. when they know the provenance of the translation suggestions they are working with. One of her findings is that "cognitive load [and processing speed] for machine translation matches is close to fuzzy matches of between 80-90% value" (p.185). For fuzzy matches above 90%, including exact matches, TM processing is faster and requires a lower cognitive load, whereas the opposite happens for fuzzy matches below 80%.

In a different study, [1] reproduces an actual scenario that can be found in industry. The author analyses time and quality when editing translation-memory suggestions vs. machine-translation suggestions, in an environment where translators could not tell the provenance of each suggestion. Under this particular condition, her "findings suggest that translators have higher productivity and quality when using machine-translated output than when processing fuzzy matches [at any percentage level] from translation memories" (p.11).

3 Research Question and Hypotheses

Inspired by these two studies and their rather contradictory findings (at least for high-percentage fuzzy matches), I set out to investigate whether the fact of knowing the provenance of the segments could provide and explanation for this discrepancy. What are the differences (if any) in the translation process between a situation where translators know the provenance of the translation suggestions they are editing and a situation where this information is not available?

In order to answer this question, I compared two translation environments. In the first environment, translators do not know the provenance of translation suggestions, whereas in the second environment translators do have access to this information. These are my working hypotheses:

Hypothesis 1 (H1): The *translation speed* is higher when *provenance information* is available.

Hypothesis 2 (H2): There is no significant difference in the *quality* level when *provenance information* is available.

Some definitions are necessary in order to operationalize the variables we want to test:

- *Translation speed* is measured as words per hour. There are separate counts for the first rendition (drafting) and second rendition (self-revising).
- *Provenance information* of translation suggestions is indicated by showing their origin (TM or MT) and, in the case of TM, by displaying its fuzzy-match percentage and highlighting the differences between the actual segment and the matching segment in the TM, as is usually the case in most TM systems.
- *Quality* is measured as a score given by two reviewers, who process all resulting translations according to predefined criteria (see section 4.3).

4 Pilot Experiment

In order to test those hypotheses and fine-tune the methodology for my future doctoral research, I ran a pilot experiment with two translators from English to Spanish. Each of them translated two similar source texts of around 500 words each, in the two different environments described below.

Environment B presents the source-text segments on the left-hand side of the screen and a pre-translated version of the source text (obtained through the pre-processing of the file with TM and MT) on the right-hand side. In this case, all no-matches were replaced with MT suggestions, and the whole text was presented as a sequence of pre-translated segments. Translators could edit the pre-translated segments as if they were revising a translated file and they had no information on the origin of each of the pre-translated segments (i.e. whether they came from a TM segment or an MT segment). For mnemonic recall, let us call this environment B (as in 'blind'). This environment tries to reproduce as close as possible the environment described in [1].[1]

Environment V is similar to the previous one in that translators also had access to the source-text segments on the left-hand side of the screen and an editing space on the right-hand side. However, its difference consists in that, instead of working by 'blindly' editing pre-translated segments, translators could see where the default

[1] Our environment B presents all the pre-translated text at once, while the environment used in [1] displays each pre-translated segment at a time and does not allow for a revising phase. In order to make both environments closer (and environment B closer to environment V), we are planning to change the settings of our Trados project in the future to have it display each segment at a time. There are no plans to restrict the revising phase in our study.

translation suggestion was coming from (either from the translation memory or from the MT engine). Additionally, in the case of TM suggestions, translators could see the highlighted differences between their actual source segment and the TM source segment. For mnemonic recall, let us call this environment V (as in 'visual'). This environment tries to reproduce as close as possible the environment described in [2].

4.1 Subjects

Both subjects are male and native speakers of Spanish. Subject1 has formal training in translation and 4 years of professional experience in several fields, especially audio-visual translation. Subject2 also has formal training in translation and around 8 years of professional experience in various fields, mainly in localisation and technical translation. Both are familiar with many different translation memory systems.

For my main experiment, I plan to have ten English-to-Spanish translators who are native speakers of Iberian Spanish: five male and five female subjects. They will be selected by means of a questionnaire and will have 5+ years of professional experience working with translation-memory systems on technical or marketing texts. Formal training in translation will not be a prerequisite.

All ten translators will translate both texts in both environments. Five translators will start working in B and the other five in V, in order to account for potential differences related to the order of the tasks

4.2 Materials

Our source texts were taken from an article in a technical magazine and deal with composite materials in car manufacturing. The main reason for choosing this kind of material was a wish to use text outside of the 'software localisation' domain – which is the object of most research studies in the field – still with (marketing) stylistic features that make it more demanding for translation. The specific article was chosen mainly because of its topic (technical, while still somehow interesting for translators) and length – allowing for the extraction of two excerpts of around 500 words.

The main article had a total of 1310 words, corresponding to 55 source segments, or 23.8 words per segment in average. In order to have two source texts of around 500 words, I used 21 segments for each of them. As a result, SourceText1 has 512 words, and SourceText2 has 510 words.

A translation memory was created by aligning the English source text with the Spanish target text (final version revised by a copy-editor and approved by the client) using SDL Trados WinAlign + manual verification of each segment. A decision was made to use the following fuzzy match distribution in the experiment:

- 7 'no matches' (replaced by MT feeds);
- 5 exact matches;
- 9 fuzzy matches, of which
    - 3 matches within the 70%-79% range,
    - 3 matches within the 80%-89% range, and
    - 3 matches within the 90%-99% range.

The order of presentation of match types during translation was defined by a random number generator and it was different for each of the environments. Then I edited the aligned memory to obtain two memories with the characteristics above, one for each environment. Segments set to have an 'exact match' suggestion were left untouched. Segments corresponding to a 'no match' were replaced through SDL Trados Studio with translation suggestions provided by the public, freely available Google Translate machine-translation engine. Finally, for creating the fuzzy matches I resorted to the following strategies: delete parts of the source and target segments, include or replace some words in the source and target segments, or edit the source text.

4.3 Data Collection

The two translation environments were created within SDL Trados Studio 2009 Freelance. The main methods for collecting data were screen recording and keystroke logging through BB Flashback Express 2. Retrospective interviews were also used to try to obtain some insight of translators' feelings and satisfaction in both tasks. Think-aloud protocols were not used as they are known to slow down the translation process and we were precisely trying to measure translation speed in a natural(istic) environment. For testing quality, all texts were rated by two reviewers.

Time was measured by watching each of the translators' performances in BB FlashBack Player and manually noting down the start and end times for each individual task. Time was counted when translators were typing, thinking, hesitating, or looking at the source text (except when they read the full source text before starting the translation, as we cannot make a correspondence between the time and specific segments). Time was not counted when translators switched to another window to look up terminology, tried to find a specific function in the tool, or spoke with the researcher. The time counter was paused when the subject started moving the mouse to go to another application (usually a web browser) outside of the translation environment. It was also paused when the subject moved to the source segment to copy text to be pasted in the browser. Time count was resumed when the subject returned to the translation environment. Time spent on searches within the translation environment (mainly with the Concordance function) was considered as translation time.

For assessing quality, all texts were rated by two reviewers, based on an error-count system. The quality level of a translation was measured through a score, which starts at 10 and decreases according to the grid shown in Table 1.

4.4 Running the Experiment

Both subjects chose to use their own laptop computers during the experiment. Before they started, we made sure they had the required version of SDL Trados and BB FlashBack installed and configured. The aim was to have translators work in an environment as close as possible to their natural work environment, meaning that they could keep their preferred configuration in terms of keyboard, screen and mouse

(either built-in or external), operating system (within the Microsoft Windows family), browser favourites, dictionaries, etc. They also had access to the Internet during the experiment.

Table 1. Quality assessment grid

Type of error	Deduction
1 misspelled word	.25
1 grammar mistake (morphology, syntax)	.25
1 use of incorrect or inconsistent terminology	.25
1 general readability (understanding) issue	.25
1 sentence structuring issue (style, register)	.25
1 instance of omitted information	.25
1 instance of incorrect or inaccurate meaning rendition	.25
1 localisation error (numerical formats, units)	.25
Other deductions	.25 each

At the beginning of the experiment, a digital voice recorder was turned on. The initial tasks subjects had to perform were: (a) copy a short passage in Spanish, and (b) translate a short passage from English to Spanish. In both tasks, the source texts were printed out and translators had to type their target texts in Microsoft Word. The purpose of these two initial tasks was twofold: to measure their baseline typing speed (and eventually assess whether this has an influence on their editing strategies) and to serve as a warm-up (and stress-down) activity. This came from a suspicion that the typing ability of each individual translator might have an influence on their performance with each kind of translation suggestion.

Next, the translators were given instructions in Spanish on how to perform the main tasks for the experiment.[2] In general terms, the instructions told the subjects that the memory they would be provided was created based on a client-approved final version of the Spanish magazine, that it contained five different kinds of matches, and that machine translation was used to replace 'no match' segments. The translation 'brief' mentioned the translators would be paid the same amount per word (no fuzzy-match discounts), so implying that they were supposed to revise all segments, including exact matches. The instructions also made it clear that their translations were going to be assessed and graded for quality by a professional reviewer, thus also implying that the translators should try to achieve maximum quality in both environments. A time limit of 1.5 hours was set for each of the texts.

During the translation of the texts in both environments, BB FlashBack was set to record screen activity; keystrokes; mouse position, movements and clicks; translators' faces; and sound (voices, keyboard, etc).

[2] Subject1 received the instructions orally, but then the researcher decided to give similar instructions in written format to Subject2, in order to eliminate potential variations due to his oral performance. In the main experiment, all subjects should receive the same instructions in written format.

5 Preliminary Results

5.1 Subject1

Tables 2 and 3 show the average speed results for Subject1.

Table 2. Average translation speed per type of segment in environment V for Subject1.

	SOURCE WORDS	TIME (sec) 1st rendition	SPEED (words/hr) 1st rend.	TIME (sec) 2nd rendition	SPEED (words/hr) Combined
EXACT (100%) MATCHES	131	155	3036	94	1895
90-99% MATCHES	91	234	1397	101	977
80-89% MATCHES	51	153	1197	27	1019
70-79% MATCHES	87	401	781	88	577
NO MATCHES (MT FEEDS)	150	783	690	132	591
	510	**1727**	**1063**	**441**	**847**

Table 3. Average translation speed per type of segment in environment B for Subject1.

	SOURCE WORDS	TIME (sec) 1st rendition	SPEED (words/hr) 1st rend.	TIME (sec) 2nd rendition	SPEED (words/hr) Combined
EXACT (100%) MATCHES	128	566	815	49	749
90-99% MATCHES	65	369	635	53	555
80-89% MATCHES	77	210	1321	30	1155
70-79% MATCHES	77	273	1014	20	946
NO MATCHES (MT FEEDS)	165	592	1004	129	825
	512	**2009**	**917**	**281**	**805**

If we look at the average results for the first rendition (drafting), we see that translation speed is higher in V (1063 words/hr) than in B (917 words/hr), a difference of 15.9 percent. If we look at the results for the first and second renditions (drafting + self-revising) combined, translation speed is still slightly higher in V (847 words/hr) than in B (805 words/hr), but the difference is reduced to 5.2 percent. Due to the dispersion of the data and the reduced number of segments in the texts, the detected difference in the overall speed between V and B is not statistically significant.

However, if we look at the different fuzzy-match levels, differences in speed are more pronounced. In environment V, it is possible to identify three groups of speed levels: exact matches are translated the fastest, fuzzy matches between 80-99% are translated at around half that speed, and lower fuzzy matches (below 80%) and MT output are translated the slowest. This is in accordance with intuitive expectation and with the results obtained by [2].

In environment B, there is a dramatic reduction in speed (from 1895 to 749 words/hr) for exact matches, suggesting that provenance information has a high impact on this kind of translation suggestion. Matches in the 90-99% range also show a dramatic reduction in speed (from 977 to 555 words/hr), again indicating that provenance information has a significant impact in this case. Matches in the 80-89% range did not show a significant variation. For lower fuzzy matches and MT feeds, it is worth noting that there was an *increase* in speed.

5.2 Subject2

Tables 4 and 5 show the average speed results for Subject2.

Table 4. Average translation speed per type of segment in environment V for Subject2.

	SOURCE WORDS	**TIME (sec) 1st rendition**	**SPEED (words/hr) 1st rend.**	**TIME (sec) 2nd rendition**	**SPEED (words/hr) Combined**
EXACT (100%) MATCHES	131	236	2000	121	1323
90-99% MATCHES	91	354	925	160	637
80-89% MATCHES	51	225	814	77	606
70-79% MATCHES	87	456	687	135	530
NO MATCHES (MT FEEDS)	150	475	1138	214	784
	510	1746	1052	708	748

Table 5. Average translation speed per type of segment in environment B for Subject2.

	SOURCE WORDS	TIME (sec) 1st rendition	SPEED (words/hr) 1st rend.	TIME (sec) 2nd rendition	SPEED (words/hr) Combined
EXACT (100%) MATCHES	128	445	1035	95	854
90-99% MATCHES	65	275	852	52	716
80-89% MATCHES	77	226	1229	143	752
70-79% MATCHES	77	289	961	79	755
NO MATCHES (MT FEEDS)	165	568	1045	161	814
	512	1802	1023	530	790

For this translator, the average results for the first rendition (drafting) show that translation speed is also higher in V (1052 words/hr) than in B (1023 words/hr), but the difference is much smaller than for Subject1, at only 2.8 percent. The combined results for the first and second renditions (drafting + self-revising) show that translation speed is now higher in B (790 words/hr) than in V (748 words/hr), with a difference of 5.6 percent. As was the case with the data for Subject1, this difference is not statistically significant.

Now let us look again at the speed differences according to the various fuzzy-match levels. Roughly speaking, the data for environment V indicate that Subject2 processed translation suggestions coming from exact matches two times faster than suggestions coming from fuzzy matches (1323 vs. 591 words/hr in average), and he translated suggestions coming from machine translation around 33 percent faster than the average speed for fuzzy matches. The faster speed for exact matches is still in accordance with our expectations, but the reasons for machine-translation suggestions being translated faster than high-percentage fuzzy matches should be investigated further.

In environment B, similarly to what happened with Subject1, the data for Subject2 indicate a dramatic reduction in the average translation speed (from 1323 to 854 words/hr) for suggestions coming from TM exact matches. All other kinds of translation suggestions had an increase in speed, with fuzzy matches in the 80-89% range showing the largest increase (42.5 percent). It is interesting to note that differences in translation speeds tend to disappear in the blind environment: exact matches were translated slightly faster, at 854 words/hr, followed by machine-translation suggestions, at 814 words/hr, with translation-memory fuzzy matches being translated a little more slowly, between 716 and 755 words/hr. If the statistical errors are taken into account, differences between the five types of translation suggestions are actually not significant.

5.3 Quality

Two revisers assessed the quality of the four translations (two per subject) using the grid provided in section 4.3. Revisers were then told to compare the two translations from the same subject and decide which one was better, if any, and to give their final grade from 0 (worst) to 10 (best). This means each reviser scored the translations twice – once according to the grid, then again holistically. The results are shown in Table 6.

Table 6. Translation quality levels for both subjects.

	Subject1		*Subject2*	
	Text 1 (environment V)	Text 2 (environment B)	Text 1 (environment V)	Text 2 (environment B)
Reviser 1	8.5	7.0	8.5	9.0
Reviser 2	7.5	7.0	8.0	8.5
Average	***8.0***	***7.0***	***8.25***	***8.75***

According to the two evaluators, Subject1 performed better in environment V, while Subject2 performed slightly better in environment B. From the evaluators' feedback, we think that quality assessment has not been done properly and the above grades need to be revised again before we can make any definite conclusions. Furthermore, we think the rating instructions need to be made clearer and a greater number of revisers shall be used.

6 Discussion

We took pains to control most of the factors that might affect our results (type of text, length of text, source language, target language, translator's experience, translation tool, etc.) and we tried to have only our main independent variable (knowledge of provenance) act on our two dependent variables (speed and quality). However, we are aware that many potential extraneous variables (confounds) were also present and had not been properly considered.

Data from our pilot experiment do not allow us to draw a definite conclusion on our first hypothesis (on speed) if we take the whole texts as a reference. Subject1 was slightly faster (5.2 percent) in environment V, while Subject2 was slightly faster (5.6 percent) in environment B. However, we can assume that the overall speed, besides individual-specific differences, depends on the distribution of different types of translation suggestions in the texts, as both subjects were faster with certain types of

suggestions. For example, if our texts contained only exact matches and machine translation feeds, our results for the entire texts would probably be different.

Although our aim was to be able to draw some conclusions from a pair of intra-subject studies (if the current pilot experiment can described in this way), the translators' personal styles (and the revisers' preferences) played a more prominent role than we had originally expected. The retrospective interviews are still being processed and a deeper analysis of this material might help shed light on some of the results. For example, Subject1 seems to have 'respected' the suggestions from the translation memory more often than Subject2, and one of the evaluators did not like the solutions present in the original memory (the version approved by the client). This fact might explain why the quality results for this translator are slightly lower.

In any case, our second hypothesis needs further verification, as the data we have on quality for each subject are not sufficient to determine even whether the translations produced in the two environments can be considered of same quality for each individual subject.

6.1 Limitations

Below is a list of known limitations of the pilot experiment and, wherever possible, some solutions to overcome them in the main experiment.

Small number of subjects. This is a common problem in translation process research, and we hope the inclusion of more subjects (10) in the future will make data statistically more relevant.

Experience increases over time. The subjects' experience (thus their speed and quality) in working in both environments (B and V) can increase over time, at least as far as post-editing MT is concerned. Therefore, the data we are gathering might be representative of performance at the beginning of a learning curve. One solution would be to train translators for some period and measure their performance after some time.

Few segments. The text chosen as source text had long segments, which obliged us to use only a few segments per type of suggestion. Since we do not want to increase the total word volume of the source texts, we will probably need to choose another article or even another text type.

Irregular segments. The shortest segment had six words, while the longest had 44, which makes them hardly comparable, as MT is known to work better with segments containing a 'single idea' and worse with long sentences. Same solution as above.

Terminology. The distribution of terms in the source text should be reconsidered for the main experiment. Even though the time used for terminology search was discounted, the time spent within the translation tool was higher when terms were more complicated. This was partly compensated for by the fact that the type of suggestion for each segment was defined randomly, but in order to eliminate extraneous variations, we will try to remove problematic terms or provide a glossary for them.

Segment identification. Sometimes it was difficult to identify which segment translators were focusing on, especially in the self-revising phase. Eye tracking is an additional data-collection method that is being considered to help solve this issue.

7 Conclusion

We set out to investigate whether provenance information about translation suggestions in translation environments that integrate TM and MT has an impact on speed and quality. We ran a pilot experiment with two subjects that translated two 500-word texts in two different environments. Through screen recording and keystroke logging, we measured the time spent for each of five different types of translation suggestions. The final translated texts were assessed for quality by human reviewers. Retrospective interviews completed the data gathering methodology with an aim at obtaining general impressions from the subject translators.

Our data show that the overall speed was not significantly different in the two scenarios and the quality was of comparable level. If we look into individual types of suggestions, data on speed also show that translators spent much longer translating (post-editing) exact matches when they did not know the provenance of the suggestions.

Although inconclusive, the results of the current study indicate that 'provenance information' is relevant for translators working with translation suggestions from TM and MT, and that this information should be taken into account when analysing and comparing the results of different experiments.

We expect this study will help increase knowledge on translation and post-editing processes, which can be beneficial for all parties involved in the translation scene, including independent translators, translation agencies, translation-tool developers and, ultimately, translation customers, as the results can contribute to devise optimal workflows and best practices.

Besides the potential impacts on earnings (and savings), the search for optimal processes can increase the volume of text that can be processed. Even more important, it is our concern to try to optimise the translation process in ways that will help increase job satisfaction among translation professionals. Finally, I hope the results will also be of intellectual importance, as we are trying to demonstrate that the impact of technology is not just in what it does, but also in what the stakeholders know about what it does.

References

1. Guerberof, A. Productivity and quality in the post-editing of outputs from translation memories and machine translation. Localisation Focus – The International Journal of Localisation 7/1: 11--21 (2009)
2. O'Brien, S. Eye-tracking and translation memory matches. Perspectives: Studies in Translatology 14/3: 185--205 (2006)

The Impact of Translation-Memory (TM) Technology on Cognitive Processes: Student-Translators' Retrospective Comments in an Online Questionnaire

Tina Paulsen Christensen & Anne Schjoldager

Aarhus University
Business and Social Sciences
Department of Business Communication
Translation and Interpreting

Abstract. The use of Translation-Memory (TM) technology and other translation software is bound to influence translators' cognitive processes. Unfortunately we still lack empirically founded knowledge of this. Our paper therefore presents and discusses the theoretical background, setup and preliminary findings of a small-scale pilot study of student-translators' retrospective comments in an online questionnaire survey regarding what they had experienced during an introductory hands-on course in TM-assisted translation. We also discuss some basic concepts and methods within translation process research, and apply a simplified model of the translation process that comprises three main phases taken from a general writing model: planning, drafting and postdrafting. As far as our student-translators are concerned, TM technology seems to affect processes in all of these phases.

Keywords: Translation-Memory (TM) technology, translation process research, cognitive processes, retrospective comments, questionnaire survey

1 Introduction

Because of the digital revolution, professional translation is no longer a purely human activity, but nowadays tends to be carried out by means of translation-memory (TM) technology. TM technology is basically a database in which source-text (ST) segments and target-text (TT) segments are paired in order for a translator to access and re-use them in a current translation. Many features of TM technology are bound to influence the translators' cognitive processes in some way or other (Garcia 2007; Biau Gil/Pym 2006; Mossop 2006). Unfortunately, we still lack empirically documented knowledge of how translators, their workflows and cognitive processes are affected by TM technology (Christensen/Schjoldager 2010). By reporting on a small-scale study of student-translators' retrospective comments, we hope to contribute with some knowledge about the impact of TM technology on

translators' cognitive processes. Our study is inspired by and draws on Dragsted's (2004 and 2006) research on segmentation in human and TM-assisted translation and O'Brien's (2006 and 2008) studies of cognitive loads in connection with various TM match types.

As we are interested in studying translators' cognitive processes when they interact with TM technology in order to help them become more aware and critical of its impact on their work, it is worth emphasising that our aim is translation-theoretical and not technological. In other words, our aim is to understand how translators think while using TM technology, not to suggest improvements to the software itself. We would also like to emphasise that the student-translators of our study are probably more conscious of any changes dictated by TM technology than are professional and experienced TM translators, who must be expected to have integrated any changes caused by TM technology into the general translation process.

In this paper, we shall briefly discuss the field of translation process research (section 2), methods used within this field (section 3) and the concept of TM-assisted translation (section 4). We shall then proceed to discuss the background and preliminary results of our own study (sections 5-9) and shall round off with some concluding remarks (section 10).

2 Translation process research

As noted by Holmes (1972/2000) in his seminal paper on the name and nature of translation studies, translation research was traditionally product oriented, i.e. focussed on linguistic and textual descriptions of translated texts. As far as we are aware, Holmes was the first to identify a process-oriented branch of translation studies that was concerned "with the process or act of translation itself" (Holmes 1972/2000: 177). Pointing out that "the 'little black box' of the translator's mind" has been the object of much speculation, Holmes (1972/2000: 177) advocates a descriptive (i.e. empirical) approach to translation process research. As pointed out by Palumbo (2009: 92), it was Krings' (1986) pioneering work on translators' use of time and reference books by means of verbal reporting (introspection) that marked the beginning of empirically founded translation process research. Krings' study and many that followed focussed on cognition, i.e. Holmes' "little black box", drawing on and adopting methods from cognitive psychology, especially verbalising methods.

As suggested by Schubert (2009), translation process research tends to focus either on external or internal processes (see also Göpferich 2008: 1). These approaches differ essentially as far as the object of study is concerned, but they complement rather than rival each other as far as knowledge is concerned. The external process may be defined as "everything in the translation process which can be observed by another person", which he also refers to as the translation workflow (Schubert 2009: 19). Similarly, Breedveld (2002: 9) describes translation not only as a mental process, but also as "a social process in which different actors interact and influence the text-in-production". Examples of such external, observable process data are translators' use of tools or their consultation with colleagues and clients. In

continuance of this, using a TM may be described as part of an external process, and perhaps the TM itself may then be seen as an extended cognitive resource or as distributed cognition (Hutchins 2000).

Internal translation processes concern mental activities, which cannot be studied directly and therefore tend to be studied by methods that are borrowed from cognitive psychology, especially verbalising methods (see section 3, below). Internal processes – such as thoughts, feelings, beliefs etc. – may be further subdivided into conscious and non-conscious (subconscious) activities. Following Göpferich (2008: 1), we shall refer to conscious internal activities as cognitive processes. According to Hutchins (2000: 1), cognitive processes are involved in memory, decision making, inferencing, reasoning and learning, for instance.

3 Methods in translation process research

Today a variety of methods are employed in translation process research. Depending on which criteria are used, the methods can be classified in different ways. Distinguishing between offline and online methods, we take our starting point in Krings' (2005: 348) model of basic methods applied in translation process research, focusing on features that are relevant for our own investigation (section 5, below). For a more detailed discussion of the pros and cons of various methods within translation process research, see Dam-Jensen/Heine (2009) and Christensen (forthcoming).

Offline data are collected after the translation process. *Online data* are produced during the translation process. As regards online methods, Krings distinguishes between data collected by way of observation of the translation process and data collected by way of verbal-report data during the process. Regarding offline methods, he distinguishes between product analysis and verbal-report data.

Verbal-report data comprise a subject's verbalised thoughts. Since we are not conscious of automated processes, it must be assumed that only conscious, i.e. cognitive, processes can be made accessible by means of verbal-report methods (Göpferich 2008; Álvarez 2007). For our purposes, we shall regard the verbalising activity as an act of metacognition, i.e. thinking about thinking.

Online verbal-report data are obtained when translators are asked to verbalise their thoughts during the task. These verbalisations, which are regarded as introspection, are recorded orally and subsequently transcribed in think-aloud or talk-aloud protocols (both abbreviated as TAPs), for instance. Introspective verbalisations are assumed to allow us rather direct access to subjects' minds, though verbalisations are not, of course, identical to the actual processes themselves. Online verbal-report methods are generally criticised for interfering with the on-going translation process because translators are distracted from the task at hand. It has also been said that, because of the additional cognitive load of online verbalising, subjects tend to give priority to procedural thinking over other cognitive processes (House 2000: 152).

Offline verbal-report data are obtained when translators are asked to verbalise their thoughts after the translation task. These verbalisations, which are regarded as retrospection, comprise specific or general comments about a given task. Comments

are elicited in interviews or in questionnaires, for instance. Interviews and questionnaires may ask both closed and open questions. Closed questions provide mainly quantitative data, whereas open questions provide qualitative data. In translation studies, questions may relate to TTs and STs, or they may relate to workflows and (cognitive) processes. Questionnaires may be answered orally with the researcher present, or, typically, they are answered in writing and given or posted to the researcher afterwards. Offline verbal-report data, particularly retrospective comments, are sometimes criticised for rendering information that may not be consistent with what actually goes on in the subjects' minds, mainly because of the unavoidable delay between the actual processes and the verbalisations. The risk of distortion is generally thought to increase with time. Thus, for instance, Ericsson (2006: 230) notes that subjects are only able to recall relatively accurately what went on in their minds immediately after completing the task in question.

4 TM-assisted translation

TM technology is a form of computer-assisted translation (CAT). CAT covers human-aided machine translation (HAMT) and machine-aided human translation (MAHT). In HAMT, translation is essentially carried out by the software itself, but human translators are required to resolve specific problems and to correct the TT afterwards, which is mostly referred to as postediting. In MAHT, translation is carried out by a human translator, but computer assistance, TM technology for instance, is an integral part of the translation process. As mentioned in section 1, TM technology is basically a database of segmented and paired STs and TTs that a translator can access and re-use in a current translation. Thus, the TM continuously provides the translator with translation proposals (matches) that derive from his/her own or other peoples' translations. In effect, the translator may spend more time revising previous translations in segments offered by the TM than s/he does translating 'from scratch' (Garcia 2010). This and other features – the enforced segmentation of the ST (Schäler 2001; Dragsted 2004, 2006) and its uncritical and form-based method of ST/TT alignment, for instance – are bound to influence the translators' cognitive processes in some way or other (Garcia 2007; Biau Gil/Pym 2006; Mossop 2006).

5 The study

As already mentioned, the aim of our study is to discover the impact of TM technology on translators' cognitive processes, i.e. internal processes that are potentially conscious and that may be subject to metacognition and verbalisation. We assume that student-translators who have just experienced TM-assisted translation for the first (or nearly the first) time may be regarded as more suitable informants of mental changes dictated by the technology than professional and experienced TM translators, who may no longer be conscious of any changes that the technology may have caused to their mental processes. With a view to fulfilling our aim, we chose to conduct a small-scale pilot study of our own students who had

participated in an introductory hands-on course in TM-assisted translation, which was part of an obligatory course in translation methodology and theory at our department in 2009.

The course comprised an introduction to TM technology in general and Translator's Workbench (Trados 2007) in particular followed by a practical task of TM translation. For the purpose of the practical task, one of the authors (Christensen) had constructed a Danish/English TM of three STs and three TTs by aligning authentic parallel texts taken from a company website. All these texts were instructions for the use of mobile phones. Two pages from one of the English texts were doctored and used as a ST in the course, so as to allow students to retrieve different kinds of matches from the TM. Out of a total of 51 segments in the ST, students could only retrieve seven perfect matches[1] in the TM, though translation proposals at the word and phrase level could easily be retrieved from the concordance facility. Students were given no explicit translation brief, as they were merely told to translate the ST for the benefit of Danish users of the mobile phone in question. The practical task lasted for approximately 70 minutes.

Course participants were 23 MA students of English, French, German or Spanish. Out of these, 22 students (a response rate of 95.7 %) filled in an online questionnaire, which was made available to them immediately after the course and for the following week. Questionnaire answers were given in closed boxes (mainly for background information) and in open boxes, where students were asked to write their thoughts about and reactions to what they had experienced during the course. The questionnaire comprised a total of 26 questions, which, for the sake of our analyses, may be divided into five thematic parts:

A. Personal information about respondents (questions 1-7)
B. Previous experience with TM assisted translation (questions 8-13)
C. Level of IT skills (question 14)
D. Comments to the course (questions 15-19)
E. Comments to the translation process (questions 20-26)

6 What students told us about themselves

In this section, we shall summarise what students told us about themselves and what they thought about the course, i.e. their answers to questions 1-19. Most students were between 24 and 27 years old. The average age was 26.4. Most (18) students studied two foreign languages at their BA level, whereas some (4) studied a foreign language and another topic, mainly communication studies. Few (6) had experience as professional translators, which was mainly as freelancers. Some (10) were aware of the existence of TM technology prior to the course, but few had tried to use it.

[1] In fact, the seven perfect matches that could be retrieved from the TM of the course were only marked as 97 percent matches by the system. The reason for this is that, in Translator's Workbench (Trados 2007), perfect matches retrieved from aligned and imported texts are given a 3 percent penalty, to distinguish them from perfect matches aligned from texts that were translated in the system itself, which are marked as 100 percent matches.

Three said that they had tried working with TM technology before, but none of these had done so professionally.

According to their answers to our question 14 regarding their IT skills, students did not seem overconfident. When asked to which extent the following statement was true for them personally: "I know a lot about computers and most current programs and, generally speaking, I'm able to solve technical problems as they occur", most said that this was true "to some degree" (9) or "to a small degree" (8).

As mentioned above, in questions 15-19, we asked students to comment on the hands-on course itself. Many reported some technical problems with their computers and/or the program (Trados), but most seem to have sorted out these problems themselves, and they were generally well satisfied with their participation in the course.

7 Why they thought TM-assisted translation was different

We shall now focus on answers to two questions, namely questions 20 and 21. Question 20 asked students directly if they had felt that translating with a TM was different from translating without a TM. All 22 students answered 'yes' to this question. Since they had all answered in the affirmative, question 21 then asked them why they had felt that TM-assisted translation was different.

Assuming that this would get us closer to understanding perceived changes in their cognitive processes, students' answers were divided into four categories: positive comments, negative comments, neutral/other comments and don't know. Eight students gave answers that comprised both positive and negative comments. Two students gave answers that comprised both positive and neutral/other comments. Four students wrote answers that were categorised solely as positive, whereas seven students offered only negative comments. Nobody said that they did not know.

Admittedly, students' answers to the question may refer not only to internal (cognitive) processes but also to external processes (workflow). As far as the 14 positive comments are concerned, we think that they may refer to both kinds of processes, when they all imply that TM-assisted translation made their work faster, more manageable, more efficient and/or more consistent. One (No. 18) even said that TM technology made translating more interesting.

As we see it, the negative comments to the impact of TM technology are more reflective of cognitive processes than the positive comments are. Thus, a total of 13 students offered negative comments that indicated some cognitive changes, mainly that TM technology made them think less for themselves and made them trust and use their own judgement less. In other words, many regretted a general loss of control. Thus, for instance, student No. 22 wrote that s/he tended to accept uncritically what was offered by the TM and only to use his/her own knowledge when no translation proposals (matches) were offered:

> "You tend to constantly trust the options of the program, so I didn't have to think a lot for myself actually. You feel that you're getting something helpful, and you make use of it, of course. For

instance, I used the concordance facility a lot, in order to find out which proposals the program could offer (and not what I myself could offer). So, only if there were no proposals from the TM, did I translate myself, and this was a lot less pleasant than copying a proposal and pasting it into the target-text segment."

Two comments that were categorised as neutral/other (No. 17 & No. 20) expressed a similar cognitive change, but in these comments the tendency to accept uncritically the translation proposals from the TM was seen as a risk that could and should be considered. Consider what student No. 17 wrote, for instance (here in our close translation and with a few clarifying explicitations in square brackets):

> "Generally speaking I was afraid of letting the TM take over my job of assessing the [target] text. In other words, I was very much conscious of the fact that I shouldn't let my brain relax too much and just accept the translation proposals without thinking. For me, the problem was that when your eyes have first registered a translation proposal, it's harder to think of other solutions. Actually, I wish I could have looked at the English [source] text and made a draft [translating without a TM] before working with the TM, but this wouldn't be efficient, of course."

Similarly, student No. 20 commented that a changed routine caused by TM technology forced him/her to spend more time on the postdrafting stage (here in our close translation):

> "Of course, you got a good deal of possible solutions, just like that. Think though also that you risk staring yourself blind on what's on the computer screen. I would therefore never hand in anything like this, without revising it on paper first!"

8 What they thought about enforced sentence-based segmentation

In question 24 we asked students to comment specifically on sentence-based segmentation in TM technology, assuming that answers would also relate to their cognitive processes. First, students were asked to indicate what they thought about sentence-based segmentation by clicking on one of three options: (1) It's an advantage, (2) It's a disadvantage, or (3) I don't know. Eleven students said that it was an advantage; one student said that it was a disadvantage, whereas ten students said that they did not know. Then, students were asked to give reasons for these answers.

We have categorised comments as either positive comments, negative comments, neutral/other comments or no comments/haven't thought about it. Two students offered comments that were both positive, negative and neutral/other; nine students wrote comments that were both positive and negative; one student offered comments that were both negative and neutral/other; seven students wrote comments that were entirely positive; and one student commented in a way that was categorised as

neutral, whereas nobody gave comments that were only negative. Two students said that they could give no reasons for their answer (no comments/haven't thought about it).

A total of 18 students offered positive comments to sentence-based segmentation in TM technology. All seem to indicate that they found it a logical and useful way of dividing up the translation task, which, in itself, does not indicate any cognitive changes. However, the fact that many students are aware of the negative sides of this enforced segmentation does indicate a certain change in their cognitive processes. Again, many students mention that TM technology involves a risk of making you oblivious to contextual and functional aspects of the translation. Thus, for instance, student No. 12 mentions that, though your task becomes more manageable, you may be forced to work in a way that changes your usual process:

> "Positive: It's more manageable. Negative: As already mentioned, you tend to focus on each individual segment and risk forgetting about the coherence of the text. It may also be a problem that you cannot see the context of the translation proposal offered by the TM. This can result in wrong translations".

Another student (No. 9) implies that sentence-based segmentation disturbs your natural (cognitive) translation process when you are working with units below the sentence level, which you sometimes do:

> "It's an advantage because you wouldn't be able to work with larger chunks at the same time anyway. It's a disadvantage because you easily lose track of the text as a text. It can also be a disadvantage if you work on segments that are smaller than a sentence, but I found that the concordance facility helps you make up for this".

A comment by another student (No. 6) to qualify her first answer of "I don't know", which was categorised as neutral/other, makes a similar point (here translated closely by us, with explicitations in square brackets):

> "From what I've experienced, I cannot really see that it is an advantage or a disadvantage. You can still view the rest of the text [on the screen], so it's still possible to relate the segment to its context, if you need to [while working] with individual sentences".

9 Preliminary results: What they told us about the impact of TM technology on cognitive processes

We shall now try to conclude on the cognitive changes caused by TM technology that were suggested by our students, using a simplified model of the translation process as an analytical tool. Viewing translation as a recursive and reiterative writing activity (see also Mossop 2000; Breedveld 2002; Jakobsen 2003), we shall assume that translation processes may be categorised as belonging to one of three main phases: planning, drafting and postdrafting. Inspired by Englund Dimitrova

(2010), we shall subdivide the drafting phase into comprehension, transfer and production.

1) The *planning phase* includes activities such as researching the topic and the respective communities of the ST and the TT, interpreting relevant norms and the communicative purpose of the TT (skopos), reading and analysing the ST, choosing a macrostrategy and, in the case of TM-assisted translation, selecting and assessing an available TM. For our purposes, a macrostrategy concerns the translator's choice of an overall plan or a set of principles for carrying out a specific translation task, which tends to be either ST- or TT-oriented (e.g. Schjoldager 2008). Our results seem to imply that students tend to forget about the planning phase. In particular, many students' answers indicate that macrostrategic decisions were no longer part of their translation process. We are aware that this may not be true for professional translators in an authentic situation. (Actually, we are hoping that it is not.)

2) The *drafting phase* concerns the translator's work with the ST itself and, as mentioned above, it may be understood as comprising three subphases:

2.1) The *comprehension phase* comprises the decoding (understanding) of ST segments (words, phrases, etc.), drawing on what the translator has learned during the planning phase and drawing on the translator's own knowledge of relevance for the task. In principle, TM-assisted translation should not differ from human translation in this respect, but many students seem to indicate that comprehension was less thorough than it usually is.

2.2) The *transfer phase* concerns the shift from one language to another, thinking in two languages at the same time, as it were, and making microstrategic decisions. Microstrategies may be defined as a set of procedures that guide the translator's decisions in connection with specific points of a translation task, including both problem-oriented decisions and other decisions. Because of its continuous and automatic offering of translation proposals (matches), we would expect TM technology to affect the translator's cognitive processes in the transfer phase, which was also indicated by many students, who seem to have copied TM proposals rather making their own microstrategic decisions.

2.3) The *production phase* comprises what the translator carries out in relation to the TT itself, including some textual decisions. This phase is generally thought to comprise some on-going revision of the TT. Following Mossop's (2007: 167) distinction between self-revision and other-revision, this on-going revision may be characterised as self-revision. As the TM translator is not only revising text that s/he is writing him/herself, we would expect TM technology to affect the production phase. This impact may be even stronger if the translator is translating by means of a TM comprising aligned and imported texts from other translators, as our students were (section 5, above). Interestingly, our results indicate that many students seem to have spent more time assessing and revising what was offered by the TM, i.e. other-revision, than on revising what they had written themselves 'from scratch'.

3) The *postdrafting phase* comprises a (supposedly) final revision of the TT, which can be carried out by somebody else than the translator, in which case it may be referred to as other-revision (Mossop 2007). As the production phase of TM-assisted translation may be expected to be rather different from that of human translation, so is probably the postdrafting phase: With its enforced segmentation of the ST and its offering of previously translated segments (matches) from other

assignments, you might expect TM translators to carry out a more thorough final revision of the TT. Though the introductory course did not give students an opportunity to carry out any postdrafting as such, many indicated that they expected the technology to change this phase too, saying that they expected the TT to need some textual revision afterwards.

10 Concluding remarks

Seeing that we still lack empirically documented knowledge of how TM technology and other translation software influence translators' cognitive processes, we have presented and discussed the theoretical background, setup and findings of a small-scale pilot study of student-translators' retrospective comments in an online questionnaire survey regarding what they had experienced during an introductory hands-on course in TM-assisted translation.

All 22 students clearly felt that translating with a TM was different from translating without a TM, i.e. human translation. Assuming that translation processes may be categorised as belonging to one of the above-mentioned phases, we suggest that, as far as our student-translators are concerned, the greatest impact of TM technology seems to occur during the drafting phase. Thus, for instance, all students report that TM technology tends to take over the translation process when they uncritically accept whatever is offered by the TM, and many report that, especially because of the sentence-based segmentation, the technology forces them to work in a way that is different from what they are used to when carrying out human translation. More specifically, as far as the three subphases are concerned, it is suggested that the comprehension phase becomes less thorough, that the transfer phase is largely neglected, as microstrategic decisions are generally copied from previous translations, and that the production phase comprises more (other-)revision than actual production. Furthermore, according to many students' answers, the planning phase appears to be almost forgotten, and many also indicate that, because of all these changes in the planning and drafting phases, the postdrafting phase will have to change too, in the sense that textual aspects of the TT must receive more attention.

In view of recent advances within machine translation (MT) combined with an increasingly competitive market, perhaps much professional translation is soon to be carried out as HAMT (Fiederer/O'Brien 2009). While such automation of the translation process will not eliminate human translators altogether, the impact on their cognitive processes is bound to increase considerably. Therefore, to help translators prepare for an increasingly digitalised future, we shall need more empirically founded studies of how they interact with TM and other translation technology.

References

Álvarez, A.M.G.: Students' translation process in specialised translation: Translation Commentary. Journal of Specialised Translation, issue 07, 139-163 (2007)

Biau Gil, José R., Pym, A.: Technology and Translation: a pedagogical overview. In: Pym, A., Perekrestenko, A., Starink, B. (Org.), Translation technology and its teaching. Tarragona, Spain, 5-19, http://isg.urv.es/library/papers/BiauPym_Technology.pdf [Accessed on March 30, 2011] (2006)

Breedveld, H.: Writing and revising processes in professional translation. Across Languages and Cultures 3 [1], pp. 91-100 (2002)

Christensen, T.P.: How to investigate mental processes in translation memory-assisted translation? (forthcoming)

Christensen, T.P., Schjoldager A.: Translation-Memory (TM) Research: What Do We Know and How Do We Know It? Hermes – Journal of Language and Communication Studies 44, 89-101 (2010)

Dam-Jensen, H., Heine, C.: Process Research Methods and Their Application in the Didactics of Text Production and Translation. Trans-kom 2 [1], 1-25 (2009)

Dragsted, B.: Segmentation in Translation and Translation Memory Systems – An Empirical Investigation of Cognitive Segmentation and Effects of Integrating a TM System into the Translation Process. Samfundslitteratur, Copenhagen (2004)

Dragsted, B.: Computer-aided translation as a distributed cognitive task. Pragmatics & Cognition 14 [2], pp. 443-464 (2006)

Englund Dimotrova, B.: Translation Process. In Gambier, Y., Doorslaer, L. (eds.), Handbook of Translation Studies, vol. 1, pp. 406-411. John Benjamins, Amsterdam/Philadelphia (2010)

Ericsson, K. A.: Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In: Ericsson, K.A., Charness, N., Feltovich, P.J., Hoffman, R.R. (eds.), The Cambridge Handbook of Expertise and Expert Performance, pp. 223-241. Cambridge University Press, Cambridge (2006)

Fiederer, R., O'Brien, S.: Quality and Machine Translation: A realistic objective? The Journal of Specialised Translation 11, 52-74 (2009)

Gambier, Y.: Translation strategies and tactics. In: Gambier, Y, Doorslaer, L. (eds.), Handbook of Translation Studies, vol. 1, pp. 412-418. John Benjamins, Amsterdam/Philadelphia (2010)

Garcia, I.: Power-shifts in web-based translation memory. Machine Translation 21, 55-68 (2007)

Garcia, I.: Is machine translation ready yet? Target 22 [1], 7-21 (2010)

Göpferich, S.: Translationsprozessforschung. Stand – Methoden – Perspektiven. Narr, Tübingen (2008)

Holmes, J. S.: The name and nature of translations studies. In: Venuti, L. (ed.), The Translation Studies Reader, pp. 172-185. London/New York (1972/2000)

House, J.: Consciousness and the strategic use of aids in translation. Tirkkonen-Condit, S.; Jääskeläinen, R. (eds), Tapping and mapping the processes of translation

and interpreting: outlooks on empirical research, pp. 149-162. John Benjamins, Amsterdam/Philadelphia (2000)

Hutchins, E.: Distributed Cognition, http://files.meetup.com/410989/DistributedCognition.pdf [Accessed 30 March, 2011] (2000)

Jakobsen, A. L.: Effects of Think Aloud on Translation Speed, Revision and Segmentation. In: Alves, F. (ed.), Triangulating Translation. Perspectives of process oriented research, pp 69-95. John Benjamins, Amsterdam/Philadelphia (2003)

Krings, H.P.: Was in den Köpfen von Übersetzern vorgeht. Eine empirische Untersuchung zur Struktur des Übersetzungsprozesses an fortgeschrittenen Französischlernern. Narr, Tübingen (1986)

Krings, H.P.: Wege ins Labyrinth – Fragestellungen und Methoden der Übersetzungsprocess-forschung im Überblick. Meta 50 [2], 342-358 (2005)

Mossop, B.: The Workplace Procedures of Professional Translators. In: Chesterman, A., San Salvador, N.G., Gambier, Y.. Translation in Context. Selected Papers from the EST Congress, Granada 1998. [Benjamins Translation Library 39] pp. 38-48. John Benjamins, Amsterdam/Philadelphia (2000)

Mossop, B.: Has Computerization Changed Translation? Meta 51 [4], 787-793 (2006)

Mossop, B.: Revising and Editing for Translators. St Jerome, Manchester, U.K./Kinderhook, USA (2007)

Nord, C.: Text Analysis in Translation Theory. Methodology and Didactic Application of a Model for Translation-Oriented Textual Analysis. Second edition. Rodopi GA, Amsterdam/Atlanta (2005)

O'Brien, S.: Eye-tracking and translation memory matches. Perspectives: Studies in Translatology 14 [3], pp. 185-205 (2006)

O'Brien, S.: Processing fuzzy matches in translation memory tools: an eye-tracking analysis. In: Göpferich, S., Jakobsen, A.L., Mees, I.M. (eds.), Looking at Eyes. Eye-tracking studies of Reading and Translation Processing, pp. 79-102. Samfundslitteratur, Copenhagen (2008)

Palumbo, G.: Key terms in translation. Continuum, London/New York (2009)

Schäler, R.: Beyond Translation Memories. In: Proceedings of the Workshop on Example-Based Machine Translation. http://www.eamt.org/events/summitVIII/papers/schaeler.pdf [Accessed 16 December, 2009] (2001)

Schjoldager, A.: Understanding Translation. Academica, Aarhus/Copenhagen (2008)

Schubert, K.: Positioning translation in technical communication studies. Journal of Specialised Translation, issue 11, 17-30 (2009)

The Process of Post-Editing: a Pilot Study

Michael Carl, Barbara Dragsted, Jakob Elming,
Daniel Hardt, and Arnt Lykke Jakobsen*

Department of International Language Studies and Computational Linguistics
Copenhagen Business School, Dalgas Have 15, DK2000 Frederiksberg, Denmark

Abstract. We report on experiments in which manual translation is compared with a process in which automatically translated texts are post-edited. The translations were performed using Translog, a tool for monitoring and collecting keystroke and gaze data. The results indicate that the post-editing process resulted in a modest improvement in quality, as compared to the manual translations. Translation times were lower for the post-editing. It was also found that post-editing involved notable differences in gaze behavior.

Key words: Translation Process, Post-editing, Machine Translation

1 Introduction

The results of empirical research on translators' productivity as they post-edit machine-translated text in comparison with their productivity when they translate text more traditionally, either manually without any special technological support or with the support of a translation memory system, have mostly been inconclusive [10, 9]. There may be many reasons why no very conclusive results have been produced. A major factor may be that translators often object to being asked to improve on a machine's inferior text.

This situation strikes us as necessarily transitional. Not very long ago there was similar resistance among translators to using translation memory (TM) systems, but that has been almost universally overcome, and everywhere the professional translation environment now includes a TM system.

Perhaps a TM system has a more human appearance than an MT system. Both the fact that it is conceptualized as a 'memory' and the fact that its database is a record of human translations, and perhaps also the fact that the human translator has full control of how the translation is constructed, contribute to making this kind of man-machine interaction acceptable and indeed meaningful to the human user.

However, the most recent TM systems now include an MT component so that users of TM systems have the opportunity to interact with the machine

* Thanks to Kristian T. H. Jensen for help with information on the results from his experiments on manual translations of these texts. Also, thanks to our colleagues at Copenhagen Business School for serving as subjects of the experiments.

in a different mode, namely by post-editing text generated not by a human translator but by the machine. This addition of MT to successful TM solutions reflects the widespread view of MT developers that MT, especially statistical MT (SMT), has improved quite radically in recent years and deserves to be more widely used and accepted. In their view, considerable productivity gain could be obtained (a) if post-editing was accepted as a meaningful method of producing a translation and (b) if acceptance was followed up by post-editing training.

In order to properly test such assumptions, we plan to conduct a longitudinal study in order to trace the effect of training on positively motivated translators. The pilot study reported in the present paper lacks this longitudinal dimension, but was undertaken in order to find out how translators with no post-editing training at all would perform when asked to post-edit MT-produced text in comparison with the performance of a group of translators who had translated the same texts manually, without any dictionary or technical assistance. We chose three English news texts which were to be translated into Danish. We specifically wanted to see how post-editing Google Translate versions of the three texts would compare with translating the three texts manually, in terms of the quality of the translations produced, and the time it took to produce them. We also investigated various features of keyboard and gaze activity.

2 Experiments

The manual translation data was elicited in experiments conducted by K. T. H. Jensen in 2008-2009 [8]. In his PhD study of allocation of cognitive resources in translation, he had 24 participants translate a warm-up text and three British newspaper texts, A, B, and C, assumed to be at different levels of difficulty. 12 participants were MA students (one male), 12 were professional translators (three male), all with Danish L1 and English L2. The English source texts averaged about 850 characters and were translated under different time constraints.

In the current experiment, we chose 8 translations from each of the manually translated A, B and C texts which had no time constraints. In our post-editing experiment, we used the same three texts (A, B, and C), and asked 7 translators to post-edit English-to-Danish machine-translated versions produced by Google Translate.

All 7 translators were native Danish speakers. Three of them had professional translation experience, two post-editors had a formal translation background (but no extended professional experience), and one post-editor was a bilingual native Danish speaker with no formal translation education. None of them had significant experience in using CAT tools. Three of the translators had already manually translated the texts 2 years before in the context of the manual translation, but we think that this did not have a measurable impact on the translation performance, given the long lapse of time between these two events and also the different nature of the two tasks.

The post-editing was performed using Translog [3], a tool for monitoring and collecting keystroke and gaze data during translation. Translog consists of two windows: the source text is shown in the top window, and the translator types the translation into the bottom target window. At the beginning of the post-editing experiment, the Source Text (ST) was displayed in the top window, and the Google Translate output was pasted into the target window at the bottom of the screen. These translations were then post-edited by the translators. Table 1 gives an overview of the properties of the manual and the post-edited translations. On average, the post-edited translations were slightly shorter than the manually translated versions, there were many more deletions during post-editing than during manual translation, there are less insertions, and when post-editing, translators used navigation keystrokes and mouse clicks much more often.

Table 1. Averaged keyboard activity data over 7 versions of three post-edited and three manually translated texts from seven translators

	Post-editing						Manual Translation				
	Google	TT len	insert.	delet.	navi.	mouse	TT len	insert.	delet.	navi.	mouse
A text	834	853	221	112	491	12	884	945	61	35	5
B text	863	903	281	127	379	21	949	1089	127	183	6
C text	865	915	181	74	390	13	905	976	66	47	3

The Google translations of the three English texts consisted of A:834, B:863 and C:865 characters, whereas the average length of the post-edited translations was A:853, B:903 and C:915 characters, and the average length of the manual translations was A:884, B:949 and C:905 characters. It is interesting to note that almost all translations (the post-edited as well as the manual translations) were longer than the Google translations.

Note that the number of insertion keystrokes minus the number of deletion keystrokes does not equal the length of the final TT translations, since highlighting a word by using, e.g., the left or right arrow in combination with shift+control would count as one (navigation) keystroke, but the deletion of a highlighted sequence can be achieved by just hitting the delete (or backspace) key once, or by overwriting it with another (sequence of) character. The latter activity would then count as an insertion, rather than (or in addition to) a deletion, even though the highlighted sequence is deleted. The table shows that the usage of the keyboard for post-editing and manual translation is quite different.

2.1 Evaluation of Translation Quality

The quality of each translation was evaluated by seven native Danish speaker evaluators. Four of the evaluators were professional translators from the CBS

teaching staff and two evaluators had at least 3-5 years of translator training at CBS, and again one evaluator had no translator background, but was a Danish native and fluent English speaker. Each evaluator was presented with a source sentence together with four candidate translations. In each case two translations had been produced using manual translation and two had been produced using post-editing. The presentation order was randomized. Evaluators were instructed to order (rank) the candidate translations from best to worst quality, with ties permitted. This method is frequently used for the evaluation of MT system output [1, 2], but is less familiar in evaluating human-produced translations.

Each sentence was ranked by at least two evaluators. Also, each evaluator was presented with two repeats of a source sentence together with the same four proposed translations. This was done to permit calculation of inter-coder and intra-coder agreement.

Inter-coder agreement: agreement is defined with respect to a given pair of candidate translations for a given source sentence. That is, for two coders c_1 and c_2, we have a source sentence s, and two candidate translations t_1 and t_2, both of which received rankings from coders c_1 and c_2. We say that the two coders agree if their rankings for t_1 and t_2 stand in the same relation. In other words, there is agreement if one of the following three conditions holds:

1. $rank(c_1, t_1) > rank(c_1, t_2)$ AND $rank(c_2, t_1) > rank(c_2, t_2)$
2. $rank(c_1, t_1) < rank(c_1, t_2)$ AND $rank(c_2, t_1) < rank(c_2, t_2)$
3. $rank(c_1, t_1) == rank(c_1, t_2)$ AND $rank(c_2, t_1) == rank(c_2, t_2)$

There were a total of 125 pairs which were evaluated by two coders. Of these, there was agreement in 57 of them, or 46%. Assuming that chance agreement is 33%, we compute a Kappa of 0.188. While this is better than chance, it is considered Slight agreement [6]. This is consistent with the general feeling of evaluators that ordering candidate translations was a difficult task.

Intra-coder agreement: there were a small number (14) of repeat sentences, where the same coder was presented with identical pairs of candidate translations. Here six were in agreement (42.8%), for a Kappa of 0.147.

The data show that intra-coder agreement is even lower than inter-coder agreement. The fact that agreement is so low suggests to us that the assessment of translation quality was simply too difficult.

3 Analysis

3.1 Translation Quality

As mentioned above, evaluators ranked 4 translations of one source sentence at a time, where 2 translations were taken from the manual translations and 2 from the post-edited translations. Subsequently, each sentence was scored according to how often it was ranked better than the translations of the other mode. For instance, if a post-edited sentence was ranked better than one manual translation and worse than the other manual translation, it received a score of 1. If a manual

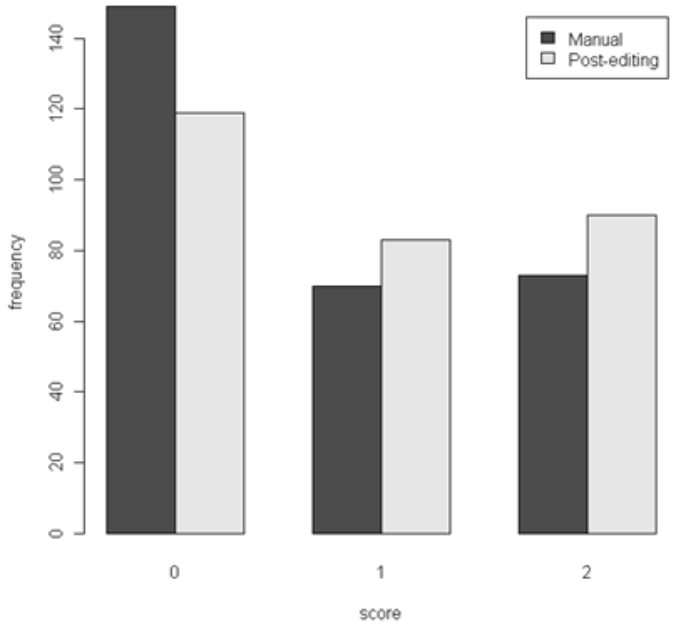


Fig. 1. A comparison of the frequency of evaluation scores of the manual translations and post-edited sentences. Higher scores are better.

Fig. 2. Average scores of post-edited and manually translated texts: A-texts (left) B-texts (middle) and C texts (right).

translation was ranked better than both post-edited translations, it received a score of 2, and if it was not better than any translations of the other mode, it received a score of 0. Accordingly a sentence can have one of 3 scores, where higher scores represent better rankings. The score of a manual or post-edited translation was as follows:

- 2: better than both of the other-mode translations
- 1: better than one of the other-mode translation
- 0: not better than any of the other-mode translations

The distribution of sentence evaluation scores is shown in Figure 1. The graph indicates that the post-edited translations are judged to be better than the equivalent manual translations. The difference is not quite significant, according to the Wilcoxon signed-rank test ($p = 0.05053$). It is however an interesting result that translation quality does not seem to be reduced by the integration of machine translation in the translation process.

The average scores over all the sentences in the post-edited and the manually translated A, B and C texts are shown in Figure 2 below.

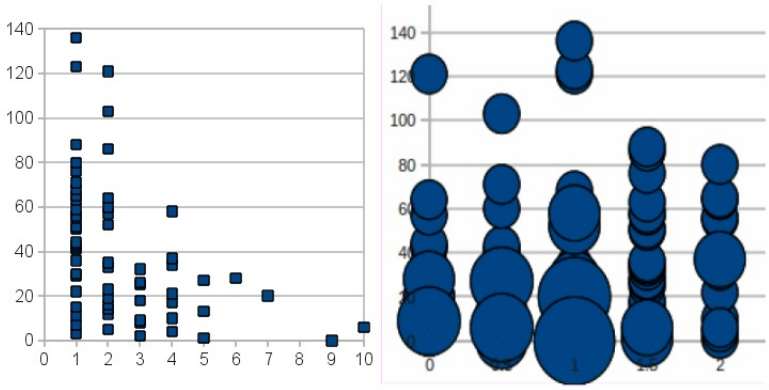


Fig. 3. Correlation between the number of edit operations (vertical axis) and correlation with translation score (right) and number of sentences (left).

3.2 Edit Distance and Translation Quality

The 21 post-edited texts consisted of 133 sentences (8 sentences in the A text, 6 sentences in the B text and 5 sentences in the C text). For each of the 133 post-edited sentences the edit distance was computed. The edit distance indicates the minimum number of changes between the Google translation and its post-editing version. Figure 3 shows that there are between 0 and up to 136 edit operations per sentence. The distribution of edit operations is shown in

figure 3 (left): As can be expected, there are only a few sentences with many operations, and there are more sentences with few operations, e.g. 1 sentence with 136 edit operations, but 10 sentences with 6 operations. We also computed the correlation between the average score of the post-edited sentences, as described in section 3.1, and the number of edit operations per sentence. Since there were two scores from two different evaluations per post-edited sentence, we computed 266 correlations between edit distance and translation score, which are shown in figure 3 (right). Bigger bubbles represent more occurrences of the operation/score relation. Surprisingly, there is no correlation between the score of the post-edited sentence and the number of edit operations, indicating that more post-editing does not necessarily lead to better translations.

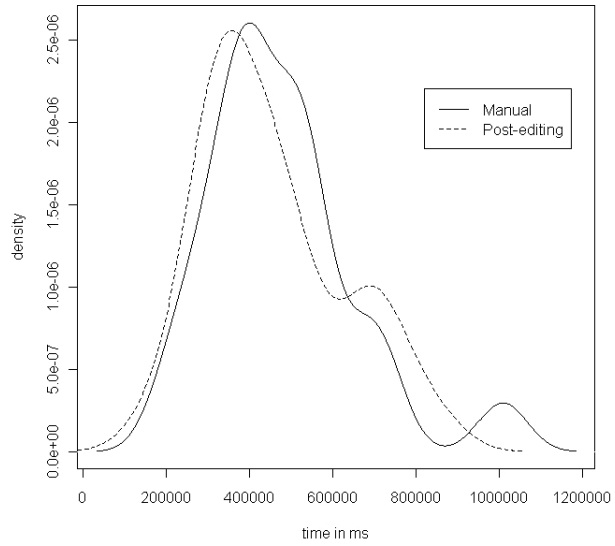


Fig. 4. The estimated distribution of the time spent manually translating and post-editing a text.

3.3 Time

One of the most obvious reasons for engaging in post-editing is the desire to save time. Figure 4 shows the estimated distribution of the time spent on manually translating a text compared to post-editing a text. The two distributions are quite similar, but there is an indication that post-editing may lead to some time saving, though not a significant difference ($p = 0.7118$). This may partly be due to the low number of participants in the tasks. On average a text was post-edited in 7 minutes 35 seconds, while a manual translation took 7 minutes 52

seconds. In this context it should be noted that while all manual translators had experience in translating, none of the post-editors had experience post-editing or using CAT tools. We expect that more post-editing experience will yield a margin of time saving.

3.4 Gaze

We recorded participants' gaze activity in the manual translation and the post-editing task. In the manual translation task, we used a Tobii 1750 eye tracker, which runs at a frame-rate of 50 Hz, and in the post-editing task, we used a Tobii eye tracker which runs at 60 Hz.[1] Both are remote eye-trackers which use binocular eye tracking. The texts were presented on a TFT display with a resolution of 1280x1024 pixels. Participants sat 60 cm from the screen, and were calibrated with a standard 9-point grid.

A basic assumption in eye movement research is that "the eye remains fixated on a word as long as the word is being processed" [5]. Gaze duration is thus taken to signal the cognitive effort associated with processing a particular item, and fixations in reading tend to be longer on items requiring effortful processing, for instance less frequent words, words containing spelling errors, ambiguous words and words which are inappropriate in a given context [7]. Evidence from reading studies suggest that the majority of words in a text are fixated during reading, but that some words are skipped and some words are fixated more than once [11]. The number of regressions has been found to increase as texts become more difficult. Conversely, the eyes are less likely to fixate highly predictable words [7]. In translation, fixation counts are generally higher than in reading, with regressions occurring more frequently [4], and the average fixation count per minute has been found to be significantly higher in complex texts than in simpler texts. Gaze times have similarly been used as indicators of particular items requiring larger cognitive effort [12].

Total gaze time on both areas of the screen (ST and TT) was approximately the same in the two tasks, 263,938ms in the manual translation task and 295,508ms in the post-editing task on average across the three texts. Since the average total task time was lower in the post-editing task (see section 3.3), a higher proportion of time was spent looking at the screen. The slightly larger gap between total task time and total gaze time in the manual translation task may indicate that more time was spent looking outside the screen, most likely at the keyboard, when the translation was produced manually. Another intuitive explanation may be that when producing a translation from scratch, translators may stare off into the space as they await inspiration - something they would not do in a more "mechanical" post-editing task. However, off-screen fixations were not recorded in any of the tasks and the distribution between gaze time and task time will need to be investigated further in future studies.

We analysed the distribution of gaze activity in Translog's ST window vs. its TT window in the two tasks, using the measures fixation count and total gaze time, to investigate which of the two areas attracted most visual attention. In the manual translations, the number of fixations was distributed more evenly

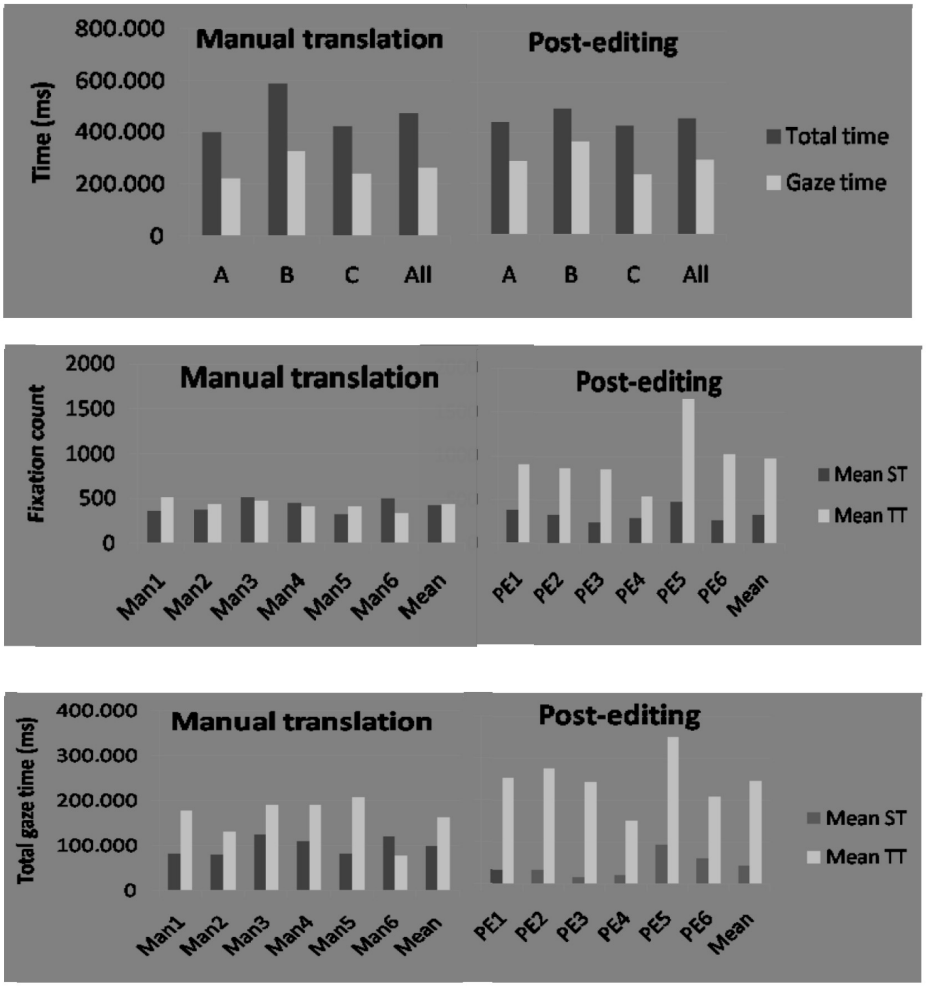


Fig. 5. Comparision of post-editing vs. manual translation behaviour with respect to 1. total translation time vs. total gaze time (top) 2. mean fixation counts on the source vs. target text (middle), 3. total gaze time on the source vs. target text (bottom)

on ST and TT than in the post-editing task. The average fixation count in the manual translation was 420 on the ST area of the screen and 434 on the TT. In the post-editing task, participants fixated the ST 334 times on average against 975 fixations on the TT. Means for six participants in each group are shown in Figure 5. The total gaze time was higher on the TT area than on the ST area in both tasks: 163,364ms on average on the TT against 100,575ms in the ST in the manual translation task, and 247,226ms on average on the TT against 43,055ms on the ST in the post-editing task.

Differences between fixation count and total gaze time in terms of ST/TT distributions show that participants had longer average fixation durations on the TT area in both tasks (Figure 5), but the tendency for most visual attention to be on the TT was most pronounced in the post-editing task, and both the fixation count and the total gaze time on the TT were significantly higher in post-editing than in manual translation according to an unpaired two-sample t-test ($p < 0.01$).

Editing SMT output thus apparently requires a higher TT reading and rereading effort than manual translation. The gaze activity in the post-editing task reflects a process, it may be assumed, of first reading a segment of raw SMT output, then comparing this against a segment in the ST that it is a translation of, and then possibly correcting the machine-translated output and reading the corrected version one or several times. In manual translation, TT gaze activity simply involves monitoring and possibly correcting one's own manual translation output, a process which, based on the eye movement data, requires less effort.

The ST was consulted more frequently (see Figure 5, middle, the dark bars in fixation count) and in particular attracted longer fixations (see Figure 5, bottom[1]) when participants produced a translation manually than when they post-edited SMT output. The number of fixations on the ST was not very different from the post-editing task (it was slightly higher, but the difference was not significant according to an unpaired two-sample t-test, $p = 0.07998$), but the duration of each fixation was longer on average, leading to significantly longer total gaze time on ST during manual translation ($p < 0.01$). This indicates that a different type of ST comprehension is involved in a post-editing task than in manual translation. Manual translation seems to imply a deeper understanding of the ST, requiring more effort and thus longer fixations, whereas in post-editing, the ST is consulted frequently but briefly in order to check that the SMT output is an accurate and/or adequate reproduction of the ST. Also, it may be assumed that in post-editing, the translator reads the SMT output in the TT window before consulting the ST, whereas in manual translation, the ST is naturally attended to first. Note, however that none of our translators had experience in post-editing. The observed behaviour might change dramatically as the translators become more acquainted with the task. This will have to be investigated further.

[1] The total gaze time is the product of fixation count and fixation duration.

4 Conclusion

MT technology has been developing rapidly in recent years, and many have suggested that it can have a major impact on productivity in the translation process, when followed by a post-editing process. However, there is a widespread belief among translators that MT has a negative effect on translation quality, and there is also skepticism that post-editing MT can be done as quickly as ordinary translation. The present study represents a preliminary attempt to address these issues. We found striking differences in both the keyboard and gaze activity of translators when doing post-editing as opposed to manual translation. Furthermore, we found that translation speeds were on average somewhat faster with post-editing, together with a modest increase in translation quality.

These results provide indications that post-editing MT may indeed be shown to have a positive effect on productivity. Given the small scale of the current study, however, no firm conclusions can yet be drawn. Furthermore, our results show that the evaluation of translation quality was extremely difficult. We believe that this difficulty derived in large part from the fact that evaluators were asked to perform relative evaluations, and nearly all the translations were of very high quality. In subsequent studies, we intend to address this issue by asking evaluators to perform more traditional categorical evaluations, in particular asking them to focus on clearly identifiable problems in translation quality. Our results, preliminary as they are, are consistent with a widespread belief that reductions in translation time are possible by doing post-editing. In subsequent work we will pose the question: under what conditions are such reductions possible without a negative effect on translation quality?

References

1. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. 2007. (meta-) evaluation of machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 136–158). Association for Computational Linguistics.
2. Elming, J. 2008. Syntactic Reordering in Statistical Machine Translation. Ph.D. Thesis. Copenhagen Business School, Denmark.
3. Jakobsen, A. L. 1999. Logging target text production with Translog. Copenhagen Studies in Language, 24. pp. 9–20.
4. Jakobsen, A. L. and Jensen, K. T. H. 2008. Eye movement behaviour across four reading tasks. In Göpferich, S., Jakobsen A. L. and Mees, I. M. (eds) Looking at eyes. Eye-tracking Studies of Reading and Tranlsation Processing. Copenhagen Studies in Language 36. 103-124.
5. Just, M.A. and Carpenter, P.A. 1980. A theory of reading: From eye fixations to comprehension. Psychological Review. Vol. 87. No. 4. 329-354.
6. Landis, J. Richard and Koch, Gary G. 1977. The measurement of observer agreement for categorical data. Biometrics, 33, 159–174.
7. McConkie, G.W. and Yang, S. 2003. How cognition affects eye-movements during reading. In Hyönä, J., Radach, R. and Deubel, H. (eds) The mind's eye: Cognitive and applied aspects of eye movement research. 413-427.

8. Jensen, K. T. H. (to appear in 2011) Allocation of cognitive resources in translation: an eye-tracking and key-logging study. Ph.D. Thesis. Copenhagen Business School.
9. Krings, H. P. 2001 (tr. and ed. by G. S. Koby et al.) Repairing Texts. Empirical Investigations of Machine Translation Post-Editing Processes. Kent, Ohio: Kent State UP.
10. O'Brien, S. 2007. Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output. Across Languages And Cultures, 7, 1, pp1–21 Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin Vol. 124. No. 3. 372–422.
11. Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin Vol. 124. No. 3. 372–422.
12. Sharmin, S. Špakov, O. Räihä, K. J. and Jakobsen, A. L. 2008. Where on the screen do translation students look while translating, and for how long? In Göpferich, S., Jakobsen A. L. and Mees, I. M. (eds) Looking at eyes. Eye-tracking Studies of Reading and Tranlsation Processing. Copenhagen Studies in Language 36. 31-51.

Patterns of shallow text production in translation

Michael Carl

Department of International Language Studies and Computational Linguistics
Copenhagen Business School, Dalgas Have 15, DK2000 Frederiksberg, Denmark

Abstract. The depth and timing of source text understanding during translation production is controversial in translation process research. Two theories seem to compete: defenders of the *deep/alternating* school assume that translators proceed in cycles of comprehension-transfer-production, while other translation scholars suggest that translations may be produced in a fashion of *shallow* and *parallel* comprehension and production. We assess these hypotheses by comparing text production activities in a copying task and in a translation task. Text copying constitutes a kind of baseline for text production in general as we can assume that any other form of text production (including translation) requires more time and effort than merely text reproduction. Surprisingly, however, we observe similar patterns of keystroke behaviour in copying and translation.

1 Introduction

Translation scholars disagree to what extent translation requires a *deep* or *shallow* understanding of the source text (ST), and to what extent translation is a stratificational or a parallel ST comprehension → TT production process. Craciunescu et al. (2004), for instance, claim that "the first stage in human translation is complete comprehension of the source language text". Only after this complete (i.e. *deep*) comprehension is achieved can the translation be produced. Similarly Gile (2005), suggests a stratificational translation process model, in which a translator iteratively reads a piece of the ST and then produces its translation: First the translator would create a "Meaning Hypothesis" for a ST chunk (i.e. a Translation Unit) which is consistent with the "context and the linguistic and extra linguistic knowledge of the translator" (p. 107) for which then a translation can be produced.

Also Angelone (2010) supports that translators process in cycles of comprehension-transfer-production. Uncertainties of translators could be attributed to any of the comprehension, transfer, or production phases, and it is claimed that "non-articulated indicators, such as pauses and eye-fixations, give us no real clue as to how and where to allocate the uncertainty" [p.23]

Some scholars challenge these views, stating that translation processes are based on a *shallow* understanding of the ST and that ST understanding and TT

production can occur in *parallel*. Ruiz et al. (2008) investigate theories of translation along the lines *shallow/parallel* and *deep/alternating*.[1] They find "code-to-code links between the SL and TL at least [on] the lexical and syntactic level of processing", and assume a parallel process, where "the translator engages in partial reformulation while reading for translating the source text". They come to the conclusion that translators switch between the two modes, but more often chose the shallow/parallel one. Also for Mossop (2003), there exist "direct linkages in the mind between SL and TL lexicogrammatical material, independent of 'meaning'": The translator "automatically produces TL lexical and syntactic material based on the incoming SL forms".

We investigate these hypotheses from an empirical angle by analyzing the interaction of reading and text production activities in a text copying and a translation tasks. We take it as uncontroversial that a copyist, in contrast to a translator, may proceed in a shallow/parallel manner: (1) apart from lexical encoding and decoding, text copying does, in theory, not require any deep ST (or TT) understanding (2) reading and writing processes can occur to a maximum amount in parallel during text copying, since no cognitive effort is required for lexical transfer, for syntactic reordering, or for revision. Copying speed would thus essentially depend on the typing skills of the copyist.

The *deep/alternating* hypothesis implies that we can see a clear distinction between reading and writing activities so that the typing speed of a translator is reduced due to the need to first understand the ST passage before starting to type in its translation.

In this paper we compare the typing activities of the copying task with typing activities in translation production and observe, surprisingly, the same patterns. We show instances of typing activity in unchallenged translation which resembles text copying into another language.

Our investigation is based on a collection of activity data from two different tasks. First, an English text of 168 words had to be re-typed (copied) by 10 experienced L2 English translators. A second English text of 160 words is the basis for two translation examples which are discussed in the first part of section 3 and which represent unchallenged and smooth translation progression. This text was translated by two experienced translators. All translation examples are English texts translated into Danish by experienced translators (more than 8 years of professional experience).

We record keystroke and gaze movements during the copying and translation tasks using the Translog software. Translog is a data acquisition software (Jakobsen, 1999). It consists basically of two windows which horizontally divide the screen into two halves; the top window plots the ST, and the bottom window is an editor in which the translation is to be typed. It is possible to register keystrokes and gaze activities in Translog, which are collected and can be replayed, and visualized in progression graphs as in figures 1, 2, and 3 below.

We compare typing and gazing behaviour during the translation and text copying tasks. We first introduce a cognitive model of text copying in section

[1] Respectively horizontal/parallel vertical/serial in in their terminology.

2 and compare our empirical copying data with the predictions of that model. In section 3 we compare the copying behaviour with patterns observed during challenged and unchallenged translation, and section 4 discusses the findings and draws conclusions.

2 Text Copying

In this section we introduce a cognitive model of text typing. We will then illustrate the model with a typing example from our data. In section 3 we will compare these experimental findings with examples taken from a translation session.

2.1 A cognitive model of typing

John (1999) suggests a three step model for typing: First a perceptual operator perceives a written word. Then a cognitive operation retrieves the spelling of the word from long-term memory, and finally a motor operator finds a key on the keyboard and hits it. John makes a distinction between copying of single (sequences of) characters and more complex symbols. For the more complex symbols, like words and syllables, a cognitive operator is required to retrieve the spelling of the word from long-term memory and to initiate the typing of each character. Hence, between the perception of a word and its typing an encoding (perception) and decoding (memory retrieval) of the symbols is performed.

With the assumption that a skilled typist produces about 30 words per minute, she comes to the following figures:

1. A perceptual operator reads a word of about six letters and encodes it in 340 msec
2. Next, another cognitive operator retrieves a spelling list of the word from long-term memory, serializes it, and trigger the typing of each character. This operation has a cycle time of 50 msec.
3. Finally, the motor operator needs 230 msec to type a character on an alphanumeric keyboard at a rate of about 30 words per minute

John assumes that each of the operators works serially in themselves (only one keystroke can be processed at any one time) but that they can work in parallel with each other, with the following seriality restrictions that:

- perception has to be completed before getting the spelling list from long-term memory and before initiation of a character can begin.
- once a character is initiated with a cognitive operator, the motor operator cannot be stopped.
- the perceptual processor stays three words (i.e. chunks) ahead of the cognitive processor.

John uses this model to analyze which of the three operations (reading/retrieval/typing) are the limiting factors in text copying. She finds that the overall typing speed depends to the largest extent on the time needed for motor activity, rather than for perception or cognitive control. In line with other researchers, John assumes that the 50msec of cognitive cycle time are constant and hence the typing skills are often the limiting factors in text production.

2.2 Copying an easy text

To illustrate the analysis of a typing process, we reproduce an example from our experimental data. A text of 9 sentences and 168 words had to be copied by 10 English L2 speaker using the Translog tool. Keyboard and gaze activities were recorded during the copying process. Example 1 shows a fragment of the 3rd sentence of the text.

Example 1:

> The rise in unemployment has spattered a once-profitable business with red ink.

The sentence consists of 13 words (including sentence final dot) with 80 characters (including inter-word blank spaces). One of the copyists has copied the sentence in 21 seconds with 5 typos. Figure 1 shows the copy-progression graph: the vertical Y-axis plots the original sentence which was to be copied; the horizontal X-axis represents a time line in msec. Fixations on source text words are marked by a blue cycle. Typing activities consist of text insertions and deletions (in red).

The figure shows a time fragment of 21 seconds between msec. 58000 to ca. 81000 in which the above sentence is copied. At the beginning, the typist first gazed at the two words "The" and "rise" before starting typing. Two typos occur in the first word when reproducing "The". These typos were immediately corrected. Perrin (2003) suggests a short-hand form to represent writing activities, where correction are represented in square brackets. In this notation, the typing pattern would be represented as: "Th[i-][r-]e" which is read as follows: First the typist writes "Thi-".[2] Then "i-", the blank space (i.e. "-") and the "i" are again deleted, then "r-" is typed and again deleted, until finally the correct "e" is typed. There are thus 4 correcting keystrokes in the production of "The". The typist goes then on immediately with the typing of "rise", without looking back into the source text. There are two fixations just before msec 62000, one on "rise" and one on "unemployment" the latter one while already typing "in". From the progression graph it appears that the word "in" was actually not looked at – however, it is likely that this word was in the parafoveal scope of the fixation on "rise".

The copy activity then goes on smoothly. There are two more typos and deletions, but the typist seems to copy the text without much hesitation, looking

[2] The blank space is represented as a dash "-" in the graph and in figures below.

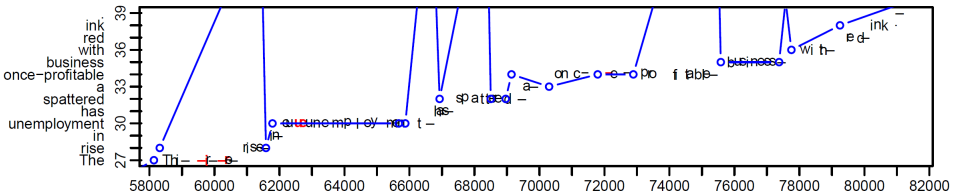


Fig. 1. A progression graph for unchallenged text copying

in general only one word ahead in the source text. In some cases the typist verifies the spelling of a word that is being typed (e.g. "spattered" around time stamp 69000) and in other instances she already scans the next word while still typing the previous word (e.g. also "spattered" around 67000 ms).

According to John's model, a skilled typist would need for keying the 80 characters 80 * 230msec plus an initial 390msec for reading the first word, 340msec for perception of the first word plus 50msec for retrieval of the spelling list. Counting each of the 7 typos as 2 keystrokes (one for insertion and one for deletion), the predicted typing time, according to John's model, amounts to approximately 22 seconds. Compared with the measured typing time of 23 seconds, the model predicts pretty well with an error rate of less than 5%. While the model, thus, seems to be quite exact for predicting the overall time needed when typing activities go smoothly, it does not seem to be so precise for predicting the gaze activities and the structure of the gaze/keystroke coordination: John's model predicts a consistent three words look-ahead. However, figure 1 shows that in many cases only one word is looked ahead from the position that is currently being copied. In addition, longer (or more difficult) words - as in the case of "unemployment" and "once-profitable" - may trigger re-fixations, and in some short words are not fixated at all, which is not predicted in Johns copying model.

3 Translation

In this section we look at translation activities. We distinguish between alternating and parallel activities. The term "alternating" is used when a translator at any one time either reads (the ST) or writes (the TT). During "parallel" activities, the translator does both, reading and writing at the same time.

3.1 Parallel reading and writing

Figure 2 shows an example of parallel translation activities. It represents a translation progression graph for the English source sentence in example 2 into Danish:

Example 2

> Police officer Chris Gregg said that Norris had been acting strangely around the hospital

Danish translation:

> P[i]olitiins[ep]pekt[rør]ør Chris Gregg sagte at Norris havde opført sig sært på h[i]o[p[si]s]sp[o]italet

Figure 2 shows a time fragment of 28 seconds (seconds 149-177) in which the translation in Example 2 is produced. The Danish translation consists of 12 words with 79 characters and 12 typos.

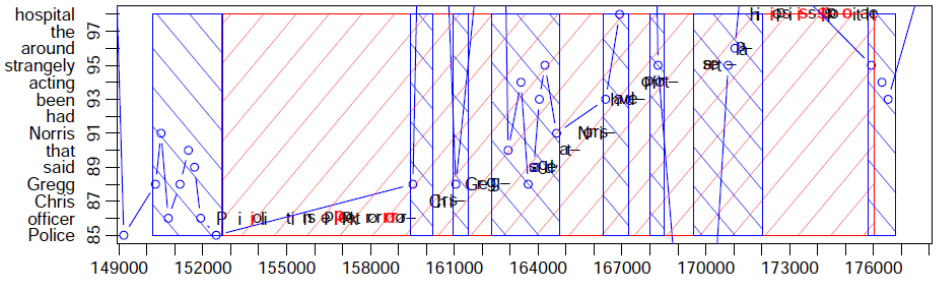


Fig. 2. The translation progression graph shows parallel reading and typing activities

As in the previous copy-progression graph (figure 1), the vertical axis in figure 2 plots the ST words while the horizontal axis represents a time interval in which the translation is produced. The (blue) dots are ST fixations and the downwards hatched boxes represent "fixation units". A fixation unit (FU) consists of a sequence of coherent fixations on the ST, where no pauses of more than 400ms occur between successive fixations. A FU, thus, represents a reading pattern of a ST chunks in which (presumably) the perception and encoding of the ST words take place. There are several FUs in the translation progression graph in figure 2, two of which represent extended reading activity. The first one occurs at the beginning of the sentence, during the time stamps 149 and 152. The translator's eyes move back and forth in the chunk "Police officer Chris Gregg said that Norris". After this the translation "Politiinspektør" is typed including a number of typos, which are immediately corrected (deleted characters are in red). The following production of the proper noun goes on smoothly.

The other extensive FU occurs between seconds 162 and 165. Here the fragment "Gregg said that Norris had been acting strangely" is read by jumping back and forth in the chunk. In contrast to the first FU, this reading activity occurs in parallel while already typing the translation of "Gregg said that".

According to John's typing model, a typist would need 24 seconds to key in this sentence. If we subtract the 3 initial seconds in which the sentence was initially scanned, we measure 25 seconds translation production time vs. 24 seconds predicted typing time. The translation was thus approximately produced at the speed of an expert copyist, also with a predicted error of less than 5%. That is, the additional cognitive effort of the translation activity took place in parallel with the typing activity and did not require additional time. Johns' model for copying apparently also correctly predicts expert translators' typing speed in an unchallenged translation situation. However, gaze behaviour (and thus mental activities) are different in translation and in copying.

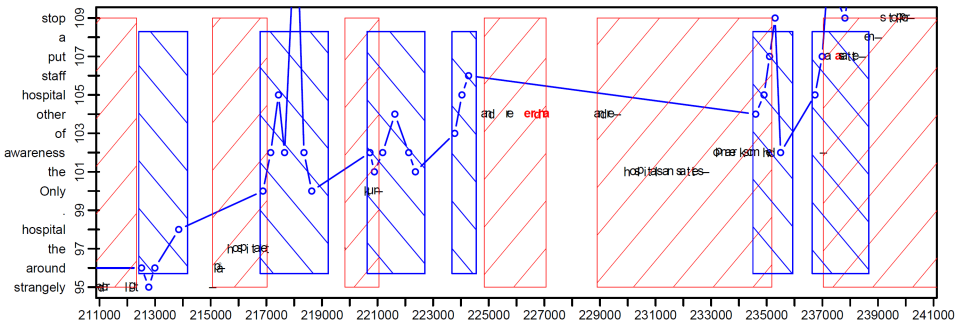


Fig. 3. The translation progression graph shows alternating reading and text production.

3.2 Alternating reading and writing

Figure 3 shows an example of mainly *alternating* translation activity, where the translator is either reading a ST segment or writing a piece of the TT. Figure 3 plots the translation progression of a sentence-final segment and the first words of the next sentence. The translator reads only a few words ahead of what she is currently translating. The produced translation is shown in example 3:

Example 3

> strangely around the hospital. Only the awareness of other hospital staff put a stop ...

Danish translation:

> underligt på hospitalet, kun [andre]andre hospitalsansattes opmærksomhed [a]satte en stopper ...

There is some reading activity (seconds 217-220) before the translator starts translating the second sentence. The two English sentences are collapsed into

one Danish sentence, connected by a comma. There is another scanning phase closely following the first one between seconds 221-225 just after typing "kun", which is the translation of "only". That is, the translator typed "kun" and presumably only then developed a strategy for reordering the translation of "the awareness of other hospital staff". Note the inversion into "other hospital staff[GEN] awareness" in the progression graph.

The translation in example 3 consists of 11 words and 82 characters and was typed in 30 seconds. Including the 6 typos - "andre" was first typed then deleted and then again typed and later "a" was typed and deleted – the typist model predicts 22 seconds to complete the typing. That is, while there might be approximately the same amount of ST reading in a parallel and in an alternating mode, the overall translation time is significantly longer in the alternating mode that in a parallel mode, since activities occur sequentially.

4 Discussion and Conclusion

In this paper we empirically underpin a hypothesis of Mossop (2003) and Ruiz et al (2005), that translation production may be based on a shallow understanding of the ST.

We compare two experimental settings, a copying task and a translation task. We record keystroke and gaze movements using the Translog software. In the copying task, a copyist reads an English text and types the same text on a keyboard, while in the translation task the text is translated into another language (Danish). In the translation task we observe both, "parallel" and "alternating" text production.

We find that text copying and translation activities may resemble each other in terms of typing speed and the number of fixations on the ST – the distribution of fixations is however different in both tasks. Our data show that translators look further ahead into the ST than copyists, both during parallel and alternating text production. Carl and Kay (2011) show that experienced translators operate more frequently in a parallel manner, while translation students resort frequently to the alternating mode.

Reading is far less steady than writing, the eyes jump over two sometimes three words, back and forth, until a piece of text is sufficiently understood to start typing out a translation. These reading patterns resemble in the parallel and in alternating mode. Whereas the alternating mode implies that the translator is either involved in ST understanding or in TT production, the observed reading patterns of 3 to 5 words ahead of the current point of text production does not suggest a deep or "complete" understanding of the ST.

From our examples we thus conclude that translators (of these texts) proceed preferably in a shallow mood which rather resembles text copying than a full text understanding.

From previous investigations it seems that more experience allows translators to work similar to a typist. A translator will aim at producing translations with minimal effort and minimal cognitive workload. Obviously a translator must first

read a ST passage when producing its translation, but usually s/he will try not to do more than that. That is, for producing the next TT word a translator will ideally and whenever possible only consult one (or a few) ST words, just enough to go on with the text production. Whenever the source and target languages are close in terms of conceptual and syntactic structure there will be a minimal lapse of time between reading of a ST words and the production of the translations. In such instances we are likely to observe a linear, almost word-for-word translation production where the typing activity occurs immediately after a ST word has been read. This translation pattern resembles those of figure 2. We take it that mental buffering and workload is minimized in this setting and productivity will basically depend on the typing skills and speed of the translator.

References

1. Eric Angelone. Uncertainty, uncertainty management and meta cognitive management in the translation task. In *Translation and Cognition*, pages 17–41, Amsterdam, 2010. John Benjamins.
2. Michael Carl and Martin Kay. Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators. *Meta*, page forthcoming, 2011.
3. Daniel Gile. *La Traduction. La comprendre, l'apprendre.* Presses Universitaires de France, Paris, 2005.
4. Gyde Hansen, editor. *Probing the process in translation: methods and results*, volume 24 of *Copenhagen Studies in Language*. Copenhagen: Samfundslitteratur, 1999.
5. Arnt Lykke Jakobsen. Logging target text production with Translog. In *[4]*, pages 9–20, 1999.
6. Bonnie E. John. Typist: a theory of performance in skilled typing. *Hum.-Comput. Interact.*, 11(4):321–355, 1996.
7. Brian Mossop. An Alternative to 'Deverbalization'. Technical report, York University, 2003.
8. C. Ruiz, N. Paredes, P. Macizo, and M. T. Bajo. Activation of lexical and syntactic target language properties in translation. *Acta Psychologica*, 128(3):490–500, 2008.

Modeling (Un)Packing of Meaning in Translation: Insights from Effortful Text Production

Fabio Alves[1], Adriana Pagano[1], Igor L. da Silva[1]

[1] Faculdade de Letras, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627 Pampulha – Belo Horizonte/MG 30575-020 – Brazil

Abstract. This paper introduces a proposal aimed at enriching the framework of translation process research with a systemic-functional linguistics perspective for investigating instances of effortful translated text production. Drawing on the concept of grammatical metaphor and its potential for modeling both monolingual and multilingual text production, it examines ongoing meaning construction in translation as a special type of language processing which involves unpacking and repacking meanings construed in the target texts upon reading the source text. By analyzing logs recorded through key logging and eye tracking, we attempt to investigate phenomena that can shed a light into human translators' cognitive processes and which are potential sources for modeling meaning construction at play during the translation process.

Keywords: Modeling of Human Translation, Empirical-Experimental Methodology, Meaning Construction, Translation Process Research, Effortful Text Production.

1 Introduction

In the literature on translation process research, pauses and recursiveness as recorded through key-logging software have been pointed out as indicators of effortful meaning production in translation tasks [1, 2]. More recently, eye tracking has been incorporated in the methodology used by translation process studies [3], whereby data obtained through key logging, eye tracking and verbal protocols are triangulated to both illuminate the translator's behavior during task execution and identify instances of text production that constitute translation problems [4].

Concomitantly, theory-informed text analysis [5, 6] has sought to approach real-time text production as captured in translation tasks in order to seek possible motivations for those instances of effortful production signaled by pauses and recursiveness, envisaging an integration of particular patterns of gaze trajectory and eye fixations into the analysis. One such theory supporting text analysis is systemic functional linguistics [7, 8], which offers a comprehensive approach to meaning making within the context of multilingual text production, its conceptualization allowing for modeling language production in translation. In fact, its architecture is particularly suitable to approach translation process research, since one of the dimensions it adopts to examine language, namely the *logogenetic* one, contemplates the unfolding of discourse, where local decisions are made against the background of

more global orientations taken by the translator on the basis of the values for context configuration adopted by him/her.

In this paper, we propose to enrich the framework of translation process research with a systemic-functional linguistics perspective for examining linguistic phenomena that can be observed to be taking place in instances of effortful translated text production (as revealed by particular patterns of pauses, recursiveness, progressive and regressive fixations within the text, and gaze trajectory across the source and target texts seen as two distinct areas of interest). Drawing on the concept of grammatical metaphor [7] and its potential for modeling both monolingual and multilingual text production [9], we aim at examining ongoing meaning construction in translation as a special type of language processing which involves unpacking and repacking meanings construed in the target texts upon reading of the source text. By analyzing logs recorded through key logging and eye tracking, we attempt to investigate phenomena that can shed a light into human translators' cognitive processes and which are potential sources for modeling meaning construction at play during the translation process.

2 Theoretical Underpinnings

Within the framework of systemic functional linguistics, *logogenetic* instantiation of text has been frequently studied within monolingual text production, with particular focus on phenomena involving recapitulation of higher rank units in language, such as clauses, in lower rank units, such as groups and words [10]. This is accounted for through the concept of grammatical metaphor, which names a phenomenon "whereby a set of agnate (related) forms is present in the language having different mappings between the semantic and the grammatical categories" [7]. This can be seen in the example below:

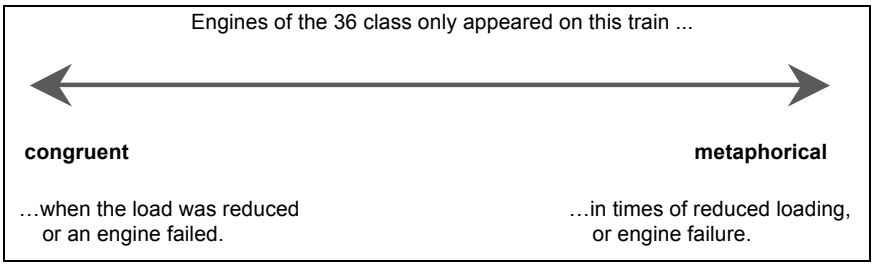


Fig. 1. Example of congruent and metaphorical wordings
Source: Halliday & Matthiessen (1999: 235).

The less metaphorical wording on the left "when the load was reduced or an engine failed" construes meaning through a hypotactic clause subordinated to the main one "Engines of the 36 class only appeared on this train". The meanings construed by this subordinate clause can also be construed through a more metaphorical wording on the right, here through a circumstantial adjunct, where agency is less explicit than in its more congruent wording. Nominalization and adjectivization pack meanings in such a

way that some of the content becomes more implicit, such as the fact that the load is reduced by an external agent implicated in the passive voice construction on the left.

Grammatical metaphor is deployed throughout the language system and accounts for the fact that states and events represented in texts can potentially be encoded through different wordings (a clause, a phrase, etc.). The choice for a more or less metaphorical wording in turn bears an impact on the degree of implicit or explicitness of the meanings construed in language.

Research on grammatical metaphor as a phenomenon having an implication in translated text [9, 11] has posited the hypothesis of (de)metaphorization as a process accounting for perceived differences between non-translated and translated text. Three sources have been identified as likely explanations for properties characterizing translated texts as opposed to non-translated text. The *typology* of the source language system may be reflected in some of the properties of the translation; the *registers* of a source text and a target text for a given context may not be the same, thus demanding decisions on the translator's part on how to construe meanings oriented to the contextual variables envisaged for the target text; and finally, *understanding* involves relating given units of text to more explicit and more literal paraphrases and in this sense demands decisions on the translator's part as to how metaphorical certain wordings need to be, can be or will ultimately have to be.

Drawing on observations of texts in comparable (translated and non-translated samples and parallel corpora (originals and their translations), understanding in monolingual and multilingual text production can be modeled based on grammatical metaphor, the translator's performance involving relating meaningful (grammatical) units to their more or less metaphorical variants [9]. When the level of metaphoricity is lower in the translated text than in the original one, explicitation of meanings implicitly encoded in the original text is performed by the translator drawing on co-textual and contextual assumptions. The following example, retrieved from a website (www.linguee.com) offering a search engine of parallel corpora shows an aligned occurrence of a clause containing the verb "to evolve" and a circumstance of location realized by a prepositional phrase. This circumstance is partly demetaphorized in its translation into Portuguese and realized as a verb ("come to be") due to typological differences between the two languages:

Source text
Some day it might **evolve** into a real citizens' initiative, found in the legislation of some Member States.
English: www.europarl.europa.eu/sides/getDoc...;language=EN

Target text
Um dia este instrumento poderá **desenvolver-se e vir a ser** uma verdadeira iniciativa popular, que aliás já consta da legislação de alguns Estados-Membros.
Portuguese: www.europarl.europa.eu/sides/getDoc...;language=PT

Back translation into English
One day this instrument may **develop and come to be** a true popular initiative, which by the way is already part of the legislation of some Member States.

Fig. 2. Example of demetaphorization in translation

The wordings in bold above show that the meaning construed in English by "evolve into" need to be construed in Portuguese by two verb groups "desenvolver-se" (develop) and "vir a ser" (come to be), which illustrates explicitation of part of the meaning packed in a circumstance in English ("into ...").

Steiner's observations from the final output or translation product perspective have also been confirmed in studies of the translation process [12], as data obtained through key logging shows a series of micro units within one or more macro units that encapsulate (de)metaphorization processes. In this sense, the concept of micro and macro-units [13] as retrievable from key-logging data allows for capturing paths of (de)metaphorization movements that may be or not perceivable in the final rendition output.

This can be supported by evidence obtained from eye-tracking data, particularly regarding both progressive and regressive fixations within text and gaze trajectory across areas of interest (source and target texts) during instances of (de)metaphorization identified though macro units in tasks logs.

In order to illustrate the methodological steps implemented for tracking (de)metaphorization movements and the analytical procedures adopted to explicate the shifts in the level of metaphoricity, this paper examines results from an experiment involving eight Brazilian professional translators who translated a popular science text from Portuguese (L1) into English (L2). The rationale for the experiment is briefly described in the following section.

3 Methodological Considerations

Focusing on the notion of grammatical shifts (parts of speech changes), Hansen-Schirra, Neumann & Steiner (2007) proposed a methodology for product-oriented and corpus-based studies of translated texts to enable the annotation and alignment of a parallel corpus of source and target texts [14]. This allows for the identification of alignment units between source and target texts, including unaligned segments in the source and target texts ("empty links") and segments which can be aligned only at a higher rank due to differences in grammatical functions in the source and target renditions ("crossing lines").

Probing translated text production but focusing on the notion of translation unit (as *foci of attention*), Alves & Vale (2009) proposed a methodology for process-oriented and corpus-based studies of translated texts to mark, annotate, extract and classify translation units (TUs) as micro and macro translation units [13]. The authors developed the Internet-based software package Litterae (available at http://letra.letras.ufmg.br/litterae/index.xml), which is able to read XML files generated by Translog 2006© and automatically provide micro units on the basis of a user-provided pause unit (in seconds). These micro units can be grouped into macro units and further analyzed as the user inserts annotation categories, such as the phase of the translation process where each micro unit is found.

As far as Hansen-Schirra, Neumann & Steiner's (2007) proposal is concerned, category (part of speech) change is examined by mapping alignment units (AU) from source texts onto corresponding occurrences in target texts (final output of translation)

[14]. Alves & Vale (2009), on the other hand, try to map translation units from source texts onto sequences of corresponding translation units in the unfolding of target text production [13]. Together, the two proposals can map alignment units in source and target texts onto translation units which can be approached as evidence of *cognitive entities* observable in the process data and, therefore, allocate an entire set of translation process data, consisting of TUs, to appropriate AUs. In this paper we follow a similar approach, based on a methodology put forward in Alves et al (2010) [12], to explore the modeling of language processing in translation on the basis of grammatical metaphor with a focus on instances of effortful text production.

3.1 Sample Experimental Design

Eight Brazilian professional translators (named from BT1 to BT8) participated in an experiment carried out in Belo Horizonte, Brazil, 2010, aimed at investigating, among others, the impact of more or less metaphorical wordings in the source text on the rendition of the target texts. Having access to only one electronic dictionary, they were asked to translate a text from Portuguese into English (inverse translation task, i.e. L1 into L2). Two source texts were used; these were two versions (A and B) of a popular science text, generated through manipulation in order to create analogous instances of more or less metaphorical wordings in each version. Task execution was recorded through key-logging and eye-tracking software. Free and guided recall protocols were carried out upon task completion, both of them being eye-tracked and audio-recorded. Key-logging data was analyzed to identify micro units within macro units. Eye-tracking data was analyzed to investigate subjects' gaze trajectories and fixations. Given time and space constraints and our attempt to illustrate (de)metaphorization movements in detail, the analysis of micro units herein reported is limited to two particular macro units in the translation process of one of the subjects, namely those concerning the first two clause complexes translated by subject BT5.

4 A Case in Point: Analysis and Discussion

As mentioned above, two versions of the source text where used in the experiment, each one having instances of more or less metaphorical wordings when compared against each other.

Figure 3 shows the first two clause complexes in Portuguese of Version B, the input for subject BT5's translation process. Back translations in English are provided, and manipulated wordings are in bold.

Version B
Clause complex 1 **A tarefa de identificar um bom café** é para os degustadores relativamente simples, **mas a atribuição de uma nota exata para cada amostra** é outra história. **Clause complex 2** **O degustador aprecia** com base em habilidades que adquire com a experiência.
Back translation
Clause complex 1 **The task of identifying a good coffee** is for taster relatively simple, but **the assignment of a precise score** to each sample is an entirely different matter. **Clause complex 2** **A taster judges [coffee samples]** on the basis of skills acquired through experience.

Fig. 3. Clause complexes corresponding to the Macro Units under scrutiny

Macro units in BT5's process were mapped on the basis of key-logged data obtained through Translog with 3-second-long pauses and grouped together using the software package Litterae. Each micro unit corresponds to a meaningful text string found in between pauses either in the drafting phase (i.e., from first keystroke until the first draft of the whole source text) or in the revision phase (i.e., any changes implemented after the rendition of the first draft of the whole source text). Figure 4 below illustrates a macro unit which consists of 8 micro units in the drafting phase and 1 micro unit in the revision phase.

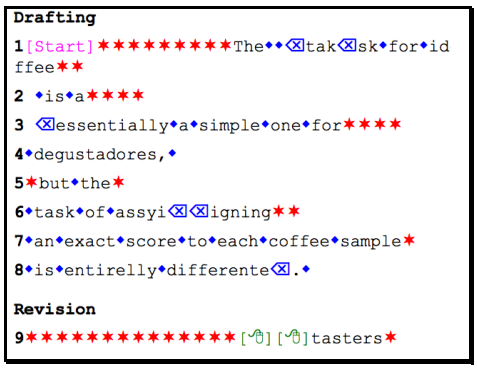


Fig. 4. BT5's micro units for the translation of clause complex 1

BT5 translated the text with little recursiveness (related to deletion of typos), and also kept one word in Portuguese (i.e., "degustadores), which is translated in the last micro unit in the revision phase. The pauses seem to be related to effort: in the first micro unit, one of the longest, as the eye-tracking data show, pauses relate to the reading of the whole clause complex; in micro units 2, 3, 5, 6 and 7, they seem to be related to trying to solve a lexical problem, such as the search for a noun in English for "degustadores" (tasters); and in the ninth micro unit, a substantially long pause as

well, eye tracking data shows that this is related to look ups within the dictionary provided.

The second clause complex can be mapped onto the following micro-units below:

Fig 5. BT5's micro units for the translation of clause complex 2

Figure 5 illustrates a macro unit which consists of 11 micro units with little recursiveness (related to deletion of typos), those units being related to choices at the word rank. Although this seems to be a very short macro unit, lasting one minute and eight seconds in the drafting phase and 1 minute and 46 seconds in the revision phase, BT5 makes considerable changes in the revision phase, as evidence by the occurrence of 5 micro units (45 % of the macro unit). These changes will be explained below.

The key-logged data reported as micro units in Figures 5 and 6 were mapped onto eye-tracking data in order to verify if gaze trajectory and eye fixations revealed effortful attempts on BT5's part to translate the second clause complex.

In Figure 6 each frame corresponds to 15-second-long gazes. Such eye-tracking data shows considerable effort in the rendition of the macro unit corresponding to the clause complex 2 in the drafting phase. Lines linking fixations from the source text through the target text area of interest (and vice-versa) show the subject's recurrent need to process small portions of the source text in order to produce the target text (see the short distance between lines in frames 2 and 3). Fixation also shows recursiveness in the reading of this macro unit, there being almost one fixation per word.

In Figure 7, each frame corresponds to a 20-second-long gaze. In the revision phase, BT5's gaze does not show recurrent movements from source to source text (and vice-versa). As expected for this phase in the process [15], most fixations are found in the target text area of interest.

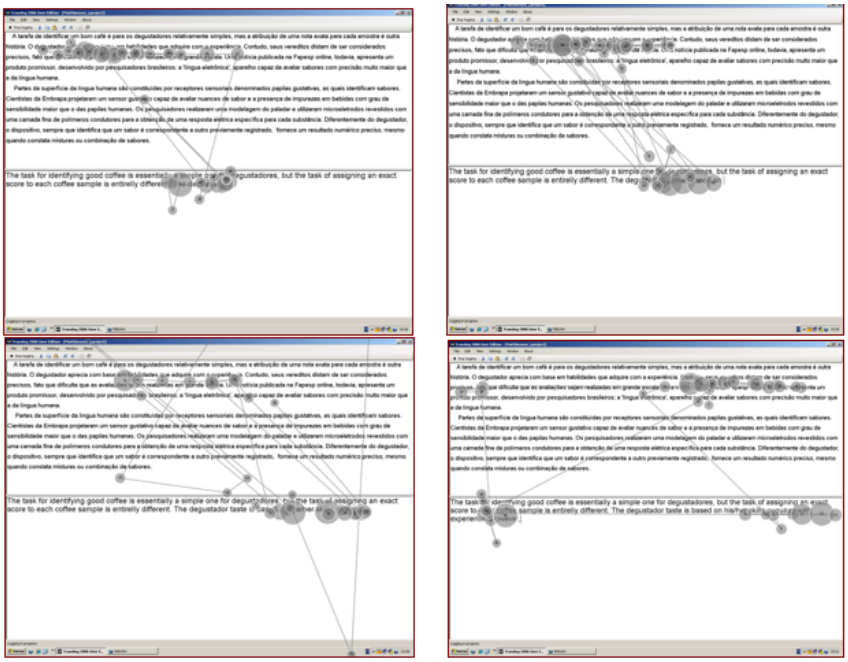


Fig. 6. Sequence of gaze plots generated by eyetracking for the performance of BT5

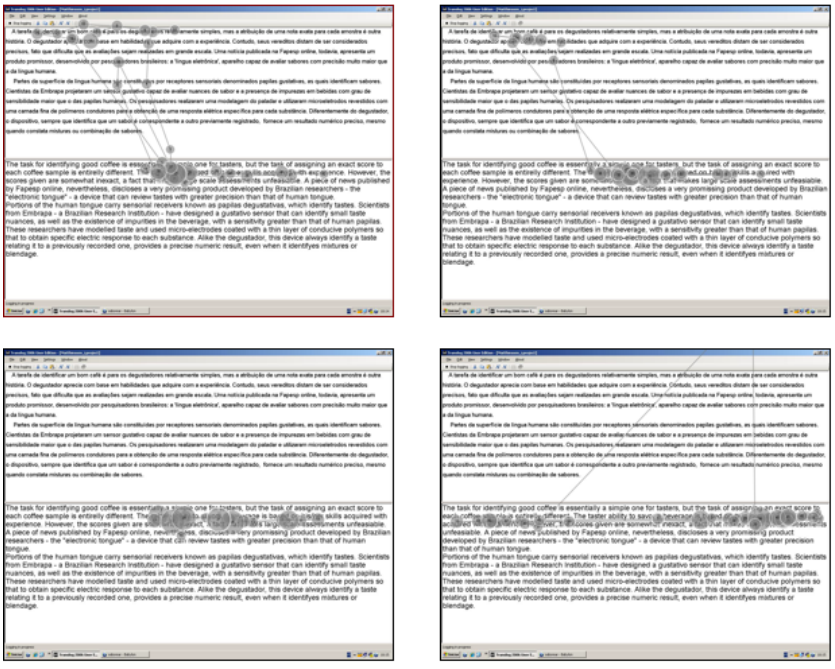


Fig. 7. Sequence of gaze plots generated by eye tracking for the performance of BT5

As a joint analysis of Figures 4-7 shows, particular patterns of pauses, recursiveness and eye gazing and fixations can be cross-analyzed to locate instances in the source and target texts where effort is stronger. Let us now turn to an analysis of motivations for the effort invested by the subject in terms of the constraints in the two language systems to which the translation problem can be ascribed.

If we have a look at the source text, there is a typological problem that exerts pressure on the translator's behavior. Where the source text reads

> O degustador aprecia com base em habilidades que adquire com a experiência.

the intransitive use of the verb "aprecia" in Portuguese, back-translatable into English as ***judges***, *savors* or ***tastes***, demands that the translator overcome two potential problems: one is the need to turn this intransitive verb in Portuguese into a transitive verb in English, due to typological constraints in the latter; and the other is to seek to avoid the repetition of **taste**, if the choice is for the verb **to taste**.

The final output in BT5's text is

> The taster ability to savor a beverage is based on his/her previous skills acquired with years of experience.

where we can see concurrent metaphorization and demetaphorization when compared to the source text meanings. "O degustador aprecia" (the taster savors) is realized as "the taster ability to savor", with a metaphorization added through the noun "ability", which offers a solution to the problem posed by the meaning construed by the intransitivity of the verb in Portuguese. "Aprecia" (savors, tastes) is realized as "to savor a beverage" with a consequent demetaphorization through explicitation of an object selected in order to use a transitive construction in English.

However, and this is where process data comes in to play a fundamental role in our analysis, metaphorical shifts in the translation product may entail further metaphoricity shifts in the process. This can clearly be seen in the micro units identified for the macro units under scrutiny.

During the drafting phase, the subject's log shows the following interim rendition:

> The degustador taste is based on his/her skills acquired with experience

where the problem of the intransitive form in Portuguese finds an interim solution in English through the selection of a copula or relational process "is based on". This is taken up in the end-revision phase, where three attempts are made with various degrees of metaphoricity:

1. The taster taste is based	+ metaphorical
2. The taster savors a beverage is based	– metaphorical
3. The taster ability to savor a beverage	+ metaphorical

Fig. 8. Target text renditions during end revision phase

Shifts in levels of metaphoricity such as the ones observed in the translation log above seem to point to a strategy implemented by the subject to deal with typological differences between the two language systems. Interestingly enough, the subject may or may not be aware of this strategic path taken. In the case of BT5, data from the verbal recall recorded upon task completion shows no evidence of awareness on the subject's part, as all he says in his protocol regarding his task is:

> "I did a dirty translation first, using, introducing some words in Portuguese ... that I was not sure that I could use ... *taster* ... and then I used *degustador* in Portuguese all the same ... and ... in order to later on in the revision ... to go back to doubts and improve the text."

Most significantly for the purposes of the present discussion, shifts in levels of metaphoricity such as the ones observed in the BT5's translation log and eye-tracking data seem to provide empirical evidence of meaning making processes at stake in translation of the kind that can be mapped by further research with potential implications for modeling human translation processes.

5 Concluding Remarks

In this paper, we have attempted to provide a brief illustration of a methodology and analytical procedures that can be adopted in order to explore a particular phenomenon in meaning production, namely grammatical metaphor. Its identification in the course of a task execution was clearly made through pauses, eye fixation and gaze plots, indicators of effortful text production. The foci of attention, mapped on time and resources invested by the translator to deal with a translation problem ascribed to such instances of effortful production, need not find a counterpart in recall protocol data, even though a discussion of this kind of data from an expert performance perspective would certainly point to more expert like behavior if evidence of meta-reflection and metalanguage can be found in the protocols.

On the whole, our approach shows the potentiality for exploring eye-tracking data to account for higher-level cognitive processes in translation, along lines somewhat different from those in standard psycholinguistic research which tend to focus on automatic aspects of language processing. The methodology also has implications for translation modeling through shifts in metaphorical wording. Finally, grammatical metaphor seems to offer a productive approach to show instances of effortful language processing in translation by mapping alignment units onto translation units. It also highlights the need for a comprehensive theory of language for translators to develop awareness and metalanguage to account for their choices. As sketched herein, the proposed methodology promises to open up a new avenue for the investigation of meaning construction in translation and should now be tested in larger samples of translation process data to be further developed.

References

1. Jakobsen, A.L.: Translation Drafting by Professional Translators and by Translation Students. In: Hansen, G. (Ed.) Empirical Translation Studies: Process and Product, p. 191-204. Samfundslitteratur, Copenhagen (2002)
2. Jakobsen, A.L.: Investigating Expert Translators' Processing Knowledge. In: Helle, V. et al. (Eds.) Knowledge Systems and Translations, p. 173-189. Mounton de Gruyter, The Hague (2005)
3. Jakobsen, A.L., Göpferich, S., Mees, I. (Eds.): Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing. Copenhagen Studies in Language 36. Samfundslitteratur, Copenhagen (2008)
4. Mees, I., Alves, F., Göpferich, S. (eds.): Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen. Copenhagen Studies in Language 39. Samfundslitteratur, Copenhagen (2009)
5. Silva, I.: Conhecimento Experto em Tradução: Aferição da Durabilidade de Tarefas Tradutórias Realizadas por Sujeitos Não Tradutores em Condições Empírico Experimentais. Unpublished Thesis, Faculdade de Letras, UFMG, Belo Horizonte (2007)
6. Pagano, A., Silva, I. Domain Knowledge in Translation Task Execution: Insights from Academic Researchers Performing as Translators. In: XVIII Fit World Congress, CD-ROM. Foreign Language Press, Shangai (2008)
7. Halliday, M.A.K., Matthiessen, C.M.I.M.: Construing Experience through Meaning: a Language-Based Approach to Cognition. Continuum, London (1999)
8. Halliday, M.A.K., Matthiessen, C.M.I.M.: Introduction to Functional Grammar. 3rd ed. Edward Arnold, London (2004)
9. Steiner, E.: Intralingual and Interlingual Versions of a Text – How Specific Is the Notion of Translation. In: Steiner, E.; Yallop, C. (Ed.) Exploring Translation and Multilingual Text Production: Beyond Context, p. 161-190. Mouton de Gruyter, Berlin/New York (2001)
10. Matthiessen, C.M.I.M.: Theme as an Enabling Resource in Ideational 'Knowledge' Construction. In: Ghadessy, M. (Ed.) Thematic Development in English Texts, p. 20-84. Pinter, London (1995)
11. Teich, E.: Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts. Mouton de Gruyter, Berlin/New York (2003)
12. Alves, F.; Pagano, A., Neumann, S., Steiner, E., Hansen-Schirra, S.: Translation Units and Grammatical Shifts: Towards an Integration of Product- and Process-Based Translation Research. In: Shreve, G.; Angelone, E. (Eds). Translation and Cognition, p. 109-142. John Benjamins, Amsterdam (2010)
13. Alves, F.; Vale, D.: Probing the Unit of Translation in Time: Aspects of The Design and Development of a Web Application for Storing, Annotating, and Querying Translation Process Data. Across Languages and Cultures 10 (2), 251-273 (2009)
14. Hansen-Schirra, S., Neumann, S., Steiner, E.: Cohesive Explicitness and Explicitation in an English-German Translation Corpus. Languages in Contrast 7(2), 241-265 (2007)
15. Alves, F., Pagano, A., Silva, I. A.: New Window on Translators' Cognitive Activity: Methodological Issues in the Combind Use of Eye Tracking, Key Logging and Retrospective Protocols. In: Mees, I., Alves, F., Göpferich, S. (eds.) Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen. Copenhagen Studies in Language 39. Samfundslitteratur, Copenhagen (2009)

NLP and cognitive modelling

Modeling Cognitive Frames for Situations with Markov Logic Networks

William R. Murray[1] and Dominik Jain[2]

[1]The Boeing Company
P.O. Box 3707 MC7L-66
Seattle, Washington 98124-2207
william.r.murray@boeing.com

[2]Intelligent Autonomous Systems Group
Technische Universität München
jain@cs.tum.edu

Abstract. Machine reasoning and human reasoning differ in the representation of concepts and means of reasoning. Machine reasoning emphasizes formal logic or probabilistic reasoning to represent concepts and to reason about situations. In contrast, human reasoning emphasizes embodied concepts based on prototypes, family resemblances, and metaphorical reasoning.

We investigate the application of Markov Logic Networks (MLNs), a kind of statistical relational model, as an improved means of modeling human concepts and inference, compared to First-Order Logic or standard probabilistic approaches such as Bayesian Networks. We apply MLNs to the task of inferring situational frames that describe short narratives. Finally, we compare our approach to abductive approaches, comparing how well each approach infers situational frames and models human cognitive frames.

Introduction

Statistical Relational Learning (SRL) is a branch of machine learning that is concerned with learning and inference over multiple kinds of objects and multiple kinds of relationships expressed, for example, using first-order logic (FOL) formulas. Traditional classifiers handle only a single kind of concept and do not reason about links between concepts. One SRL approach, Markov Logic Networks (MLNs), represents joint probability distributions with weighted FOL formulas. Each weighted formula represents a constraint, the weight indicating the hardness of the constraint. Markov Logic Networks subsume HMMs, Bayesian Networks (BNs), Dynamic Bayesian Networks, and (in the limit of infinite weights) FOL. Markov Logic Networks can be viewed as a new kind of common interface language for artificial intelligence [1].

Our hypothesis is that MLNs better model human cognitive frames than either logic or probabilistic graphical models, for they combine the complementary

advantages of both approaches. Logic allows generalities to be expressed, e.g., mammals have fur, have live birth, and do not fly. Adding probability decreases the brittleness of these generalizations, allowing different likelihoods of being true. Essentially, MLNs allow us to compactly represent generalizations while still allowing for exceptions. The exceptions are just less probable. Probabilistic graphical models such as Bayesian Networks can also express uncertainty, but are restricted to to a fixed set of propositional relationships. Newer SRL approaches designed to handle Horn-clause or other generalizations are described in [2]; these are similarly experimental and at this point the MLN technology seems to have gained a wider following and increasing scope of application [3].

Another reason MLNs hold promise for modeling human beliefs is their ability to model knowledge bases that are contradictory. With contradictions, no interpretation satisfies all the rules, but MLNs allow those interpretations that best fit the rules to be the most likely. Humans may similarly hold contradictory beliefs, yet manage to reason and act under uncertainty. In contrast, first-order logic does not allow *any* inconsistencies, since then any formula can be proven.

In the research described, we model simple narrative situations described as stories. MLNs are used as a computational mechanism to formally represent human concepts mentally represented by humans as cognitive frames, with prototypes [4]. We are working towards a more accurate model of human concepts, but still one amenable to formal representation and reasoning.

Representing Cognitive Frames in Markov Logic Networks

Humans represent concepts such as situations with cognitive frames [4], a prototype-based psychological model of the world. Our motivation for improving how computers formally model cognitive frames is to facilitate collaborative human-computer applications and improved human-computer interaction (HCI). Adaptive systems, such as intelligent tutoring systems, and natural language systems, such as discourse management systems, are examples of the application areas that would benefit. More opaque formal methods (e.g., neural networks or SAT planning techniques) may be sufficient for AI systems that have no need of human interaction. We expect that collaborative systems requiring human knowledge acquisition, and human-in-the-loop interaction will be required in the majority of AI applications: First, human intelligence and knowledge is still required for common-sense reasoning, and second, fully autonomous systems are not yet trusted in critical applications.

Our focus on this paper is representing cognitive frames as human concept representations formally in computers. The reasoning modeled is that of pattern recognition, where missing details are filled in by default. We do not address metaphoric reasoning at this point, but see [10] for an initial investigation using formal theorem proving.

If we hear that we are having dinner at a restaurant we will expect to receive a menu, have to order, have our food served by the waiter, and then have to pay the waiter or cashier at the end. The actions, physical setting, roles, and props are stereotypical. Frames vary across cultures and an expected aspect of cultural membership is knowledge of its conventions [11]. Examples of varying cultural

expectations include tipping (who gets how much), conversational conventions (greetings, turn-taking, and how loud conversation may be), and restaurant pricing (e.g., take-away food versus food eaten in the restaurant).

Formal (computer) representations of human concepts are models of models of the world, as shown in Figure 1. The computer's formal model represents a human cognitive frame. In turn, a human cognitive frame (idealized cognitive model [4] or schema) is a prototypical representation of their experience, which crucially depends on interacting and experiencing the world in a human body. An ability to formally represent and reason about human concepts, and what they depict in the world, is required for improved human-computer interaction. Without this ability, communication is restricted in a way that is similar to the cultural misunderstandings between people, only more seriously aggravated by the lack of shared human experience.

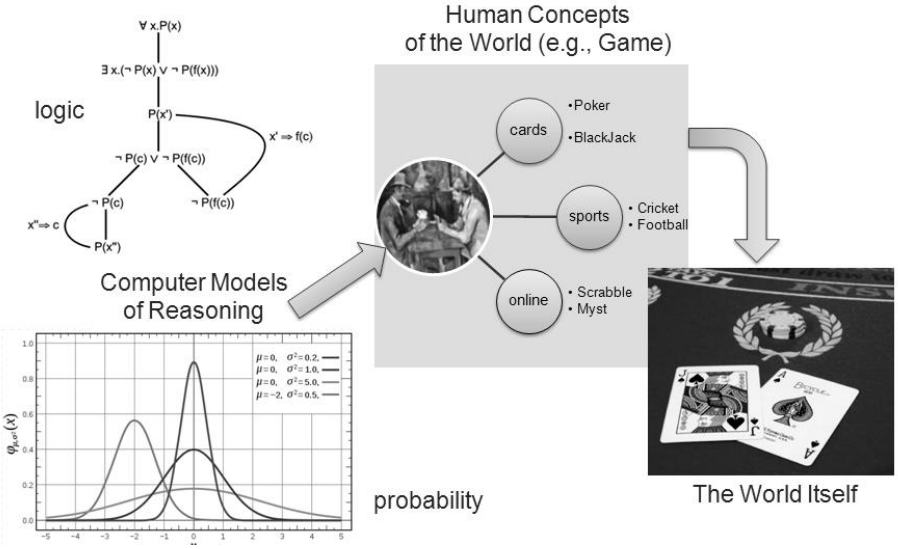


Figure 1. HCI applications need to represent human cognitive frames, and reasoning. Formal representations of world entities and relationships are not sufficient.

Modeling Cognitive Frames that Represent Situations with MLNs

In this first step to investigate the ability of MLNs to model situational frames, we leverage existing training and test data from [5]. A set of 25 training sentences and 25 test sentences were interpreted with Markov Logic rules. The task was to learn frame classifications appropriate to the sentences based on the physical setting of the actions, the kinds of actions that happened, the props (objects) used in the actions, and the roles played by humans in the situations. First, we explain our motivation for using Markov Logic Networks. Second, we provide a brief formal introduction to MLNS. Finally, we describe our problem domain, problem representation, and results obtained.

How Markov Logic Networks can Improve the Modeling of Cognitive Frames

Informally, statistical relational learning (SRL) and statistical relational reasoning can be viewed in two ways:

1. SRL is an extension of machine learning that:
    a. Handles multiple types of concepts and multiple types of relationships (links) between them. Traditional concept learning handles only one type of concept at a time, and has very limited handling of links between concepts.
    b. Focuses on representation and reasoning over links, as well as concepts.
2. Reasoning in SRL combines probability and first-order logic. It lifts Bayesian Networks and other propositional approaches to first-order logic, essentially allowing probabilistic statements to be expressed over broad classes of objects simultaneously. The reasoning can be viewed simply as "softening" first-order logic, allowing both hard and soft constraints to be expressed in first-order logic. However, once detailed knowledge engineering begins, a deeper understanding is required to correctly model domains.

So why are Markov Logic Networks a step towards a better model for human cognition? Logic alone has these limitations:

1. It assumes necessary and sufficient conditions, which is a poor model for many natural kinds [4] and basic categories. It is a good model for formal domains such as mathematics and works well for engineering artifacts and exactly prescribed procedures, until real-world concerns begin to interfere with the abstract models.
2. Logic assumes entities and predicates with immutable identities. In the real world, objects, people, definitions, and frames are continually changing. For example, computer and mobile technology is continually improving and evolving. People age and acquire new beliefs. Word meanings, such as what it means to be a game, change with new technology. Even technological terms (e.g., planet) change, and concepts such as 'art' continue to be culturally contested [11]. Finally, the means of communication and its socially accepted protocols continue to evolve rapidly in the online world.
3. Logic assumes that predicates are either true or false and that no entity is a better example of the predicate than any other. In other words, there is no notion of a prototype or ideal exemplars. In contrast, we know that robins are better prototypes for birds and Michelangelo's works are better exemplars of art than surrealism or an art student's still life.
4. It assumes a sentence semantics and ideal of truth defined by a Kripke model, that is an interpretation of constants (e.g., predicates, entities, and functions) that satisfy a set of axioms. Instead, for humans "truth" itself is a concept,and a statement may be only considered "technically true", if the situation is quite different from what that would imply [4].
5. It assumes a single context in which sentences are either true or false. Instead, humans 'frame' facts and sentences in different ways, according to their

backgrounds, and these frames alter their interpretation of sentences and whether they regard them true or not (e.g., 'Universal healthcare is socialism,' or 'Wealthy people do not pay their fair share,' according to one's politics in the US).

6. Without extensions, logic does not handle defaults, uncertainty, time, change, necessity, beliefs, or other modal assertions about propositions. Of course, we can add extensions for each of these individually, but combining these becomes more and more unwieldy.

Instead, to retain a principled (mathematically well-founded and computable) approach we might turn to probability. This approach would have the advantages of:

1. Better supporting schema recognition as a kind of pattern recognition. By formally modeling prototypical concepts and radial concepts we can provide a graded categorization of instances that more accurately models the way humans think.
2. Better expressing uncertainty: We can express likelihoods ('It will rain with a 40% chance tomorrow') or degrees of belief ('John believes that his proposal will be accepted with 50% certainty.').
3. Remaining amenable to automated reasoning through the use of algorithms such as belief propagation for Bayesian reasoning or Markov-Chain Monte Carlo sampling for Markov Random Fields.

Before the advent of SRL approaches, most tractable approaches to handling probability were restricted to graph-based models handcrafted for one set of individual objects. Directed-graphs gave rise to Bayesian models. Undirected models gave rise to Markov Random Fields. The newer SRL approaches generalize these approaches by specifying templates that apply to sets of objects of defined types.

Statistical relational reasoning approaches, such as Markov Logic Networks, combine logic and probability. The probabilistic aspects ameliorate the brittleness of purely probabilistic approaches. First-order logic allows a more compact expression of rules. The rules are no longer hard and fast as now soft constraints and even contradictory beliefs can be expressed. Different sets of beliefs are conceptually represented as alternate *possible worlds*. Weights on first-order logic formulas express how likely each of these possible worlds is. Practically, not all of the possible worlds are generated to avoid combinatorial explosions.

The sentences appear deceptively simple, e.g., "Bill took a bus to a restaurant. He drank a milkshake. He pointed a gun at the owner. He got some money from him." but have been used to test inductive logic programs (ACCEL) and abduction in Markov Logic networks [8]. Only a small set of frames are used (robbery, shopping, dining, bus travel, taxi travel, and air travel) but more than one frame can apply in a situation.

The sentences were automatically translated into FOL with WordNet concepts as ontology terms using the controlled natural language CPL [7]. A controlled natural language is a subset of English that can be translated into logic. The narrative sentences are directly readable into CPL with only minor changes from the original. Examples of these changes include adding "Then" or "Finally" to the beginning of a sentence to indicate event sequencing, and correcting word spellings, e.g., changing "busdriver" to "bus driver", so that CPL can identify the correct WordNet terms.

Markov Logic Networks: A Brief Formal Introduction

More formally, Jain, et al., [9], provide a concise overview of Markov Logic Networks:

Markov logic networks (MLNs) are probabilistic logical models that combine the semantics of probabilistic graphical models (namely Markov random fields, a kind of undirected graph) with the full power of first-order logic. An MLN can be seen as a set of constraints on the set of possible worlds that is implicitly defined by a set of logical predicates and a set of constants, as each logical atom that can be constructed using these domain elements is viewed as a boolean variable. Specifically, the constraints are formulas in first-order logic with attached numeric weights that quantify their hardness.

Formally, a Markov logic network L is a set of pairs (F_i,w_i), where F_i is a formula in first-order logic and w_i is a real number, the weight of formula F_i. Together with a finite set of constants C, an MLN defines a Markov network $M_{L,C,}$ the ground Markov network, as follows:

1. $M_{L,C}$ contains one binary node for each possible grounding of each predicate appearing in the formulas of the Markov logic network L.

2. $M_{L,C}$ contains one feature (a real-valued function) for each possible grounding of each formula F_i in L. The value of this feature is 1 if the ground formula is true, and 0 otherwise. The weight of the feature is w_i.

The ground Markov network's set of variables **X** is the set of ground atoms that is implicitly defined by the predicates in the MLN and the set of constants C. The Markov logic network specifies a probability distribution over the set of possible worlds $\mathcal{X}$, i.e. the set of possible assignments of truth values to each of the ground atoms in **X,** as shown in Equation 1 below, where the inner sums are over indices of MLN formulas and $n_i(x)$ is the number of true groundings of the *i*-th formula in a possible world of $\mathcal{X}$.

$$P(X = x) = \frac{1}{Z} \cdot \exp\left(\sum_i w_i \cdot n_i(x)\right) = \frac{\exp\left(\sum_i w_i \cdot n_i(x)\right)}{\sum_{x' \in \mathcal{X}} \exp\left(\sum_i w_i \cdot n_i(x')\right)} \quad (1)$$

Inference

Since (1) provides the full-joint distribution over the variables in $\mathcal{X}$, it can be used to compute arbitrary conditional probabilities: The probability that a formula F_1 holds given that formula F_2 does can be computed as

$$P(F_1 \mid F_2, L, C) = P(F_1 \mid F_2, M_{L,C}) = \frac{P(F_1 \wedge F_2 \mid M_{L,C})}{P(F_2 \mid M_{L,C})}$$
$$= \frac{\sum_{x \in \mathcal{X}_{F_1} \cap \mathcal{X}_{F_2}} P(X = x)}{\sum_{x \in \mathcal{X}_{F_2}} P(X = x)}$$
$$= \frac{\sum_{x \in \mathcal{X}_{F_1} \cap \mathcal{X}_{F_2}} \exp\left(\sum_i w_i \cdot n_i(x)\right)}{\sum_{x \in \mathcal{X}_{F_2}} \exp\left(\sum_i w_i \cdot n_i(x)\right)} =: \frac{W_{F_1 \wedge F_2}}{W_{F_2}} \qquad (2)$$

where $\mathcal{X}_{Fi}$ is the set of possible worlds in which F_i holds, and W_{F1^F2} and W_{F2} are the sums of exponentiated sums of weights for possible worlds where F1 ^F2 holds and where F_2 holds respectively.

The Problem Domain: Describing Simple Narratives with Situational Frames

In this section we illustrate the narratives and the situational frames used to describe them. Most narratives are quite short such as: "Jack went to the liquor store. Then he found some bourbon on the shelf." while some are longer, e.g., "John bought a ticket. Then he went to the restaurant. Then he ordered a milkshake. Finally he boarded the plane." The first story is described by the dining frame alone while the second is described by both the dining frame and the air travel frame together.

The Problem Representation: From Controlled English to MLN Input Data

Our approach is to first take each story and translate it from controlled natural language to a formal logic translation. We show the representation with a Prolog notation, for a short story ('Bill went to the liquor store. Then he pointed a gun at the owner.') taken from the training data. First, the CPL translation to logic:

```
(1) "Bill went to the liquor store."
   isa(liquor-store01,off-licence_n1),
   isa(go01,go_v1),
   named(Bill#n,["Bill"]),
   tense-is(sentence,past),
   agent(go01,Bill#n),
   "to"(go01,liquor-store01).
(2) "Then he pointed a gun at the owner."
[ Resolved "he" -> Bill#n ]
   isa(Bill#n,Bill#n),
   isa(gun01,gun_n1),
   isa(owner01,owner_n1),
   isa(point01,indicate_v2),
   tense-is(sentence,past),
   agent(point01,Bill#n),
   object(point01,gun01),
   is-at(point01,owner01),
   next-event(go01,point01).
```

Next, from this representation we extract key features to provide as *evidence* (input) to the Markov Logic Network learning algorithms. A LISP program performs this extraction, e.g., looking for WordNet instances of places for settings and other objects as props. Actions are WordNet verbs. The results are shown below:

```
// Frame classification
Describes(Travel,Story_6)
Describes(Robbery,Story_6)

//Roles...
HasRole(Story_6,Owner_N1)

//Setting
HasSetting(Story_6,Liquor_Store_N1)

//Props...
HasProp(Story_6,Gun_N1)

//Players...
HasPlayer(Story_6,Bill)

//Events...
HasEvent(Story_6,Go_V1)
HasEvent(Story_6,Indicate_V2)

//Players causing events...
HasAgent(Story_6,Bill,Go_V1)
HasAgent(Story_6,Bill,Indicate_V2)

//Players undergoing events...

HasObject(Story_6,Indicate_V2,Gun_N1)
```

In the training examples, such as the example above, we specify the situational frames that apply. In the test examples, they are not specified but inferred from the Markov Logic Network learned from the training examples.

MLN learning algorithms from the Alchemy software package [3] learn weight parameters from the training data. The MLN with learned weights can then identify the best frames for new situations given *partial* information about events, props, roles, setting, and event sequencing. Multiple frames can match a short narrative.

We infer situations from their various attributes and learn the importance of each attribute from examples. The intention is to emulate the induction people perform over many experiences of the same kind of situation. The MLN

```
HasEvent(sit,+event) => Describes(+fr,sit)
HasProp(sit,+prop) => Describes(+fr,sit)
HasSetting(sit,+place) => Describes(+fr,sit)
HasRole(sit,+r) => Describes(+fr,sit)
```

expresses logical rules that associate various aspects of situations with frames that may describe them. The first rule asserts that an event implies the frame that interprets the situation. The second rule asserts that a prop (object) in a situation implies the frame that interprets it. The last rule asserts that the setting (physical place) implies the frame. All of these rules are only sometimes true and to different degrees. SRL

software such as Alchemy [3] learns appropriate weights measuring how strongly each rule is true from user data. The weights vary for each kind of event, prop, or setting. We describe our results below.

Results

We trained both generative and discriminative models using the methods available in the Alchemy system [3]. Generative training can be used to obtain models that can be applied in various ways, for the representation of a full-joint distribution allows arbitrary conditional distributions to be inferred (as given in Equation 2). By contrast, discriminative learning specifically optimizes conditional distributions that can be used to discriminate between classes such as the kinds of frames we seek to identify in story understanding. Models trained in this way will typically perform well on the single task they are intended to perform, in this case frame identification. However, they cannot reasonably be used for other applications, such as asking how likely an action is for a particular frame, or what props are most likely to occur in a particular setting. For these other kinds of queries a generative model is needed.

We achieved very good results with a discriminative model and the MLN above. The four formula templates in the MLN evaluate how well individual events, props, settings, and roles predict individual frames. The first template (with the predicate HasEvent) expands to a separate formula for every type of event and every type of frame (as indicated by the + operators). The other three templates with predicates HasProp, HasSetting, and HasRole act similarly for props, settings, and roles. Together, the four templates are sufficient to obtain convincing results. The dataset comprising 25 stories used a reduced set of frames that did not differentiate between different ways of travelling for this model. When applied to our test set comprising another 25 stories, the learned model achieved a precision of 91.18%, recall of 77.50%, and F1 score of 83.78%, higher than the highest F1 score (79.56%) realized for the MLN-based abductive approaches to story classification described in [8] and [14]. The ACCEL logic-based abduction system [5], with a coherence metric targeted specifically to story understanding, performed however, with an F1 score of 89.39% [8]. Our results are comparable, but do not take advantage of ACCEL's narrative coherence heuristic, which is tailored to judge how well an explanation satisfies all story sentences. Instead, we used MaxWalkSAT to compute the most likely explanation for the evidence (events, props, settings, and roles) provided in each of the stories. The recognized frames are viewed as retrieved entities. *Precision* is defined as the fraction of frames inferred that are correctly recognized and *recall* is defined as the the fraction of correct frames that are inferred.

The MLN for the generative model used predicates similar to those used before, and the following formulas:

```
// soft constraints
 FramedBy(sit, +f)
 FramedBy(sit, +f) ^ FramedBy(sit, +f2)
*FramedBy(sit, +f) ^ *Prop(sit, +p)
```

```
*FramedBy(sit, +f) ^ *OccursAt(sit, +pl)
*FramedBy(sit, +f) ^ *AgentAction(sit, +ac)
*FramedBy(sit, +f) ^ *HasRole(sit, +r)
```

We also experimented with adding new predicates, e.g., predicates for interactions between one character and another, or between a character and situational props. AgentAction is one of these new predicates and is used to indicate the actions taken by the main character in a situation. We found that some of the new predicates were, given the sparsity of the data, not helpful with respect to learning the key regularities in this domain. The structure of the model is an extended Bayesian network-like model, which would be capable of representing, for example, conditional distributions of FramedBy, Prop, OccursAt and AgentAction given FramedBy, as well as the marginal distribution of FramedBy. With the second formula, we included features that capture the co-occurence of frames, which, however, goes beyond the expressiveness of Bayesian networks.

Our models are created without the manual fine-tuning of weights used in [8] and are structurally engineered in a more straightforward manner. As mentioned previously, the generative model can handle queries other than frame identifications, but at the expense of lower performance on the frame identification task. An example query is shown below:

If we are traveling by air, how likely is it that a suitcase occurs as a prop?

We need to specify that only one frame applies, otherwise the results are conflated with props that occur in other frames that could also otherwise apply:

```
P(Prop(Sit,x) | FramedBy(Sit,Air_Travel_Frame) ^
!FramedBy(Sit,Dining_Frame) ^ ...)
      =
Prop(Sit,Suitcase_N1) 0.987174
Prop(Sit,Bag_N1) 4.9995e-05
Prop(Sit,Bourbon_N2) 4.9995e-05
Prop(Sit,Bread_N1) 4.9995e-05
...
Prop(Sit,Straw_N1) 4.9995e-05
Prop(Sit,Ticket_N1) 8.70283e-05
Prop(Sit,Token_N1) 0.0012721
Prop(Sit,Station_N1) 0.000457362
```

In contrast to the suitcase, the odds of all the other possible props: bourbon, bread, a gun, milk, money, a straw, milkshake, a ticket, or a token are all far smaller.

The results indicate that we can extract some predictions that we would intuitively expect, similar to the prototypes that humans would construct: suitcases are associated with air travel, guns with robbery, and so on. The reason that the results are not more certain is that the amount of data is very sparse: Typically machine learning assumes hundreds to thousands of instances; our data allowed only 25 training and 25 test cases. Furthermore, the extraction of the true principles from the data is hindered by the use of coarse approximations (in our case, the optimization of the pseudo-likelihood [12]). Additionally, the inference tasks can be quite challenging even for state-of-the-art techniques such as MC-SAT [13], a Markov chain Monte Carlo method owing to the occurrence of individual atoms in a very large number of formulas, which can cause the mixing time of Markov chains to be very high.

Related Work

FrameNet [6] is a formal representation of common human situations. It provides a frame hierarchy and frame slots (semantic roles) for a very large number of situations. The frames are much more elaborate than those represented here, with precise descriptions of situational characteristics with frame slots. However, no axioms describe the frames or slots, and a computational means for recognizing and using the frames is not provided.

The University of Texas machine learning group has done work in abduction for the situational data we use in our metrics [8]. Their approach has been to represent each situation with hand-crafted deductive rules and then provide a more general approach that can algorithmically augment these rules for inference in the opposite direction (abduction). The approach also relies on Markov Logic Networks, but their intent is to explore abduction within MLNs, rather than to explore improved modeling of human cognitive frames.

Our approach differs from FrameNet in addressing the computational mechanisms to be applied for recognition of situations and predictions from knowing frames that apply to situations, i.e., roles and events. Our approach differs from that of [8] in not providing handcrafted rules for each situation but instead inferring characteristic events and rules for each situation to model the situation prototypes. Then once prototypical situations have been recognized (more than one may apply at a time) these situations are used for predictive purposes. An explicit approach to abduction or a need to add abductive rules is not part of the process.

Summary

We have begun to explore the promise that Markov Logic Networks have as an approach to model cognitive frames. In our research, situational frames are inferred from the props, actions, roles, and the settings of situations. We used a pre-existing machine learning database of narrative situations from the University of Texas at Austin for testing and development.

Markov Logic Networks combine logical expressivity with probabilistic reasoning. Standard approaches to knowledge representation have focused primarily on modeling human concepts with necessary and sufficient conditions, or by using extensions to formal logic such as default reasoning, while ignoring work on prototypes, exemplars, and metaphorical reasoning. Probabilistic alternatives alone do not capture regularities useful for expressing rules that are typically true, such as the heuristics and cognitive stereotypes that humans acquire from experience.

The SRL approach of MLNs allows reasoning about both concept classifications and relationships between concepts. It further allows representing generalities (through logic) while allowing exceptions (through probability). We have applied MLNs for modeling cognitive frames, similar to the frames represented textually in FrameNet. We have not yet explored deeper modeling issues, such as representing radial concepts [4], contradictory beliefs, or metaphorical reasoning [11]. Clearly much more work needs to be done investigating MLNs, both in these areas and to tie

MLN formalizations of cognitive models into natural language processing or into adaptive systems where computational models of human conceptual reasoning are critical to success.

References

1. Domingos, P., Lowd, D.: Markov Logic: An Interface Layer for Artificial Intelligence. Morgan & Claypool, San Rafael, CA (2009)
2. Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning. The MIT Press, Cambridge, MA (2007)
3. Alchemy - Open Source AI, http://alchemy.cs.washington.edu/ See the publications list for the scope of application.
4. Lakoff, G. Women, Fire, and Dangerous Things. University Of Chicago Press; 1997 edition.
5. Ng, H. T., and Mooney, R. J. (1992). Abductive plan recognition and diagnosis: A comprehensive empirical evaluation. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, 499–508.
6. Fillmore, Charles J. & Baker, Collin F. 2010. A Frame Approach to Semantic Analysis, in Heine, B. & Narrog, H. (eds.) *Oxford Handbook of Linguistic Analysis*, from OUP.
7. Schwitter, R. Controlled Natural Languages for Knowledge Representation. COLING 2010: pp. 1113-1121 (2010)
8. Kate, R. J., and Mooney, R. J. 2009. Probabilistic abduction using Markov logic networks. In *Proceedings of the IJCAI-09 Workshop on Plan, Activity, and Intent Recognition (PAIR-09)*.
9. Jain, D., Kirchlechner, B. and Beetz, M. Extending Markov Logic to Model Probability Distributions in Relational Domains. In *Proceedings of the 30th German Conference on Artificial Intelligence (KI-2007)*, pp. 129–143, 2007.
10. Murray, W. R. Conceptual Metaphor and Scripts in Recognizing Textual Entailment. The 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008).
11. Kövecses, Z. Metaphor in Culture: Universality and Variation. Cambridge University Press. (2006).
12. Richardson, Matt and Domingos, Pedro (2006). Markov Logic Networks. Machine Learning, 62, 107-136, 2006.
13. Poon, Hoifung and Domingos, Pedro (2006). Sound and Efficient Inference with Probabilistic and Deterministic Dependencies. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 458-463), 2006. Boston, MA: AAAI Press
14. Singla, P. and Mooney, R. Abductive Logic for Plan Recognition. To appear in Proceedings of the Twenty-Fifth National Conference on Artificial Intelligence. 2011. AAAI Press.

Learning Theories for Noun-Phrase Sentiment Composition

Stefanos Petrakis and Manfred Klenner

Institute for Computational Linguistics, University of Zurich, Zurich, Switzerland,
{petrakis,klenner}@cl.uzh.ch,
http://www.cl.uzh.ch/

Abstract. The work presented here is an approach to Sentiment Analysis from a rule-based, compositional perspective. The proposed approach is characterized by three major points: (a) rules are automatically learned from annotated corpora using Inductive Logic Programming and represented as Prolog sets of clauses, (b) the focus is on the noun-phrase (NP) level, and (c) learning is performed on deep-parsed structures. We describe the process of annotating a collection of some 3000 German NPs of medium to quite complex structure, as well as the empirical evaluation of our implementation, in comparison with commonly used classifiers and a handcrafted rule-based system.

Keywords: sentiment analysis, inductive logic programming, machine learning

1 Introduction

Sentiment detection is a young field of research that has been progressing rapidly into maturity during recent years. During its development, the term itself, sentiment detection (or analysis[1]), has conceptually annexed neighboring research fields, namely subjectivity analysis and opinion mining [6]. The work presented here focuses on "pure" sentiment detection, striving to correctly evaluate the polarity orientation (e.g. positive/negative/neutral) of sentiment in text.

The majority of approaches to sentiment detection that have been proposed in the past are (a) focusing on the document level and (b) employing standard statistical machine learning techniques to perform classification of polarity[2]. In contrast to these approaches, a sub-current of research within the field has focused on explicit compositional treatment of sentiment on the sentential as well as the subsentential levels by utilizing more of the available linguistic structures [2], [3], [4]. The compositional nature of sentiment is clearly manifested in common phenomena like negation ('these were not good news'), intensification ('an excellent liar') and diminishment ('an unrealistic hope') among others.
The work presented here embraces this compositional view of sentiment analysis.

[1] In the rest of the paper we use these two terms interchangeably.
[2] [5] provide a detailed list of such approaches.

We describe a method for automatically producing theories of sentiment composition for noun-phrases(NPs). This is a supervised machine learning approach which intends to model relations between the constituents of NPs. From such a model, sentiment on the NP level could then be systematically calculated. While the language of choice for the specific experiment is German, the idea is applicable to any given language. Learning is performed on relationally structured data, a corpus built specifically for this purpose.
In the rest of this paper we shall describe the corpus annotation process as well as the relational learning component our system employs. We compared an implementation of our proposed approach with standard statistical sentiment classifiers as well as a handcrafted rule-based system and will present the results of the empirical evaluation. We will conclude with insights gained and remarks regarding the applicability and significance of this novel idea.

2 Sentiment Composition

The compositional aspect of sentiment can be expressed by the following statement, a slightly modified version of the principle of compositionality[7]:

> *"The sentiment of a complex expression is determined by its structure and the sentiment of its constituents"*

Correctly analyzing the sentiment of a chunk, phrase, sentence or clause can be accomplished by identifying and utilizing the relations between the constituents of the given textual unit. A number of basic examples of such relations on the NP level are given in [4]:

ADJ	NOUN	→ NP	Example
NEG	POS	→ NEG	disappointed hope
NEG	NEG	→ NEG	a horrible liar
POS	POS	→ POS	a good friend
POS	NEG	→ NEG	a perfect misery
POS	NEU	→ POS	a perfect meal
NEG	NEU	→ NEG	a horrible meal

Fig. 1. NP Sentiment Composition

The patterns for sentiment composition of NPs observed in Figure 1 are quite intuitive and one could easily formulate them as rules. For example, the case of "disappointed hope" in the first line is one of many instatiations of the general pattern where a NP comprises a negative (NEG) adjective (ADJ) and a positive noun. However, manually producing such rules for NP sentiment composition is not always going to be a trivial task since NPs do not always come in such simple form.
Consider the following example of a NP:

"...a case of less than extreme violence that was not an expression of aggression, only the reaction to a stressful and fear-inducing situation..."

The interactions between smaller parts of this example like "not" and "expression of aggression" or even bigger ones like "was not an expression of aggression" and "only a reaction..." are essential factors for the overall sentiment that springs out of this complex NP. Expressing the set of all these interactions as a single rule that would capture this specific and also similar cases is a challenging task. There are many more examples that can illustrate how the task of manually producing rules for NP sentiment composition can be mentally-intensive as well as time-demanding.

2.1 Theory vs. Praxis

Identifying relations between words, chunks, phrases or sentences, whether labeled as interactions or dependencies, allows us to model sentiment on a high dimensional space. Whether this is necessary or not, is an open question, one that we tried to answer empirically by measuring the performance of such an implementation.
However, what is of vital importance for our research is the actual theory at hand. We are interested in research that can be applied in the real world, but what we consider as our priority in this specific research attempt is to produce a concise and highly detailed theory of compositional sentiment analysis.

2.2 Focus on the Phrase Level

There are two primary incentives why one would go for sentiment detection on the phrase level:

- There is an actual need for analyzing input at this level. Examples of text at the phrase level are found abundantly and frequently on the web in the form of statusa, tweets and other similar, fragmented, stand-alone units of texts.
- For an integrated system that performs sentiment detection in a bottom-up direction, correctly detecting sentiment on the sentence and higher levels means above all correctly detecting sentiment at the lower levels. We intend to construct such a system in the future which is why having a reliable compositional component for the phrase level is an important requirement.

2.3 Related Work

We consider as related to our work approaches that are either specifically focusing on the phrasal sentiment or are able to handle such input and at the same time accomplish their goal in a entirely or partly compositional way.

The extensive theoretical framework proposed by [2] is manually devised in contrast to ours. The authors also mention that their implementation is based on a non-robust parser which introduces a significant number of errors in the system. As we will describe later, our system is using a robust dependency parser [9].

In [10] polarity classification is based on a number of features that make use of inter-constituent relations. These features are hard-coded and in no case exhaustive. Additionally, their system performs as a first step subjectivity detection (polar/neutral) which we do not consider a crucial step.

The system described in [3] focuses on the sentential and subsentential level and uses among others a set of manually written rules for inferring sentiment in a compositional manner. These rules, in comparison to our approach, apply only to flat structures, i.e. shallow parsed phrases, although the authors hint that deep-parsing could improve their system's performance.

3 Resources

To produce the annotated data needed for learning theories of sentiment composition we: (a) collected a sufficiently large number of NPs, (b) automatically parsed and tagged words with dependency and part-of-speech (POS) information, (c) annotated words with (prior) polarity, (d) selected the most complex and interesting NPs out of the set and (e) manually labeled these with NP polarity information. In the following paragraphs this process is described in more detail.

3.1 Polarity Lexicon

The lexicon we used is based on the manually curated polarity lexicon for German used by the PolArt system [4]. That lexicon has been built using the lexical database GermaNet as reference and contains more than 8,000 words (nouns, verbs and adjectives), which makes it - as of this writing - the most extensive manually produced polarity lexicon for German. Note that the version we used is a modified one as it contains an automatically extended list of adjectives[3] as described in [8].

3.2 Collecting NPs

We decided to use the DeWaC corpus [1] as raw material since it provides a practically unlimited - for our purposes - number of NPs. We parsed and tagged a huge number of sentences from DeWaC using the Pro3Gres parser for German [9]. After that, we extracted only NPs from that set of deep-parsed sentences. These resulting NPs were now containing dependency and POS information and

[3] A necessary extension, as it was found during evaluation that the number of adjectives it originally contained was quite restricted.

we also added polarity information from our polarity lexicon for German. At this point, we had at hand a huge pool of NPs that were annotated with the information we were interested in for the learning phase. But, in order to learn interesting and sufficiently complex rules from the training material two more post-requirements were introduced; we decided we needed NPs that were (a) as long as possible and (b) as complex and rich in polarity as possible.

The first requirement (a) was straightforward to use as a filtering criterion. The second requirement (b) was further formulated as a set of three sub-criteria, picking NPs that contained:

- a SHIfter and one or more polarized word(s)
- an INTensifier and one or more polarized word(s)
- a NEGative and a POSitive word

where a SHIfter is a word (from the polarity lexicon) that can invert the polarity of another word (as in "*not* good") and an INTensifier is a word that can enhance or reduce the polarity of another word (as in "*excellent* liar").

We applied a selective search based on requirements (a) and (b) inside our pool of NPs and finally extracted 4200 NPs ranging from small to greater complexity and containing few to many polarized words.

3.3 Guiding the Annotation

These 4200 NPs were given to a group of annotators, - two annotators assigned per phrase - that were instructed to annotate them for polarity (positive/negative/neutral). The greatest of the challenges was to guide the annotators to adopt a common point of evaluating sentiment. For given phrases, adopting a subjective view of the sentiment captured by the phrase can easily lead to a variety of evaluations. We wanted to avoid such situations and where possible 'enforce' a uniform evaluation and annotation approach.
To help us and the annotators in this we introduced the notion of "political/common-sense correctness" as a means to keep things under control. A simple but good example of where this criterion could prove useful is:

"...eating cold pizza in the morning..."[4]

While the notion of eating cold food is in principle negative, we should not neglect the fact that (a) some dishes are meant to be served cold and, most importantly, (b) some of us simply enjoy a slice of yesterday's pizza in the morning. However, the criterion of "political/common-sense correctness" should stop an annotator from marking such a phrase as positive and in this way maintain some level of uniform way of evaluating sentiment.

The second point that we explicitely demanded from the annotators was to operate in a "context-independent" way. That meant that they should always

[4] Objections to whether there is sentiment at all in this phrase are reasonable. We think of such cases sentimentally interesting, in the sense that they can provoke sentiment.

try to annotate without filling in missing context information and focus only on the words and sentiment that was present in the phrase. We believe this also to help avoid varied responses.

4 Learning Theories for Sentiment Composition

The art of manually writing rules for sentiment composition should not be underestimated. It is however undoubtedly tedious at times and quite demanding most of the times, especially as the complexity of the structures one wants to model increases.
It is therefore desirable to be able to construct such theories automatically.

4.1 ILP and Aleph

Learning relational models from structured data via inductive inference is the focus of the subfield of Machine Learning called Inductive Logic Programming (ILP), a long standing paradigm for inferring sets of rules that model relations. We used the tool named Aleph[5] from [11], which is based on the following standard ILP idea:

- Given logic programs B(ackground) and E(xamples)
- Find a logic program H(ypothesis), $H \in \mathcal{H}$
- Where $B, H \rightarrow E$
- For given P(ositive) examples,
  $\forall e \in P, B \wedge H \models e$
- For given N(egative) examples,
  $\forall e \in N, B \wedge H \not\models e$

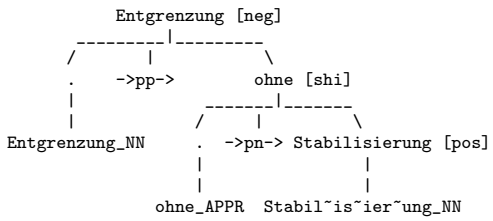
4.2 Example of an Induced Rule

Given the phrase

> "Entgrenzung ohne Stabilisierung"[6]

which was labeled as negative by the annotators, (this phrase serves as a positive example $e \in P$ for negative polarity), and had the following parse tree[7] (this information is part of the background knowledge B)

[5] Available from http://www.comlab.ox.ac.uk/activities/machinelearning/Aleph/aleph.
[6] "Debordering without stabilization" (translation from German).
[7] Word polarity annotations are also visible in square brackets.



Aleph induced the following rule[8] (this rule is part of the induced hypothesis H), for negative NP sentiment composition:

```
np_pol_neg(A) :-
    depends_BonA(A,B), has_pol(B,shi),
    has_pol(A,pos).
```

which reads:

a phrase that is headed by a POSitive noun A,
which dominates a SHIfter B,
should be labeled as negative.

The crucial point of interest is that Aleph in specific and ILP in general, allows us to extract rulesets from resources such as the annotated corpus we prepared. The most important features of these rulesets, or theories, are: (a) they are readable by and therefore comprehensible by humans, an expert could edit and enhance such an automatically learned theory, (b) they are complex enough to justify the usefulness of this whole process, in comparison with a manually edited system, (c) they can be manipulated in interesting ways, e.g. by applying global constraints on a learned theory, something that other tools of the trade don't readily provide and (d) they are theoretically intriguing, they describe sentiment composition in a thorough and detailed way.

5 Empirical Evaluation

We used 2809 NPs from the total of 4200, the ones where the annotators agreed on. This set was divided into a training subset of 2100 NPs and a testing subset of 709 ones. The training subset contained 1391 NPs labeled as being of negative polarity, 142 labeled as neutral and 567 labeled as positive. All of these NPs were used as positive examples ($e \in P$), for their respective classes. The same NPs were grouped together in pairs and used as negative examples ($e \in N$) for their opposing classes, e.g. the positive and neutral polarity labeled NPs were used as negative examples for the negative polarity class. Aleph learned various theories[9] and we selected the best performing one (in comparison). This learned

[8] Expressed in Prolog syntax.

[9] Aleph comes with a multitude of configuration parameters, modifying and tuning each one of them can lead to numerous different theories being produced.

theory was a ruleset of 157 Prolog clauses, 78 for negative NP polarity and 77 for positive NP polarity, ranging from short and simple clauses with a few predicates to longer and more complex ones. No rules for neutral NPs were part of this specific induced theory, a fact we attributed to the low number of positive examples (142) for the neutral polarity class.

5.1 Testing

The theory Aleph learned, an actual Prolog program, was tested on the 709 NPs of the testing dataset. The performances reported can be seen in the following table.

	Precision	Recall	F-Score
NegPol	0.914	0.747	0.822
PosPol	0.388	0.614	0.476

We can easily observe that for the positive polarity class (PosPol) the F-Score is relatively low, influenced by the low precision score. The negative polarity class (NegPol) on the other hand shows a quite satisfying performance. In the next section we compare other systems' performances on the same task.

5.2 Comparing

We chose to compare our system with a number of readily available classifiers as provided by the WeKa toolkit[10] (DecisionTable, SVM, BayesNet, NaiveBayes, ADTree, BFTree, J48, RandomTree, etc.). In order to train these classifiers we converted the training set into feature-vector format maintaining the same information available to Aleph.
In addition, we also tried out PolArt [4], a rule-based system that uses a similar polarity lexicon and a manually produced set of rules. The following table contains the results from running PolArt and various classifiers on the testing dataset.

		Precision	Recall	F-Score
PolArt	NegPol	0.751	0.885	0.813
	PosPol	0.603	0.388	0.472
DecisionTable	NegPol	0.807	0.893	0.848
	PosPol	0.628	0.457	0.529
BayesNet	NegPol	0.783	0.688	0.733
	PosPol	0.394	0.516	0.447
ADTree	NegPol	0.788	0.904	0.842
	PosPol	0.610	0.383	0.471
BFTree	NegPol	0.778	0.872	0.822
	PosPol	0.531	0.367	0.434
NPCompoILP	NegPol	0.914	0.747	0.822
	PosPol	0.388	0.614	0.476

[10] Available from http://www.cs.waikato.ac.nz/ml/weka/.

We observe that the results are comparable with the learned theory's performance (NPCompoILP). The DecisionTable classifier is the best performing classifier and outperforms NPCompoILP, even with a small difference. The rest of the classifiers seem to perform as good if not worse than NPCompoILP.

6 Concluding Remarks

We are aware that the idea of automatically learning theories of sentiment composition presented here is more of a proof of concept that a full grown system. We still need to evaluate more exhaustively our system, especially in order to analyze and improve the performance drop observed for the positive class. It is also imperative to test our method on languages and resources that are commonly used by the community. Finally, we intend to extend this method to other types of phrases, namely verb phrases and move our focus to higher levels of text like sentences.
Overall, we consider the approach presented here as an attractive one. It fulfills our fundamental requirement for an explicit compositional treatment of sentiment. Furthermore, it accomplishes that in an automatic way, without compromising the wish for a concise, linguistically-grounded, complex theory for sentiment composition. At the same time, based on the evaluation results we can see that the suggested approach shows practical competence, at least for the case of NPs. It performs on par with standardly used classifiers which increases even more it's appeal.

Acknowledgments

This work is funded by the Swiss National Science Foundation (grant 100015_122546/1).

References

1. Baroni M., Bernardini S., Ferraresi A., Zanchetta E.: The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3): 209-226 (2009)
2. Moilanen K., Pulman S.: Sentiment composition. Proceedings of the Recent Advances in Natural Language Processing International Conference: 378–382 (2007)
3. Choi Y., Cardie C.: Learning with compositional semantics as structural inference for subsentential sentiment analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing: 793–801 (2008)
4. Klenner M., Petrakis S., Fahrni A.: Robust compositional polarity classification. Proceedings of the Recent Advances in Natural Language Processing International Conference: 180-184 (2009)
5. Tang H., Tan S., Cheng X.: A survey on sentiment detection of reviews. Expert Systems with Applications 36 (7): 10760-10773 (2009)
6. Pang B., Lee L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2 (1-2): 1-135 (2008)

7. Dowty D.R., Wall R.E., Peters S.: Introduction to Montague semantics (1981)
8. Clematide S., Klenner M.: Evaluation and Extension of a Polarity Lexicon for German. Computational Approaches to Subjectivity and Sentiment Analysis, ECAI: 7-13 (2010)
9. Sennrich R., Schneider G., Volk M., Warin M.: A New Hybrid Dependency Parser for German. Proceedings of the Biennial GSCL Conference: 115-124 (2009)
10. Wilson T., Wiebe J., Hoffmann P.: Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing: 347-354 (2005)
11. Srinivasan A.: The aleph manual. Computing Laboratory, Oxford University (2005)

How Many Jokes are Really Funny? Towards a New Approach to the Evaluation of Computational Humour Generators

Alessandro Valitutti

Department of Computer Science,
University of Helsinki, Finland
alessandro.valitutti@cs.helsinki.fi

Abstract. The classic approach to the evaluation of computational humour generators is based on the calculus of the funniness averaged over a random set of generated items. This paper introduces a different approach according to which the key parameter to be evaluated is not the averaged funniness, but the rate of stimuli with a fix value of funniness. The variable employed in the evaluation, called *humorous frequency*, was tested on a tool for the generation of a specific class of puns, through the lexical variation of familiar expressions. In particular, the effect of the use of taboo words (e.g. sex words or insults) on the values of humorous frequency was evaluated. The results are a promising first step towards the wider use of humorous frequency in evaluation of computational humour generators.

Keywords: computational humour, pun generation, humour evaluation

1 Introduction

Recognizing a good joke is easier than judging its author. While the quality of a joke is in its funniness, the quality of a joke teller is in the capability of produce as many good jokes as possible.

The evaluation of humorous productivity is even harder in the case of automatic joke tellers. In the last two decades, in the context of computational humour [15][13][1], several program were developed for the automated generation of humorous texts. Unlike human counterparts, computational humour generators are still incapable to adapt their behaviour to the users response and correspondingly improve the performance. Therefore, the evaluation of these systems can be reduced to the analysis of their output.

The most commonly used variable to evaluate the quality of a single humorous item is *funniness*. It is well known that a humorous stimulus can provoke different responses in the recipient, from a "faint smile" to "loud guffaws" ([11], p.155). So, funniness is generally defined as the variable measuring the different intensity in the mirth elicited by the humorous stimulus. In the case of a set of

stimuli, the classic approach to the evaluation is based on the calculus of the funniness averaged over the set.

This paper introduces a different approach according to which the key parameter to be evaluated is not the averaged funniness, but the rate of stimuli with a fix (typically the highest) value of funniness. In other words, given a set of randomly generated puns, the focus is not in the averaged value of funniness but in the number of puns that are judged to be very funny. A possible consequence is the capability to distinguish between systems producing a high rate of mediocre jokes and those generating a low rate of good jokes. The variable proposed in this work, called *humorous frequency*, was tested on a tool for the generation of a specific class of puns, through the lexical variation of familiar expressions (e.g. movie titles or proverbs). In particular, the effect of the use of taboo words (e.g. sex words or insults) on the values of humorous frequency was evaluated. The results are a promising first step towards the wider use of humorous frequency in evaluation of computational humour generators.

2 State of the Art

To date there are only a limited number of researches on the computational generation of puns. In the context of this work, the term "pun" is employed to denote a short verbal expression produced through linguistic manipulation and recognizable as humorous. In [16], Ritchie provides a systematic review of the most remarkable pun generators developed in the last 20 years. Particularly interesting is the comparative analysis of the evaluation of these prototypes, as described by their corresponding authors. Each work was focused on a different specific type of humorous expressions.

Attardo and Raskin [14] described LIBJOG, a program for the generation of light bulb jokes (e.g. *How many <ridiculed-people>*[1] *does it take to screw a light bulb? Five − one to hold the bulb and four to turn the chair*). Tom Swifty, described by Levison and Lessard [9], produces a quoted utterance joined to a funny remarking phrase (e.g. *'Turn up the heat', said Tom coldly.*). In both systems no evaluation of the output was carried out.

In a similar way as Tom Swifty, HCPP (Homonym Common Phrase Pun) [21] can produce puns composed by a simple sentence and a noun phrase (e.g. *Joan charms a man in the sack. A male bag.*). McKay [12] presented a slightly more sophisticated system, WisCraic, capable to produce a more wide range of puns (e.g. in the form of question-answering). In these two cases, the evaluation was not very detailed. In the case of HCPP, no comparison with a control group was performed, and the criteria for the choice of the pun sample were not explained. The main limitation of WisCraic is that evaluation results are not provided with any measure of significance.

Another project was HAHAcronym [19], whose goal was to develop a system to automatically generate humorous versions of existing acronyms, or else to

[1] This tag denotes a group of people stereotyped and made object of ridicule.

produce a new funny acronym constrained to be a valid vocabulary word, starting with concepts provided by the user (e.g. *MIT = Mythical Institute of Theology*). The funniness of the generated items was evaluated against a baseline sample of acronyms produced without humorous constraints. The significance of the results was indirectly provided (i.e. the difference between the means in the two samples was less than the values of standard deviation, calculated in the two sets). In this case, the variability according to subject was not taken in account and not correspondingly justified. The results are then useful to identify the items averagely judged as funniest but the rate of agreement between subjects is not rated.

The JAPE program [2] produces a specific type of punning riddles (e.g. *How is a nice girl like a sugary bird? Each is a sweet chick.*). The evaluation of this system seems to be the most detailed in the literature of the field. As in the case of WisCraic, two characteristics of the pun generated by this prototype were evaluated: *jokiness* (i.e. the property to be recognised by the hearer/reader as a joke) and *funniness* (e.g. expressing the rate of recipients humour appreciation), both variables were tested against the corresponding control samples. In this case, though, the significance values in the results were provided. Unlike HCPP treatment, the choice of the material was described in detail and, in addition to the HAHAcronym evaluative approach, the variation of funniness (averaged over the subjects) was evaluated through the Wilcoxon Signed Rank Test [6], therefore providing a significance value.

3 Research Question and Proposed Approach

All the above approaches to the evaluation of humour generators provided an essential contribution to the understanding of the "performance" of a pun generator and the "quality" of produced puns. In this work, the exploration focuses on a specific question that perhaps is still underestimated: the frequency of puns with high funniness. In other words, given a set of randomly generated puns of a specific generator, how many items are recognized as highly funny?

The mentioned studies seem not to provide a convincing answer to this question. In the evaluation of JAPE, the hypothesis test applied to the two pun sets (i.e. the control set and the set of puns generated with the humorous constraints) confirms that the value of funniness (averaged on each set and on the sample of subjects) is significantly higher in the second condition. But in that measure it is not possible to discriminate between the weight of items with high funniness and the number of highly rated items. In the HAHAcronym evaluation, the sum of rates of funniness by all subjects was calculated. Also in this case, the average value over the set of puns was calculated.

In this study, the performance of the pun generator is defined as the fraction of funny items in a randomly generated set of puns and called *humorous frequency* (HF). In this context, the funniness is defined as a Boolean variable. The measure can be performed through the same rating employed in other evaluation (i.e. a scale with 4 or 5 values), with a threshold in order to filter the highly rated items.

So, the value HF, calculated over a set of puns and averaged over a sample of subject, can be interpreted as the probability of the pun generator to induce the humorous effect.

In [18], a set of criteria is proposed for the assessment of creativity from the behaviour of computer programs. In particular, there is a correspondence between HF definition and Criterion 4, according to which high quality items should make up a significant proportion of the results. In the hypothesis that funniness can be considered as a specific type of quality, HF can be considered a particular instance of Criterion 4.

4 The FEVer Generator

4.1 Variation of Familiar Expression Variation (FEV)

The FEVer program is designed to generate a specific type of humorous expressions: simple one-line puns (FEVs) obtained through the variation of familiar expressions (e.g., proverbs, movie titles, name of famous persons, etc.). The variation is performed through the replacement of one word of the original expression with a semantically different but phonetically similar word. Examples of FEVs are: *"Tomorrow is another bay"*, *"Back to the Suture"*, and *"Fatal Extraction"*.

Pun generation is then reduced to a process of lexical selection performed according to morphological, phonetic and semantic constraints.

The morphological constraint consists of the requirement that the replacement word has the same part of speech (POS) of the original word. Without this condition, the expression obtained through the replacement can be hardly recognised. In order to perform the POS analysis, the free available tool TreeTagger was employed [2].

In the procedure of lexical selection, the tagset was restricted to 4 tags corresponding to "noun", "adjective", "verb", and "adverb". The reason is that the lexical resources employed for the semantic selection are based on dictionaries in which words are tagged only with these sublist of POS.

4.2 Phonetic Similarity

The phonetic constraint consists of a phonetic similarity and is aimed to induce the recognition of the original familiar expression. In order to consider a FEV communicated via speech, it is better to avoid the case of phonetic identity, and take in account, for the two words in the replacement, the relation of "partial phonetic similarity", called *paraphony* (or *paronymy*).

Two words are paronyms when their phonemic representations are similar but not identical. Paraphony is a specific type of *heterophony* (i.e. the general relation between words with different phonetic expression). *Paronomasia* (or *punning*) is defined as the use of words similar in sound to achieve a specific

[2] `http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger`

effect as humor. Puns created through paronomasia are called *paronomasic*, *imperfect*, or *heterophonic puns* [7]. In general, two paronyms are then perceived as phonetically similar. More specifically, there are different possible criteria according to which the perception of phonetic similarity can occur. In the present context, paronymy is defined according to the specific task of familiar expression recognition. Two words are defined as paronyms (or phonetically similar) if the lexical substitution allows the listener to recognize, with a significant probability, the familiar expression currently employed.

A possible approach for the identification of homophones, heterophones, and paronyms consists of the measure of phonetic distance between words. Apart from the trivial case of homophony (corresponding to phonetic distance 0), it is possible to identify a specific range of paraphony. In this work, phonetic similarity is treated as a dichotomous variable. Phonetic distance, defined as a real value in [0,1], can be mapped to phonetic similarity variables defined in other works as numerical variable [3][5][10]. For instance, homophones are identified here by phonetic distance 0, corresponding to numeric phonetic similarity 1. Another reason for the adoption of a numerical measure of phonetic distance is to provide ranking to quality of expression variation. For example, if two substitutions are characterized by the same values of semantic constraints, the word couple with less phonetic value is preferable because the corresponding familiar expression is more recognizable. Finally, the required definition of phonetic distance needs to be computationally tractable in order to provide an automatic measure. According to these requirements, the approach adopted was based on the notion of *Levenshtein distance*.

Phonetic Distance The algorithm for the measure of the phonetic distance is a specific implementation of the Levenshtein distance [8]. It is based on a sequence of elementary operations applied on the phonetic expression of a word in order to obtain another word. Each step (i.e. application of an operation) is associated to the value of a cost function. The sequence of steps, required to transform the first word in the second one, and corresponding to the minimum total value of cost, defines the distance between two words. Three types of elementary operations are considered: *substitution*, *insertion* and *deletion*.

The cost value associated to the substitution operator was assigned according to the phonetic type, tonic accent, and vowel length. The algorithm reduces the phonetic distance between words to the distance between syllables, and the syllabic distance to the distance between single phonemes, as illustrated below:

– **Distance between phonemes.** Phonemic distance gets values between 0 and 1, according to the phoneme type. For example, the distance between two vowels is lower than the distance between a vowel and a consonant; two dental consonants (e.g. 't' and 'd') have lower distance than a dental and velar (e.g. 't' and 'k'), etc. The comparison between vowels takes in account both the accent and the length: if the vowels have both a tonic accent or are both short vowels, the distance is lower than other cases.

– **Distance between syllables.** syllabic distance is defined as the Levenshtein distance between the two corresponding sequences of phonemes. For example, if two syllables have the same number of phonemes, the main contribution to the distance comes from the phonemic comparison. Instead, if two syllables have different phonemic length, then the difference in the number of syllables may weight more than the phonemic comparison.
– **Distance between words.** Phonetic lexical distance is defined as the Levenshtein distance between the two corresponding sequence of syllables. The decomposition of words in syllables is performed automatically. The distance is normalized to the length of the longest syllabic sequence. In particular, the zero value corresponds to the case of two perfect homophones (e.g. *weight* and *wait*).

For each word, according to the procedures described above, a list of words was created and sorted according to the increasing values of phonetic distance. Given the high number of items on which calculate the distance, the process is time consuming. Therefore the phonetic distance between word pairs and the list of words was indexed and sorted according to the increasing value of phonetic distance. Given a one-syllable word, the phonetic distance with each of the other words was calculated and the list was sorted according to the distance value. Only values until 0.2 were considered because, after a qualitative survey, it was noticed that for higher values the couples of words are perceived as too different.

Phonetic Dictionary The information on mapping between words and their phonetic transcription was extracted from a phonetic dictionary. The CMU pronouncing dictionary (available at `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`) was used. It is a machine-readable pronunciation dictionary for North American English that contains over 125,000 words and their transcriptions. Each transcription represents the phonetic analysis of a word, and it is represented in the dictionary as a mapping from each word to its pronunciation in the given phoneme set. The current phoneme set contains 39 phonemes. Vowels may carry primary or secondary stress.

Caracteristics of the Current Algorithm This specification of Levenshtein algorithm is focused on the task of punning through lexical substitution. For this reason, it is fundamentally different from other specifications corresponding to different tasks such as speech recognition. If there is a rhyme between the original word and the replacement word (e.g. same position of tonic accent and homophony between right-side of words from the stressed syllable), this feature can be very effective for the expression recognition, even if the words have different syllabic length. On the other hand, in speech recognition of a single word the syllabic length can be more important, and then the search of the possible word corresponding to a given phonetic expression has to be performed in the set of words with the same syllabic length.

[10] adopted a similar approach to the implementation of phonetic constraints for punning. A specific part of their research was aimed at improving the phonetic functionalities of the JAPE punning-riddle generator [2] and subsequently integrated in the joke generator STANDUP [17]. More specifically, a new approach, based on the measure of phonetic similarity, followed the previous approach, based on rules. The phonetic tool developed in this work presents some key differences from the corresponding one of Manuring et al. A first group of differences concerns the choice of cost values to be assigned to each phoneme. In STANDUP, each phonetic type was tagged with a set of properties and a corresponding cost value. Thus, the cost function of a phoneme pair is calculated from the individual values of each element. In the system, instead, the cost value was assigned directly to the phonetic pair (e.g. consonant/vowel, some/different consonant group, etc.). Furthermore, values and set of phonetic properties taken in account in the two systems are different. In the present system the comparison between syllables, and words are considered separately. Thus the Levensthein algorithm was applied at two levels, first in the comparison between syllables and then using the resulting cost values for the comparison between words. An algorithm was specifically developed to perform the automatic syllabification of words. At present, no evaluation was performed to compare the performance of the two systems, due to the differences in the corresponding types of pun generation. A future work can be focused on the possible integration of the two approaches.

Another possible improvement consists of taking account, in the measure of phonetic similarity, not only of the word to be replaced but also the words of the expression context. In this way, a good phonetic similarity with contextual words can make the expression recognizable even if the similarity with the target word is not good.

4.3 Database of Familiar Expressions

A crucial resource accessed by FEVer is a set of expressions that are recognized as familiar by English speakers. A specific type of familiar expressions was considered: famous movie titles. 290 titles were selected from the Internet Movie Database (`www.imdb.com`). The list includes titles of the most famous movies in all sorts of categories based on votes from users.

5 Evaluation

The FEVer program was evaluated through the measure of HF (defined in section 3), in different conditions, according to phonetic distance (defined in section 4.2) and taboo-ness. The latter term denotes a boolean variable indicating if a taboo word (e.g. a sex word or a term used as insult) is employed as replacement word in the familiar expression.

It is assumed that, for some type of recipient and in some given context, the use of these words might make the replacement funny, because it has an

active role in the realization of the effect described in relief/release theories of humour [4]. The advantage in the use of taboo words is that there is no need for contextual information: a simple form of humour can be expressed in this case.

5.1 Methodology

A random set of 600 puns was generated. Only word substitutions with the same part of speech and syllabic length were taken into account, in order to preserve well-formedness and help the recognizability of the original title. In half of items (300) the title was obtained using a taboo word in the replacement (*taboo-ness = true*). The subset containing taboo words was further split into 5 clusters, each corresponding to a different range of phonetic distance between the original word and the new word. Phonetic intervals had a length of 0.15, with a range from 0.00 to 0.75. Higher values for the phonetic distance were not considered because, in that range, the new word is perceived as too different to induce the recognition of the original word. An analogous split into 5 subsets was performed for the list of items not including taboo words.

To sum up, 10 clusters of 60 elements were selected. The elements of each cluster were randomly selected. Finally all clusters were randomly mixed to avoid a cluster recognition effect. For example if the subject identifies a series of items as elements of the cluster with low phonetic distance and taboo words, (s)he might propagate the same information to the remaining elements, without really focusing on their content.

A sample of 40 subjects was considered. They were all students and researchers at Twente University in the Netherlands, with a good knowledge of English, and only some being English native speakers. Before providing the questionnaire with the expression list, there was a brief conversation with each participant in order to explain the modalities of the annotation, to be sufficiently sure that reading the expressions with taboo words would not be embarrassing or offensive.

The subjects were required to read the list of puns and, for each of them, select one of the following comments:

1. *It is funny*
2. *It might be funny but not to me or not now*
3. *Not sure*
4. *It is not funny at all*

5.2 Data Analysis

In Table 1, the values of HF according to taboo-ness and ranges of phonetic distance are shown. In the calculation of HF, three different types of puns (and corresponding clusters containing them) were taken into account: puns with taboo words (**Taboo**), puns without taboo words (**Non-Taboo**), and puns with or without taboo words (**All**). As shown, the highest value for HF corresponds to the taboo cluster with the lowest range of phonetic distance.

P-Distance	Taboo	No-Taboo	All
$0.00 - 0.15$	0.210	0.063	0.137
$0.15 - 0.30$	0.101	0.032	0.066
$0.30 - 0.45$	0.065	0.029	0.047
$0.45 - 0.60$	0.061	0.032	0.047
$0.60 - 0.75$	0.047	0.020	0.034

Table 1. Values of HF according to different pun clusters.

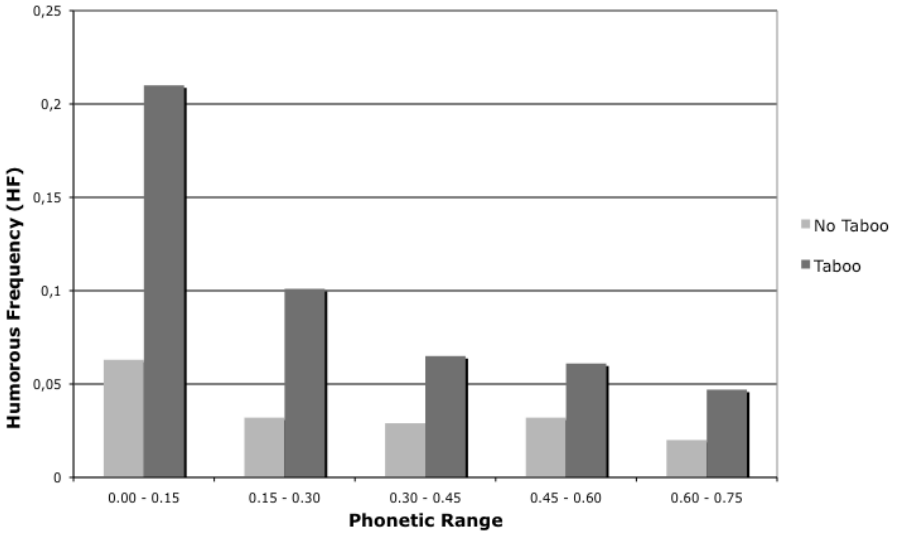


Fig. 1. Graphs of HF corresponding to different ranges of phonetic distance.

The same information is graphically represented in Figure 1 as a multiple histogram, according to the two conditions *Taboo* (i.e. puns with taboo words) and *No Taboo* (i.e. puns without taboo words). To quantify the correlation of HF with phonetic range and taboo-ness, Pearson's coefficient r_P (one-tailed, $p < .02$) was employed. The resulting value is $r_{P-all} = -0.51$, representing a "good" correlation. In the case of taboo cluster, the correlation is slightly higher ($r_{P-taboo} = -0.52$). Instead, in the case of the non-taboo cluster, the correlation is not confirmed ($r_{P-no-taboo} = -0.29$). In Table 2 some example of puns judged by subjects as funny are listed. The puns are sorted according to agreement (defined here as the number of subject recognising the same pun as humorous). It is notable that, for most of the items shown, high agreement corresponds to low phonetic distance and presence of taboo words (i.e. taboo-ness = 'yes').

Pun	Agreement	Substitution	P-Distance	Taboo-ness
Woman: Impossible	22	mission/woman	0.22	no
The Lost World : Jurassic Porn	19	park/porn	0.12	yes
Passion: Impossible	19	mission/passion	0.13	yes
Lust Busters	16	ghost/lust	0.13	yes
The Sexed Sense	16	sixth/sexed	0.13	yes
Finding Homo	16	nemo/homo	0.13	yes
Ass Age	15	ice/ass	0.04	yes
Lara Croft: Tomb Rubber	14	raider/rubber	0.12	yes
Kissing: Impossible	14	mission/kissing	0.26	yes
Forrest Dump	12	gump/dump	0.10	yes
How to Light a Guy in 10 Days	12	lose/light	0.30	no
Notting Feel	10	hill/feel	0.15	yes

Table 2. Puns sorted according to positive agreement.

In order to make this study more directly comparable to previous research, a further analysis, similar to the one achieved in the evaluation of JAPE program [2] was performed. In this case, only items generated within the lowest range of phonetic distance were considered. Two sets corresponding to 'no-taboo' and 'taboo' conditions were prepared. Finally, the Wilcoxon Signed Rank Test [6] was applied to the averaged HFs. The results confirm that the taboo-ness significantly increases HF ($p < .001$).

6 Conclusions and Future Work

The introduction of HF in the evaluation of verbal humour generators is the central contribution of this work. In particular, the use of this variable has two aimed methodological consequences. The first one is the attention shift from the humorous text, produced by the tool, to the capability to generate humour. The second consists of the adoption of a causal and probabilistic point of view in the evaluation of humour generators. HF can in fact be interpreted as a measure of the probability to generate an expression with a fixed (typically high) intensity of funniness, and therefore induce the corresponding humorous effect.

HF has been tested on a particular generator, the FEVer program, and employed to evaluate the contribution of two lexical parameters (phonetic distance and taboo-ness) to the performance of the system. In this context, the experiment is treated as a use case of the proposed evaluative approach. More generally, this methodology can be employed in the evaluation of a broader class of verbal humour generators.

A next achievement will be the development of a gold standard corpus of humorous expressions, in order to evaluate the possible upper bound in the value of HF. Obviously, this value can be useful to explore to what extent it is possible to generate humorous texts without taking account the context in which

they are communicated. For example, a potentially good joke can be appreciated only by a specific type of recipients. Furthermore, a humorous expression should be communicated in the appropriate conversational context. In that case, the measure of HF would not be sufficient for the evaluation of a context-dependent humour generator.

The information provided by HF can be exploited in the investigation of both automatic and assistive scenarios of humour generation. An interesting case of automatic scenario is provided by the use of pun generation in a conversational context. In friendly conversations people can sometimes fail in their purpose to be humorous when communicating a wittiness, and might attempt a few times before telling something agreed by they recipients as funny. In the context of interaction with a humorous conversational agent, users might tolerate a certain number of mistakes in the generation of humorous utterances. Consequently, a maximum threshold for HF can be fixed.

Humour generators can be used as assistive tools for the human creation of humorous texts. For example, a copywriter may want to check a list of randomly generated expressions, in order to select the most suitable item for the creation of an advertising slogan. In this context, the knowledge of HF can help the user to decide the maximum number of items to be examined in the current session.

Therefore, the evaluation of humour generations based on HF can be a useful way for exploring the improvement of these tools towards new challenging creative tasks.

Acknowledgements

I would like to express my gratitude to Carlo Strapparava and Oliviero Stock for their academic supervision and personal support. I am grateful to Anton Nijholt and all members of Human Media Interaction group for their participation to the evaluation experiment. Finally, I would express my gratitude to Graeme Ritchie for his precious comments to an earlier draft.

References

1. Binsted, K., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., O'Mara, D.: Computational humor. Computer Society 21(2), 59–69 (2006)
2. Binsted, K., Pain, H., Ritchie, G.: Children's evaluation of computer-generated punning riddles. Pragmatics and Cognition 2(5), 305–354 (1997)
3. Crestani, F.: Using semantic and phonetic term similarity for spoken document retrieval and spoken query processing. In: Bouchon-Meunier, B., Gutierrez-Rios, J., Yager, R. (eds.) Technologies for Constructing Intelligent Systems, pp. 363–376. Springer-Verlag, Heidelberg, Germany (2002)
4. Freud, S.: Der Witz und Seine Beziehung zum Unbewussten. Deutike, Leipzig and Vienna (1905)

5. Garcia, P., Rubio, A., Diaz-Verdejo, J., Benitez, M., Lopez-Soler, J.: A transcription-based approach to determine the difficulty of aspeech recognition task. IEEE Transactions on Speech and Audio Processing 7(3) (1999)
6. Greene, J., D'Oliveira, M.: Learning to use statistical tests in psychology. Open University Press, Milton Keynes, UK (1992)
7. Hempelmann, C.F.: Paronomasic Puns: Target Recoverability Towards Automatic Generation. Ph.D. thesis, Purdue University (2003)
8. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
9. Levison, M., Lessard, G.: A system for natural language generation. Computers and the Humanities 26, 43–58 (1992)
10. Manurung, R., Ritchie, G., Pain, H., Waller, A., Black, R., O'Mara, D.: Adding phonetic similarity data to a lexical database. Language Resources and Evaluation 42(3), 319–324 (September 2008)
11. Martin, R.A.: The Psychology of Humor: An Integrative Approach. Elsevier (2007)
12. McKay, J.: Generation of idiom-based witticisms to aid second language learning. In: [20] (2002)
13. Mulder, M., Nijholt, A.: Humour research: State of the art. Tech. rep., University of Twente, The Netherlands (2002)
14. Raskin, V., Attardo, S.: Non-literalness and non-bona-fide in language: approaches to formal and computational treatments of humor. Pragmatics and Cognition 2(1), 31–69 (1994)
15. Ritchie, G.: Current directions in computational humour. Artificial Intelligence Review 16(2), 119–135 (2001)
16. Ritchie, G.: The Linguistic Analysis of Jokes. Routledge, London (2004)
17. Ritchie, G., Manurung, R., Pain, H., Waller, A., Black, R., O'Mara, D.: A practical application of computational humour. In: Cardoso, A., Wiggins, G.A. (eds.) Proceedings of the 4th International Joint Conference on Computational Creativity. pp. 91–98. London (2007)
18. Ritchie, G.: Assessing creativity. In: Wiggins, G.A. (ed.) Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science. pp. 3–11 (2001)
19. Stock, O., Strapparava, C.: HAHAcronym: Humorous agents for humorous acronyms. In: [20] (2002)
20. Stock, O., Strapparava, C., Nijholt, A. (eds.): Proceedings of the The April Fools Day Workshop on Computational Humour (TWLT20). Trento (2002)
21. Venour, C.: The computational generation of a class of puns. Master's thesis, Queen's University, Kingston, Ontario (1999)

Entropy-Driven Evaluation of Models of Eye Movement Control in Reading

Mattias Nilsson and Joakim Nivre

Department of Linguistics and Philology, Uppsala University,
Box 635, 751 26 Uppsala, Sweden
{mattias.nilsson,joakim.nivre}@lingfil.uu.se

Abstract. This paper presents a novel method for the evaluation of probabilistic models of eye movement control in reading based on the idea of computing the test sample entropy, relative to a model, instead of the accuracy of argmax prediction. We relate the notion of entropy to probabilistic saccade models and show how entropy-based metrics can be applied to evaluate such models, similarly to how language models are assessed in natural language processing applications. To demonstrate the approach, a probabilistic model of saccade targeting is presented and evaluated against a large eye-tracking corpus. The uncertainty associated with the observed eye movement behavior, as perceived by the model, is measured and reported. The average perplexity per fixation is reduced by 29%, in comparison to a naive baseline in this pilot demonstration.

Keywords: Cognitive modeling; Eye movements; Reading; Entropy; Language modeling

1 Introduction

Advances in computational modeling of eye movement control in reading help to gain a better understanding of how the perceptual, cognitive, and motor processes that guide the readers' eyes interrelate and unfold in real time [1–8]. Although current eye movement control models often make different assumptions with respect to the relative strength of the influence of these processes on eye movements, it is commonly acknowledged, however, that making an eye movement involves two basic decisions that computational models must reflect on, the decision to initiate a saccade, and the selection of a visual target for the saccade. Thus, given some text as input, the goal of these models is to generate predictions for the duration and location of fixations, in approximation to human reading behavior. A crucial question, then, is what it means for a model to *approximate* human reading behavior, and how to establish evaluation methods and criteria that correspond to this goal.

Generally, the evaluation of computational models of eye movement control in reading is based on fitting model predictions to empirical observations. The quality of a model, however, is usually only assessed with respect to the same behavioral data that is being used to tune the parameters of the model. The

ability to perform well on independent or *held-out* data is usually never tested and the potential risk of overfitting the models has not been of much concern to date. In other words, current models of eye movements in reading are typically not being evaluated with respect to generalizability and hence we know relatively little of how well these models approximate human reading behavior in general, i.e., with respect to their predictive performance on new and previously unseen data.

Nilsson and Nivre [9] address these problems in relation to a model of saccade targeting in reading based on supervised machine learning techniques. To avoid potential problems of overfitting they maintain a strict separation of training, validation and test data through model development and evaluation, which makes it possible to assess the generalization error of the model. One potential drawback of their approach, however, is the attempt to evaluate exact predictions for the saccade patterns of an individual reader. We believe this is an unreasonable approach given the large variation that characterizes normal reading behavior, both between and within readers [10]. Predicting the actual eye movements that an individual reader makes while reading a new text is arguably a hard problem, and given that even one and the same individual is unlikely to produce the same saccade behavior over different readings of the same text it may not even be meaningful to attempt to do so.

In this paper we attempt to overcome this problem by using a novel approach for assessing probabilistic models of eye movement control in reading based on the information-theoretic notion of entropy. The method we propose is to evaluate such models, not with respect to their argmax function for prediction, but with respect to the entropy they assign to a given test sample. The basic idea is simple enough: a "good" model of eye movement control is a model that assigns high probability, and thus low entropy, to representative data. The entropy of a test sample relative to a model, in this context, measures the uncertainty the model assigns to the observed saccade behavior. The lower the uncertainty, the better the model approximates eye movement behavior. In other words, measuring the entropy of a model is a way of assessing how similar the model behavior is to the observed human behavior. While this approach crucially relies on the principle of keeping training and test data apart, it also overcomes the prediction problem of Nilsson and Nivre [9]. In this paper we demonstrate the approach by implementing a simple probabilistic model of saccade targeting and evaluating it using entropy-based measures against the Dundee eye tracking corpus [11]. It is worth emphasizing, however, that in this study we focus only on models of *where* the eyes move during reading, and we will not be concerned with the temporal aspect of how long the eyes remain paused at fixated words.

The rest of this paper is structured as follows. Section 2 provides a simple characterization of probabilistic saccade models as models that assign probability scores to fixation sequences over the words in a text. In section 3, we relate the notion of entropy to probabilistic saccade models, and show how entropy-based metrics can be applied to evaluate these models. In section 4 we turn to demonstrate the approach by implementing and evaluating a probabilistic model

of saccade targeting in reading. The results of the experiments are then reported and discussed in section 5. Section 6 concludes the paper.

2 Probabilistic Saccade Models

Experimental findings in eye movement and reading research suggest that eye movements in reading are both goal directed and discrete [12]. This means that the saccadic system selects visual targets on a non-random basis and that saccades are directed towards particular words rather than being sent a particular distance. As there is no particular combination of perceptual or linguistic factors which ensures that a certain word will be selected as the target for a saccade, we may think of the decision of where to send the eyes as a probabilistic process influenced by a number of different variables [13, 3]. Under this view, there are a number of candidate words during any fixation, each having a certain chance of being selected as the target for the subsequent saccade.[1] For our purposes we will assume, without imposing any further constraints, that a probabilistic saccade model assigns probabilities to fixation sequences resulting from saccade generation over the words in a text. Let us use the following simple representations of a text and a fixation sequence. Let a text T be represented as a sequence of word tokens $w_1, w_2, \ldots, w_n$, and a fixation sequence F for T be represented as a sequence of token positions in T, $i_1, i_2, \ldots, i_m$. For example, the short text *John gave Mary the book* is represented by $T =$ *John*, *gave*, *Mary*, *the*, *book*; and a fixation sequence over this sentence corresponding to *John* – *gave* – *John* – *Mary* – *book* is represented by $F = 1, 2, 1, 3, 5$. The requirement of a probabilistic saccade model, as construed here, is to compute a probability score for any arbitrary fixation sequence F, conditional on some text T:

$$P(F|T) = P(i_1, i_2, \ldots, i_m|T) \tag{1}$$

At some level of abstraction, then, the task of a probabilistic saccade model is similar to the task facing language models in natural language processing. The goal of a language model is generally to determine the probability of any sequence of words $w_1, \ldots w_n$, $P(w_1, \ldots, w_n)$ and the standard methods for assessing the quality of such models are based on information-theoretic notions like entropy and perplexity [15]. Inspired by the use of entropy-related measures for assessing language models, we attempt to justify in the next section our use of these measures for evaluating probabilistic saccade models of reading behavior.

3 Entropy-Driven Evaluation Metrics

The entropy H of a discrete random variable X with the probability mass function $p(x)$ is defined as:

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i) \tag{2}$$

[1] We assume here that the observed fixation distribution corresponds to the planned saccade target distribution. This is a simplification, however, since saccades may overshoot close targets and undershoot far targets [14].

where $p(x_i)$ is the probability of the event x_i. Shannon [16] defined entropy as the average minimum number of bits needed to encode a string of symbols or transmit a message based on some probability distribution. More generally, entropy is viewed as a measure of the *uncertainty* or *randomness* associated with a random variable and can be interpreted as quantifying the expected or average surprise over all possible events. Lower values of entropy indicate less average surprise, or likewise, higher average certainty.

Let us assume that eye movement behavior can be associated with a random variable X and a set of *events* whose probabilities of occurrence are $p(x_1), p(x_2), \ldots, p(x_n)$. Let us further assume that we would like to measure how much uncertainty is associated with the particular eye movements events, i.e., we are interested in knowing the entropy $H(X)$ of the probability distribution over the events. Because the true probability distribution, $p(x)$, of the random variable is not available to us, we can not *know* the true entropy of X. However, given some model $m(x)$ of $p(x)$, where $m(x)$ is some probability distribution defined over the same event space as $p(x)$, we can approximate the true distribution $p(x)$ and thus also $H(X)$. One way to conceive of this is as as if we are evaluating our model, the estimated probability function $m(x)$, with respect to the true probability function $p(x)$.

The above reasoning can be applied to probabilistic saccade models. In this case the random variable ranges over a sequence of fixations (defined over the tokens in a text), and the probability mass function $m(x)$, which is an approximation to the true, but unknown, saccade model $p(x)$, is given by our probabilistic saccade model. What we still need to determine, however, is how to assess the fit of $m(x)$ to $p(x)$, when the true probability distribution $p(x)$ is unknown.

There is a useful and well-known variant of entropy called cross entropy which allows us to compare probability distributions of a random variable X. By replacing the surprise value in the definition of entropy with an estimate derived from a model $m(x)$ we get the cross entropy of the model $m(x)$ with respect to the true model $p(x)$:

$$H_c(p, m) = -\sum_i p(x_i) \log_2 m(x_i) \tag{3}$$

The cross entropy $H_c(p, m)$ is an upper bound on the true entropy $H(p)$, which means that the cross entropy of a model $m(x)$ on some distribution $p(x)$ is greater than or equal to the actual entropy of the true distribution $H(p)$:

$$H(p) \leq H_c(p, m) \tag{4}$$

In principle, then, the cross entropy can be used as a model evaluator for probabilistic saccade models on the assumption that better models will have lower cross-entropy. For example, given two models m_1 and m_2 of p, the more accurate model will be the one with lower cross entropy. It still seems, though, that we can not calculate this quantity without knowing p. That is, the cross entropy, $H_c(p, m)$, presupposes that the true probability distribution, $p(x)$, is known. The

standard way to get around this problem is to estimate the true probability distribution using a representative test sample, which in our case is equivalent to a sample of eye movements made over the reading of some text. The cross entropy can then be approximated by a quantity often referred to as the logprob or LP for short. Thus, given a test sample of a sequence of fixations F over some text T, the approximation to the cross entropy of a model M on F is given as:

$$LP(M, F, T) = -\frac{1}{n} \log_2 p(F|T) \tag{5}$$

The formula in equation 5 gives the entropy, or average surprise, of the test sample, relative to the model M, fixation sequence F and text T. For convenience, we will simply write $LP(M)$ when the text and fixation sequence is given in the context or irrelevant. Thus, the formula can be interpreted as the average surprise associated with the saccadic movements in the test sample, reflecting from the models' point of view the predictability of a saccade before it occurs, or equivalently, the surprisal perceived just after the saccade occurred. Intuitively, a better model assigns lower logprob to the test sample, being less surprised on average. In other words, a general goal for any saccade model is to keep the average surprise with respect to the observed behavior in the test sample as low as possible. Notably, the use of this approximation to cross entropy for model evaluation relies crucially on the sampling assumption. That is, we must use a training sample to learn the parameters of the saccade model, and then a different but representative sample to test or evaluate the model.

In the general case when we regard a fixation sequence over the entire text simply as a single string, the logprob gives the average surprise in bits per fixation. This text-based computation of the logprob thus takes into account saccades that span sentence-boundaries. This means, for example, that the probability of the first fixation in a sentence (except the first sentence) will be based on the last fixation in the previous sentence. Sometimes, however, it may be more appropriate to measure the entropy on a per-sentence basis, if, for instance, the eye tracking data we use consist of sentences that were displayed in isolation or without any connecting discourse. In this case we may sum the log-likelihood for each individual fixation sequence that corresponds to a sentence, and then divide the total sum over all sequences by the number of sentences. In this sentence-based computation of the logprob, fixation sequences (and sentences) are instead assumed to be probabilistically independent of each other. Note also that in this case we get a macro average of the entropy per fixation sequence and sentence.

An alternative, but equivalent measure to entropy which is commonly reported for language models is the model perplexity. The perplexity of a model M with respect to a test sample is defined as 2 raised to the power of the logprob:

$$PP(M) = 2^{LP(M)} \tag{6}$$

Better models of the unknown probability distribution will assign lower perplexity to the events in the test sample. Perplexity, like entropy and logprob, can be computed per fixation or per sequence (sentence). It is worth noticing that

the metrics we have discussed here are all defined in terms of probability and in principle we may just as well use the probability of the test sample relative to some model as a metric, although entropy and perplexity scores are often considered easier to interpret.

To sum up, we have described how the information-theoretic quantities entropy and perplexity can be applied as evaluation metrics for probabilistic saccade models in reading. This use is based on the assumption that the models are evaluated on a representative test sample which is different from the sample that was used to train the parameters of the model. The intuition underlying the use of these measures is that the more a model captures of the variation in human saccade behavior, the lower the entropy or logprob will be. Finally, it is worth emphasizing that the general use of entropy and perplexity for model evaluation is widespread in natural language processing and artificial intelligence and is thus not new in itself. It is the application of these metrics to eye movement models that is the novel avenue we explore here.

4 A Model of Saccade Targeting

In order to demonstrate the entropy-based approach to model evaluation we turn now to present, in brief, a simple probabilistic model of saccade targeting which we then experimentally evaluate in section 5. The model we implement learns to relate, for a given reader, a set of local visual, linguistic and history-based properties over the reading of texts to five different *types* of saccades in reading. The parameters of the model are learnt by training a probabilistic classifier on authentic eye tracking data. From a cognitive modeling point of view, one may ask if this approach bears any relation to how we believe human saccade strategies are learnt. At some higher level of abstraction we think the approach is not unreasonable. More precisely, the intuition behind the approach we implement here is that saccade targeting is a probabilistic process guided by and shaped through years of past reading practice and experience. Here, this learning process is approximated and automated using machine learning methods. In relation to previous models of eye movement control in reading, this model of saccade targeting is less elaborate but similar in spirit to the spatial system implemented in SERIF [3]. Our goal here, however, is to explore the use of entropy as an evaluation metric using a simple probabilistic model as the basis for assessment.

4.1 Saccade Types

Given a sample of reading data, the model learns to categorize saccades of different lengths into one of five classes based on the type of the saccadic eye movement:

1. Forward
   - Move forward to next token

Feature	Description
Length	Length in characters
Frequency	Normalized word frequency (per million occurrences)
2-gram probability	Conditional probability given preceding token
3-gram probability	Conditional probability given two preceding tokens
PoS	Part of speech
Surprisal	Syntactic surprisal
Distance	Length, in number of tokens, of previous saccade
isFirstFixation	True if token was not fixated before

Table 1. Features defined over the fixation context.

2. Regress
   - Regress to any token
3. Refixate
   - Fixate current token
4. Skip
   - Move forward to the token to the right of the next token
5. Other
   - Move forward to any other token

4.2 Learning

The parameters of the saccade model are estimated by training a multinomial logistic regression classifier [17] on eye tracking data. Training instances used for estimating the model parameters have the form $(f(c), t)$ where

1. $f(c)$ is a feature vector representation of a fixation context c.
2. t is the observed saccade type out of c, t $\in$ {forward, regress, refixate, skip, other}.

The fixation context c of the feature vector representation $f(c)$ spans over two tokens to the left of fixation, the currently fixated token, and three tokens to the right of fixation. This is a rough approximation to the asymmetry of the perceptual span in reading, which extends further to the right of fixation (when reading from left to right). A set of features defined over the fixation context captures both visual, linguistic and history-based properties of the current fixation. The features used in this model are given in table 1. Word frequencies are based on occurrences in the British National Corpus (BNC) and bigram (2-gram) and trigram (3-gram) probabilities are based on the Web 1T 5-gram corpus [18]. The part of speech tags were derived using the part-of-speech tagger TnT [19], and the surprisal estimates were computed from an incremental top-down parser using a probabilistic context-free grammar (PCFG) [20]. The surprisal at a word w_i refers to the negative log probability of w_i given the preceding words, computed here using the prefix probabilities of the parser. A number of recent studies

in psycholinguistics have established a positive relation between surprisal and word-reading times [21, 22, 20]. It is less well known, however, if surprisal may also influence the decision of *where* to move the eyes. Both the part-of-speech tagger and the parser were trained on the Wall Street Journal portion of the Penn Treebank [23].

4.3 Evaluation Algorithm

After estimating the regression parameters for a given reader, we evaluate a model by computing the probabilities assigned by the model to the observed saccadic movements made by the reader on a held-out test set. More specifically, we apply the evaluation algorithm in figure 1 to compute the logprob of the model over the independently drawn test sample F of fixations $i_1, \ldots, i_n$. The algorithm assumes two functions, $c(d)$ and $m(t)$, where the former returns the saccade type, t, associated with the saccade distance, d, between the current and the next observed fixation, and the latter function returns the probability p, relative to the model, of the saccade type t. We start by initializing the logprob

```
E(i_1, ..., i_n)
1:    lp ← 0
2:    k ← 0
3:    while k < n
4:        d ← (i_{k+1} − i_k)
5:        t ← c(d)
6:        p ← m(t)
7:        lp ← lp + − log_2 p
8:        k ← k + 1
9:    return lp/n
```

Fig. 1. Algorithm used to evaluate model.

(lp) to 0 and k to 0 (a hypothetical *null* fixation). As long as there are more fixations we compute the distance to the next target word, d, and apply $c(d)$ to get the associated saccade type t and then use the function $m(t)$ to find the probability for the observed target p. We sum the surprise on seeing the target t (the negative log of p) to the current lp and once the loop terminates, the average logprob of the test sample is returned.

5 Experimental Evaluation

We performed a set of experiments using the model outlined in the previous section and data from the English section of the Dundee corpus. This data set contains the eye tracking records for ten different individuals reading newspaper

editorials from The Independent newspaper. The corpus consists of 20 different texts and approximately 2400 sentences. All 20 texts were read by all ten individuals and the eye movements were recorded using a Dr. Bouis Oculometer Eyetracker, sampling the position of the right eye every millisecond (see Kennedy and Pynte, 2005, for further details).

For the experiments reported here, we used the first 16 texts for training and the following two texts, 17 and 18, for evaluation. Each reader in the corpus was modeled individually. Even though all of the readers read all the texts in the corpus, not all of the readers necessarily read all the sentences in the texts. This means that some of the sentences in the corpus did not receive any fixations at all. In addition, readers quite regularly made one or more off-screen saccades which interrupted the normal sequence of saccades and reading fixations. To minimize any effect of these inconsistencies we removed sentences that did not receive any fixation, as well as sentences containing any sequence of three or more off-screen saccades. Therefore, the total number of sentences in the test sample are not the same between the different readers (the size of the test samples range between 207 and 222 sentences).

5.1 Baseline

To provide some basis for comparison we use a simple and naive baseline model with a uniform probability distribution over the five saccade types. In other words, the baseline model naively assumes that all saccade types are equally probable, the intuition being that a model moving its eyes at random should assign higher entropy to the test sample in comparison to a model that is more informed of how linguistic and other features may influence eye movement decisions.

5.2 Results

Each of the ten models was evaluated with respect to the entropy assigned to the test sample and with respect to the entropy reduction on the test sample relative to the baseline.[2] Table 2 shows the results, reported both in number of bits per fixation and per sentence for all ten models. The final row also gives the average scores across all models.

As the table shows, the entropy per fixation ranges from 1.6 bits in the best case (reader i) to 2.1 bits in the worst case (reader g), with an average of 1.8 bits over all ten models. This is equivalent to an average model perplexity of 3.48 per fixation. In other words, the average model is as confused or *perplex* with regard to the test sample as if it was choosing between 3.48 equiprobable saccade moves at each fixation.

At a closer look, one might potentially want to attribute the difference in entropy between the best and the worst model to the large difference in the total

[2] We use the term entropy in the following discussion, although what we are actually referring to is an approximation to the cross entropy, i.e., *logprob*.

Reader	#Sentences	#Fixations	Fixation		Sentence	
			Entr.	Entr. R.	Entr.	Entr. R.
a	211	2807	1.9	0.44	25.1	5.8
b	217	3745	1.7	0.63	29.2	10.9
c	221	3754	1.8	0.60	29.2	10.3
d	211	4252	1.8	0.55	35.7	11.1
e	207	2399	1.8	0.57	20.4	6.5
f	216	3924	1.9	0.46	33.8	8.3
g	211	4601	2.1	0.24	45.4	5.2
h	207	3498	2.0	0.34	33.5	5.7
i	222	2906	1.6	0.74	20.7	9.7
j	219	3976	1.8	0.49	33.2	8.9
Average	214.2	3616	1.8	0.51	30.6	8.2

Table 2. Test results for all individual models of readers *a*-*j*; number of bits per fixation and per sentence (Entr. = Entropy, Entr. R. = Entropy Reduction).

number of fixations made by these two readers. As shown in the table, reader g has the highest fixation frequency (4601) of all readers while reader i has the second to the smallest fixation frequency (2906). The relation between entropy and fixation frequency is unlikely that simple, however. Thus, for example, reader d has a substantially higher fixation frequency (4252) than reader a (2807). Yet the model of reader d gets roughly the same entropy score (1.8), even a bit lower than the model of reader a (1.9). In fact, there is only a moderate correlation between the fixation frequency and the average surprise per fixation (Pearson's r: 0.41).

Turning now to consider the entropy reduction per fixation, we first observe that all of the models reduce the uncertainty of the test sample; the entropy reduction ranges from 0.74 bits in the best case to 0.24 bits in the worst case, again for reader i and g, respectively. The average entropy reduction per fixation across all models is 0.51 bits which is equivalent to a perplexity reduction of about 29%, in comparison to the naive baseline model used here.

The results computed on a per sentence basis show that the average surprise per sentence (fixation sequence) across the readers is 30.6 bits, which essentially tell us something of how hard it is to predict the exact sequences of saccade moves that readers make over sentences. Indeed, 30.6 bits of entropy is roughly equivalent to the password strength of a five-character randomly generated password, using only (case-insensitive) alphanumeric symbols. The average reduction in uncertainty per sentence is 8.2 bits compared to the baseline, which corresponds to a perplexity reduction of 294 per sentence.

In figure 2, we decompose the results further and show the entropy per fixation grouped by saccade type, averaged over all models. The plot provides a useful summary of the relative difficulty in modeling different types of saccades. As seen here, the models seem reasonably good at predicting forward-directed saccades, i.e., *Forward*, *Skip*, and *Other*, but worse at predicting refixations, and

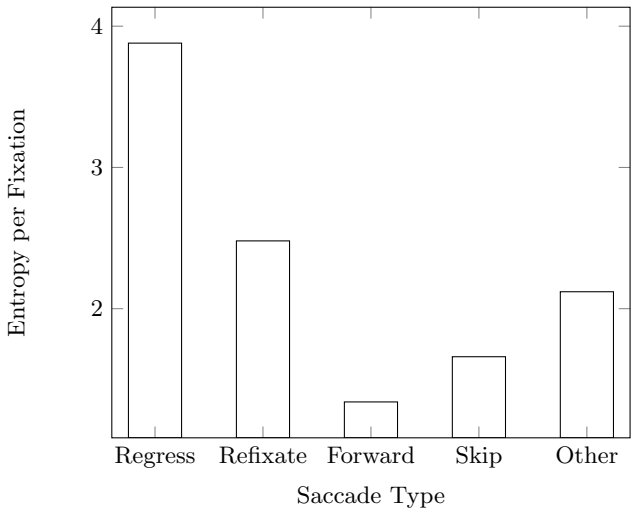


Fig. 2. Entropy per fixation grouped by saccade type averaged over all models of readers *a-j*.

regressions in particular. The average surprise associated with refixations and regressive saccades are 2.48 and 3.88 bits, respectively.

6 Conclusion

We have presented a novel method for evaluating probabilistic models of eye movement control in reading inspired by standard evaluation methods in natural language processing applications. This method is based on the assumption that the lower the entropy of a test sample, relative to a probabilistic saccade model, the better the model is in approximating human reading behavior. This approach overcomes a limitation of previous classifier-based saccade models by assessing how surprised a given model is in seeing the observed eye movements, rather than assessing how often the model predicts the exact same behavior as is observed. We demonstrated the basic approach by implementing and evaluating a model of saccade targeting that achieved an average 29% perplexity reduction on held-out data compared to a naive baseline.

References

1. Reichle, E., Pollatsek, A., Fisher, D., Rayner, K.: Toward a model of eye movement control in reading. Psychological Review **105** (1998) 125–157
2. Engbert, R., Nuthmann, A., Richter, E., Kliegl, R.: SWIFT: A dynamical model of saccade generation during reading. Psychological Review **112** (2005) 777–813

3. McDonald, S.A., Carpenter, R., Schillcock, R.C.: An anatomically-constrained, stochastic model of eye movement control in reading. Psychological Review **112** (2005) 814–840
4. Feng, G.: Eye movements as time-series random variables: A stochastic model of eye movement control in reading. Cognitive Systems Research **7** (2006) 70–95
5. Reichle, E., Rayner, K., Pollatsek, A.: A cognitive-control, serial-attention model of eye-movement control during reading. Cognitive Systems Research **7** (2006) 4–22
6. Reilly, R., Radach, R.: Some empirical tests of an interactive activation model of eye movement control in reading. Cognitive Systems Research **7** (2006) 34–55
7. Richter, E., Engbert, R., Kliegl, R.: Current advances in swift. Cognitive Systems Research **7** (2006) 23–33
8. Yang, S.: A oculomotor-based model of eye movements in reading: The competition/activation model. Cognitive Systems Research **7** (2006) 56–69
9. Nilsson, M., Nivre, J.: Learning where to look: Modeling eye movements in reading. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009). (2009) 93–101
10. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin **124** (1998) 372–422
11. Kennedy, A., Pynte, J.: Parafoveal-on-foveal effects in normal reading. Vision research **45** (2005) 153–168
12. Radach, R., McConkie, G.: Determinants of fixation positions in reading. In Underwood, G., ed.: Eye guidance in reading and scene perception. Oxford, England: Elsevier (1998) 77–100
13. Brysbaert, M., Vitu, F.: Word skipping: implications for theories of eye movement control in reading. In Underwood, G., ed.: Eye guidance in Reading and Scene Perception. Elsevier science Ltd. (1998) 124–147
14. McConkie, G., Kerr, P., Reddix, M., Zola, D.: Eye movement control during reading: I. The location of initial eye fixations on words. Vision Research **28** (1988) 1107–1118
15. Jurafsky, D., Martin, J.H.: Speech and Language Processing. Prentice-Hall (2000)
16. Shannon, C.E.: A mathematical theory of communication. Bell System Tech. J. **27** (1948)
17. Witten, I.H., Eibe, F.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
18. Brants, T., Franz, A.: Web 1T 5-gram Version 1. Linguistic Data Consortium. (2006)
19. Brants, T.: TnT – a statistical part-of-speech tagger. In: Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP). (2000)
20. Roark, B., Bachrach, A., Cardenas, C., Pallier, C.: Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2009) 324–333
21. Boston, M.F., Hale, J., Kliegl, R., Patil, U., Vasishth, S.: Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. Journal of Eye Movement Reasearch **2** (2008) 1–12
22. Demberg, V., Keller, F.: Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition **109** (2008) 193–210
23. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19** (1993) 313–330

Evaluation on Lexical Category Acquisition

Bichuan Zhang[1], Xiaojie Wang[1], Guannan Fang [1],

[1] Center for Intelligence science and technology,
Beijing University of Posts and Telecommunications, Beijing
bugnec@gmail.com, xjwang@bupt.edu.cn, gnfang@gmail.com

Abstract. We present Cohesivity, a gold-standard based evaluation measure for lexical category acquisition. Cohesivity meets three criteria: informativeness, diversity and purity; and use two kinds of gold-standard categories for measuring the syntactic and semantic similarity. We compare Cohesivity to a number of popular cluster evaluation measures commonly used in previous work and demonstrate that it satisfies several desirable properties of category acquisition, using simulated clustering results. Finally, we use Cohesivity to evaluate a real lexical category acquisition task.

Keywords: evaluation measure, lexical category acquisition, informativeness, diversity, purity

1 Introduction

By their third year of life, children have achieved the remarkable feat of how words are combined to form complex sentences. A particularly hard case is the discovery of lexical categories (noun, verb, or more fine-grained set of intuitive word categories). A number of computational models of category learning have been developed, most of which regard the problem as one of grouping together words whose syntactic behavior is similar. Typically, the input for the model is taken from a corpus of child-directed speech, and these computational models have used distributional cues for category induction [1, 2]; these models confirm the ability of identifying categories, and show that distributional cues are informative for categorization.

A problem common to all existing models is the evaluation of the categories induced. A number of different approaches to evaluation have been proposed in the past [3]. Early work used an informal evaluation of manually comparing the clusters produced by the models with the authors' intuitive judgment of the lexical categories. A second form of evaluation is to use some data that has been manually or semi-automatically annotated with part of speech (POS) tags, and to use some information theoretic measure to assess the correlation between the 'correct' data and the induced categories [3]. This type of evaluation is not ideal for assessing the categories, as the measure used was designed for clustering not for category acquisition and these may include linguistically valid distinctions not recognized by the gold standard [4]. A third evaluation is to use the derived classification in a class-based language tasks, and to measure the performance of the language tasks [3]. This evaluation is often very expensive, such as evaluating word prediction, inferring the semantic properties of novel words, and grammaticality judgment, for example [5].

It is therefore desirable to develop an evaluation measure that makes reference to gold standard that is suitable for lexical category acquisition. On the other hand, good results of the categories induced should be informative about the properties of their members, rich diversity about the quantity of clusters, and pure about similarity among their members. Finally, the ideal measure needs to be different from the evaluation measures for clustering such as F-measure, V-measure etc.

This paper proposes a new evaluation measure, Cohesivity, which meets these criteria: informativeness, diversity and purity. It relies on a basic idea from first language acquisition, viz. children form categories in natural languages, a process of grouping together items used in linguistic communication. It require a gold standard which can be fine-grained than POS tags (e.g. semantic knowledge), and therefore is suitable for evaluating deep into the categories.

After discussing some preliminary issues in section 2, in section 3 we describe several popular evaluation measures. In Section 4, we describe Cohesivity and how it is calculated in terms of POS and semantic lexicon. We describe several popular evaluation measures and draw some comparisons to Cohesivity and discuss how some desirable properties for lexical category acquisition are satisfied by Cohesivity vs. other measures in Section 5. In Section 6, we present an example application of Cohesivity on discovering lexical categories in child-directed speech.

2 Lexical Category Acquisition

Children form categories by grouping together those words that occur in the same environments [6]. There are several classes of category acquisition theories based on the use of distributional regularities, semantics, and phonology of category acquisition. Despite the potential importance of semantic and phonological cues, everyone seems to agree that grammatical categories must ultimately be defined in grammatical terms.

Lexical categories are defined in terms of the syntactic and semantic environments in which they can occur. Several computational models have used distributional information for categorizing words [1, 2, 7, and 8]. The majority of these models partition the vocabulary into a set of optimum clusters (e.g., [7] and [8]). Chrupala and Alishahi [5] propose an incremental entropy model for efficiently clustering words into categories given their local context.

3 Existing Evaluation Methods

These strategies proposed above exploits knowledge about the syntactic and semantic information of language. The knowledge implied in lexical categories can be summarized in two points. First, that words belong to discrete categories such that substituting words within the same category keeps syntactic and semantic correctness. Second, that word types may belong to more than one category. There is no standard and straightforward method for evaluating the unsupervised models of category learning [3]. A number of different approaches to evaluation have been proposed in the past.

3.1 Without Using a Gold Standard

Automatically induced categories can be evaluated based on how useful they are in performing different tasks. Clark [8] takes an approach, where the perplexity of a finite-state model is used to compare different category sets.

Frank et al. [4] proposes an evaluation measure without a gold standard, which meets two criteria: substitutable precision and recall. At the same time, it yields results that correlate with gold-standard-based measures.

Chrupala and Alishahi [5] confirm that automatically induced categories can be evaluated based on how useful they are in performing different tasks. They propose an evaluation approach by comparing the efficiency of induced categories against other category sets (including POS tags) in a variety of language tasks. They emphasize that the ultimate goal of a category induction model is to form categories that can be efficiently used in a variety of language tasks for which they compare the performance based on various category sets.

3.2 Gold-standard-based Measures

The approach embed in a variety of application to measure the total performance of the application should be the correct way to evaluate the performance of a lexical category acquisition model. Such end-to-end evaluation, also called in vivo evaluation, is the only way to know if a particular improvement in a component is really going to help the task at hand.

Unfortunately, end-to-end evaluation is often very expensive. Thus often researchers have tested the output of their models against gold-standard category assignments, such as that available in the CHILDES database [9]. These gold-standard categories are based on the intuition of human annotators and are representative of adult knowledge. Therefore, for example, at the age of two, English-learning children have not fully acquired the verb category [10], and functional categories such as determiners are acquired even later [11].

Many unsupervised models of lexical category acquisition treat the traditional part of speech (POS) tags as the gold standard, and measure the accuracy of the categories based on how closely they resemble the POS categories (e.g. [1, 12, and 13]).

Mintz et al. [2] create one measure for the quantitative evaluation of analysis which called Purity. Purity is calculated for each linguistic category of interest and ranges from 0 to 1. Purity carries information about how well these clusters successfully group together words of the same linguistic category, but we have no idea about informativeness of all categories.

Matched accuracy (MA), also called many-to-one accuracy is widely used in the field of Natural Language Processing in part of speech tagging, in which the tokens of a text are automatically annotated (“tagged”) with cluster numbers. Given these labels, accuracy can be measured as usual, as the percentage of tokens correctly labeled. Multiple clusters may have the same label if several clusters match the same gold standard category. This can lead to a degenerate solution if the model is allowed an unbounded number of categories, in which each word is in a separate cluster.

Another frequently used evaluation measure is the information retrieval metric of F-Measure [14] [15]. Rosenberg and Hirschberg [16] point out that F-measure assumes (the missing) mapping between clusters and classes. Also, in their experimental assessment they show that when the number of clusters not representing a particular class was increased the F-measure did not decrease.

Variation of Information (VI) [17] is a clustering evaluation measure that measures the amount of information lost and gained when moving between two clusterings. A lower score implies closer clusterings, since each clustering has less information not shared with the other: two identical clusterings have a VI of zero. However, VI's upper bound is dependent on the maximum number of clusters, making it difficult to compare clustering results with different numbers of clusters [18].

Rosenberg and Hirschberg suggest another information-theoretic metric for clustering evaluation: V-measure (VM). VM is the harmonic mean of homogeneity and completeness which evaluate the quality of the clustering in a complementary way. Like VI, VM uses the conditional entropy of clusters and categories to evaluate clusterings. However, it also has the useful characteristic of being analogous to the precision and recall measures.

Evaluation of lexical category acquisition is known to be problematic, due to the fact that the categories induced by the model are unlabeled, and do not exactly correspond to any of the gold standard part of speech categories. In fact, many language tasks might benefit from finer-grained categories than the traditional POS tags used for corpus annotation, e.g., sets of words denoting vehicles, types of food, tool names, etc..

All in all, using any set of gold standard categories for evaluating a lexical categorization model has the disadvantage of favoring one set of principles and intuitions over another; that is, assuming that there is a correct set of categories which the model should converge to. Besides POS tags, finer-grained categories than the traditional POS tags (e.g. thesaurus) should be treated as gold standard.

Thus we would like a metric that can be used to quickly evaluate potential improvements in a lexical category acquisition model. Thus it can be commonly used as a quick check on an algorithm; an improvement in this metric can then be confirmed by an end-to-end evaluation.

4 Cohesivity and Its Calculation

Cohesivity is a gold-standard-based measure which explicitly measures how successfully the criteria of informativeness, diversity and purity have been satisfied. Our goal is to categorize word usages based on the similarity their surrounding words. To discuss lexical category acquisition evaluation measures we introduce these criteria for a categorization solution.

First, the categories should be informative about the properties of their members. Assuming that there is a correct set of categories which the model should converge to, a clustering result satisfies this criterion if all the words that are members of the correct set are elements of the same category induced. Assigning every word into a single category guarantees perfect purity; however, this lead to a degenerate solution.

It is meaningless for the lexical category acquisition task if the model is allowed an unbounded number of categories, in which few words are in a separate cluster.

Second, the categories should be rich diversity about the quantity of clusters. The case mentioned above, only one word in each category, violates the original intention of exploiting distributional cue to categorize words. Those clusters in which there is only one word will be removed from the categories induced. So a clustering result satisfies diversity if the model acquires more categories that are informative.

Third, the categories should be pure about similarity among their members, which including syntactic and semantic similarity. As described in section 3, many models of lexical category acquisition treat the traditional part of speech (POS) tags as the gold standard, and measure the accuracy and completeness of their induced categories based on how closely they resemble the POS categories; However, POS only shows the syntactic similarity of words, besides POS tags, manually or semi-automatically annotated semantic lexicon (e.g. thesaurus) should be treated as a gold standard, which is a representation of the semantic similarity of all the word categories.

An appropriate evaluation created for formalizing all the three criteria and quantitative evaluation of analysis is called Cohesivity. Cohesivity is calculated for all inducted categories; intuitively, given two models and the same corpus, the better model is the one that can induce richer set of categories and there are more members in each category as well as members are more similar to each other.

In this paper, for Chinese lexical category acquisition task, the Peking University POS tagset [19] and TongYiCi CiLin [20], a Chinese semantic thesaurus, are used. TongYi Ci CiLin (TC) has gathered about 52,206 Chinese lexicons and classified them in a 3-level hierarchy. Level 1 (Classes), consists of 12 classes, level 2 (Sections), consists of 94 classes, and level 3 (Category), consists of 1428 classes. In our experiment, we have examined results up to the second level.

The formula for computing Cohesivity is given as follows. Suppose there are K categories induced, for each found category C_i containing N_i words, the POS tags has L labels, the TC tags has M labels, let R_{Pj} and R_{Tk} denote the number of words present in both C_i and gold standard label j or k respectively:

$$P_{Pi} = \max\left(\frac{R_{Pj}}{N_i}\right), j = 1,...,L \tag{1}$$

$$P_{Ti} = \max\left(\frac{R_{Tk}}{N_i}\right), k = 1,...,M$$

Purity of one category is computed as the harmonic mean of distinct similarity of syntactic and semantic scores, just as precision and recall are commonly combined into F-measure [14]. As F-measure scores can be weighted, Cohesivity can be weighted to favor the contributions of syntactic and semantic similarity.

$$Pu_i = \frac{(1+\beta) P_P \cdot P_T}{\beta \cdot P_P + P_T} \tag{2}$$

Similarly to the familiar F-measure, if β is greater than 1 syntactic similarity is weighted more strongly in the calculation, and vice versa.

Pu_i is the purity of category C_i ,then

$$Cohesivity = \sum_i \frac{N_i \log N_i}{Z} Pu_i \tag{3}$$

$$Z = \log \overline{N}\ \overline{N} = \frac{\sum N_i}{K}$$

To calculate Cohesivity for the models, we needed to tag the gold standard label for each category. For the words with multiple labels, we calculate them for each corresponding label, and find the maximum value; and out of vocabulary (OOV) words will just be considered to belong in no labels or manually tag its label.

For our specific goals, Cohesivity is more suitable than existing methods because we could consider different criteria comprehensively, also Cohesivity is more efficient computationally and convenient than end-to-end evaluation.

5 Comparing Evaluation Measures

All the approaches mentioned in section 3.2 represent plausible ways to evaluate performance of their induced categories based on how closely they resemble the POS categories like traditional clustering problems. However, the lexical category acquisition task is related to, but distinct from, clustering task. Our semantic similarity criterion is not measured at all in these evaluation methods, such as F-measure, V-measure and VI, that previous authors have commonly used. That is, they do not address the question of whether finer-grained categories are included in a single category; in fact, when categorizing word usages based on the similarity of their content and their surrounding context, not POS but finer-grained categories e.g., sets of words denoting vehicles, types of food, tool names, etc, are induced in one cluster.

V-measure evaluates completeness, by examining the distribution of cluster assignments within each class. In Cohesivity we use a combination criterion of informativeness and purity rather than completeness. In lexical category acquisition task, our goal is to assign words of a fine-grained class (e.g. MOTION class for the verbs “walk” and “run”) into a category. The purity is likely to improve monotonically, while informativeness is likely to decrease in opposition with the number of categories decrease. A category satisfies completeness if all the words that are members of a finer-grained category are elements of the same cluster induced, and then purity and informativeness both achieve high scores.

Assigning each word to a distinct cluster guarantees perfect purity. However, this solution scores very poorly for lexical category acquisition, in fact we acquire no category at all. We believe it is a degenerate care where there is any category which has only one word. The degenerate solutions is far from ideal, so those clusters in which there is only one word will be removed from the categories induced. After that we will exam the diversity criterion. The categories should be rich diversity about the

quantity of clusters, indicating how successful the model categorizing the words. Assuming that in the degenerate case when there is only a single category, it is a poor solution for category acquisition and the solution likely to improve when more categories are acquired by dividing the category into subsets in right way; in another degenerate case when there are as many clusters as words, it is also a poor solution for category acquisition since there is not any category and the solution likely to improve when more categories are acquired by merging the similar clusters into superset.

Solution A
F-Measure=0.6
V-Measure=0.14
VI=2.74
Cohesivity=11.3

Solution B
F-Measure=0.6
V-Measure=0.39
VI=1.94
Cohesivity=11.3

Solution C
F-Measure=0.5
V-Measure=0.30
VI=2.82
Cohesivity=16.7

Solution D
F-Measure=0.5
V-Measure=0.41
VI=2.53
Cohesivity=21.4

Solution E
F-Measure=1
V-Measure=1
VI=0
Cohesivity=4

Solution F
F-Measure=1
V-Measure=1
VI=0
Cohesivity=6

Solution G
F-Measure=1
V-Measure=1
VI=0
Cohesivity=8

Solution H
F-Measure=1
V-Measure=1
VI=0
Cohesivity=12

Fig. 1. Examples of the categories

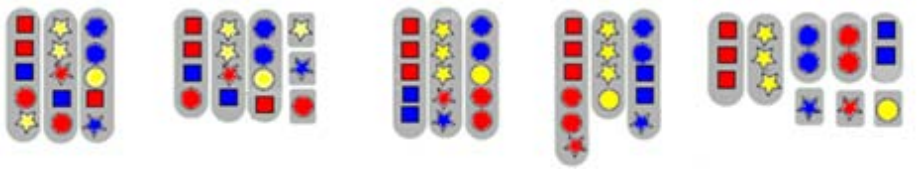
We believe that Cohesivity provides two significant advantages over other gold-standard based metrics that make it a more useful tool. First, other metrics do not explicitly calculate the degree of semantic similarity of the lexical categories. Second, other metrics do not explicitly address the interaction for assessing our three competing criteria of informativeness, diversity and purity.

We present some illustrative examples inspired by Rosenberg and Hirschberg [16]. For the purposes of this comparison we will be using F-measure, VM, VI and Cohesivity as the measures to describe the performance of lexical categories (Fig. 1 shows the examples). In the figures, the shaded regions represent lexical categories, the shapes represent POS tags, and the colors represent TC tags. In a perfect clustering, each shaded region would contain all and only the same shapes and colors.

Existing evaluation methods only treat the POS tags as the gold standard, so in the first analysis, we set the color to be the same. The situations are presented in figure 2. The F-Measure of solution A and B is 0.6 and Cohesivity is 11.3. The V-Measure of solution B (0.39) is greater than that of solution A (0.14), and the VI of solution B (1.94) is better than that of solution A (2.74) (lower VI is better). Solution B is a better clustering solution than solution A, in terms of both homogeneity and completeness. F-Measure and Cohesivity do not reflect this difference. Solutions C and D represent a case, in which clusters are divided into more pure subset; in real

application, clusters in which there is only one word will be removed, however, here we consider the small clusters have few but more than one word. In this example, the F-Measure of both solutions is 0.5. Solution D is a better clustering than solution C, there are no incorrect clusterings of different POS tag in the small clusters, which means high score of purity and diversity, though low score of the informativeness. The remaining solutions E to H represent the performance of the categories in the evaluation of the four metric. In lexical categories acquisition task, solution G and H are better than solution E and F, in terms of both informativeness (crudely, "each cluster contains more members") and diversity ("contains more clusters"). The two criteria are not measured at all in the calculation of F-Measure, V-Measure and VI.

In the second analysis, we show how Cohesivity works by computing the weighted harmonic mean of syntactic and semantic similarity in Fig. 2. The results from the analysis appear in figure 2 (β =1). Solution E is the best one, any other solution of the same elements will get less Cohesivity score than solution E. Compared to solution E, there are no incorrect clusterings of different POS tag in solution C, and there are no incorrect clusterings of different TC tag in solution D. And solution B is a better clustering than solution C and D, which gets higher score of purity and diversity, though low score of the informativeness.



Solution A	Solution B	Solution C	Solution D	Solution E
8.4	11.8	10.35	10.26	17.1

Fig. 2. Examples of Cohesivity performance

From these analyses, we have shown that, though it is difficult to assessing our competing criteria, Cohesivity provides an elegant solution to lexical category acquisition task that is more suitable than previously defined cluster evaluation measures. Lexical category acquisition task requires a gold standard which can be more fine-grained than POS tags and Cohesivity uses two kinds of gold-standard categories for evaluating deeply into the categories. We believe that Cohesivity addresses some of the problems that affect other cluster measures, and by evaluating the criteria of informativeness, diversity and purity, Cohesivity is more comprehensive than those that evaluate only one.

6 Applying Cohesivity

In this section, we describe a simple lexical category acquisition experiment and evaluate its results using Cohesivity, highlighting the interaction among informativeness, diversity and purity. We present results from the experiment in order to show how Cohesivity can be used to drawn comparisons on the acquisition task.

6.1 Dirichlet Process Mixture Model

Bayesian non-parametric models have received a lot of attention in the machine learning community. These models have the attractive property that the number of components used to model the data is not fixed in advance but is actually determined by the model and the data [23].

Some clustering algorithms applied to lexical category acquisition task so far require the number of categories as input. This is problematic as we do not know how many categories exist in the data. The fact that Dirichlet Process Mixture Models (DPMMs) do not require the number of target clusters in advance, renders them particularly promising for the many NLP tasks where clustering is used for learning purposes [21, 22, and 23]. With DPMMs, as with other Bayesian nonparametric models, the number of mixture components is not fixed in advance, but is determined by the model and the data. The parameters of each component are generated by a Dirichlet Process (DP) which can be seen as a distribution over the parameters of other distributions. In turn, each instance is generated by the chosen component given the parameters defined in the previous step:

$$G \mid \alpha, G_0 \sim DP(\alpha, G_0) \tag{4}$$

$$\theta_i \mid G \sim G$$

$$x_i \mid \theta_i \sim F(\theta_i)$$

Where G_0 and G are probability distributions over the component parameters (θ), and $\alpha > 0$ is the concentration parameter which determines the variance of the Dirichlet process. Instance x_i is generated by distribution F, parameterized by θ_i.

There are several methods to find approximate solutions of DPMMs, standard algorithms based on MCMC, such as those described by Neal [21], are computationally expensive and can take a long time to converge to the stationary distribution. Variational techniques [22] are an attractive alternative, but are difficult to implement and can remain slow. The development of inference algorithms for Dirichlet process models is currently an active research area and many improvements have been recently suggested in [24].

Daumé [24] have presented an algorithm for finding the MAP clustering for data under a Dirichlet Process mixture model. It has been shown to be extremely efficient and general by the author. We use this fast search technique to find an approximate MAP cluster assignment for the DPMMs.

6.2 Experiment

We use the Zhou corpus from CHILDES database [9] as experimental data. This Mandarin corpus is collected by Jin Zhou in East China Normal University. It has 50 files. Each file includes transcripts from a dialog tape between a mother and her child.

30 mother-child pairs with 10 children in each group were selected at 14, 26 and 32 months, and 20 pairs at 20 months. We use the mother's speech from transcripts.

In Mintz[2]'s analyses,, the features for all words are their immediate lexical contexts and associated frequencies in corpus data, which capture the context in which the words occur in text. Only the contexts of the 200 most frequent words are recorded; Mintz suggested the 200 most frequent words account for over 80% of the tokens, and less frequent words have very low frequencies. In fact, we believe that the 200 most frequent words contain insufficient distributional information for inducing categories; only that extremely less frequent words really need not to be considered, since DPMMs do not weigh the features, a large number of sparse features is likely to influence inappropriately the clustering discovered.

Table 1. Lexical category acquisition affected by feature dimension. See text for detailed explanations.

Feature	Number of categories	Avg size of categories	Matched accuracy		Cohesivity(beta=1)
			POS	TC	
100	37	19.7	60.1%	37.8%	261.3
200	52	14.6	66.6%	41.0%	300.1
300	49	15.9	64.0%	35.6%	326.4
400	51	15.3	67.0%	35.6%	327.4
500	61	13.0	67.5%	40.8%	342.4
600	73	11.0	62.9%	39.2%	332.1
700	80	10.1	63.9%	41.1%	323.9
800	75	10.7	62.7%	40.3%	321.5

We compared different results with varying dimensions of features. Table 1 shows performance of our lexical categories varies respond to the feature dimension. We evaluated the results using Cohesivity. Let us trace out in greater detail our experimental results. As feature dimension increases, the number of lexical categories increases as well as the average accuracy based on both the POS categories and semantic categories. When feature dimension is set to a low value, not enough information implied in the context is acquired, so lots of words that are not related are assigned into same category; in contrast, if we set it to be a large value, DPMMs will suffer from sparse features, and lots of small categories are formed, though it get high purity score, the Cohesivity value is low.

The average measure based on POS is better than the one based on TC tag, suggesting that though the lexical category acquisition perform well in acquiring lexical category in a POS level, which means the distributional information about word surrounding context is a useful source of information for children to form categories of various kinds of linguistic constituents (e.g., nouns and verbs), more information should be involved in inferring much finer-grained categories such as the semantic properties of an unknown word based on its context.

Considering the above properties of induced categories integrated, Cohesivity is a quick check, and can be used to quickly evaluate potential improvements in a lexical category acquisition model. It is also straightforward to note that the models do better as feature dimension increases, reach the peak about 500 and then began to decline.

Given the naive approaches taken in the experiment, this is expected Cohesivity has been notably successfully applied in acquisition task by meeting the three criteria. This example allows us to observe how transparently and conveniently Cohesivity can be used to compare the behavior across distinct lexical categories.

7 Conclusions

In lexical category acquisition task, the gold-standard categories are formed according to "substitutability": if one word can be replaced by another and the resulting sentence is still grammatical and semantically correct, then there is a good chance that the two words belong to the same category.

Thus automatically induced categories should be evaluated based on how useful they are in performing different tasks. However, end-to-end evaluation is often very expensive. It is therefore highly desirable to develop an evaluation measure that makes reference to gold standard; the ideal measure needs to be applicable to a wide range of different acquisition models.

We have presented a new external evaluation measure, Cohesivity, and compared it with existing commonly used evaluation measures. Cohesivity is based upon three criteria for clustering usefulness, informativeness, diversity and purity. We have also demonstrated Cohesivity's usefulness in comparing clustering success through a simple experiment by evaluating a DPMMs model for lexical category acquisiton.

We believe that Cohesivity addresses some of the problems that affect other cluster measures. Lexical category acquisition task is related to, but distinct from, clustering task. It require a gold standard which can be fine-grained than POS tags (e.g. semantic knowledge), and Cohesivity use two kinds of gold-standard categories for evaluating deep into the categories. The acquisition task is not completeness, and those categories in which there is only one word will be removed; acquisition models which get high Cohesivity scores will avoid this situation. By evaluating the criteria of informativeness, diversity and purity, Cohesivity is more comprehensive than those that evaluate only one.

Acknowledgments. This research has been supported by NSFC90920006 and RFDP 20090005110005.

References

1. Redington, M., Crater, N., Finch, S.: Distributional information: A powerful cue for acquiring syntactic categories. Cognitive Science: A Multidisciplinary Journal, 22(4), 425–469 (1998)
2. Mintz, T., Newport, E., & Bever, T.: The distributional structure of grammatical categories in speech to young children. Cognitive Science, 26, 393–425 (2002)
3. Clark, A.: Combining distributional and morphological information for part of speech induction. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (pp. 59–66) (2003)

4. Frank, S., Goldwater, S., Keller, F.: Evaluating models of syntactic category acquisition without using a gold standard. In Proceedings of the 31st Annual Meeting of the Cognitive Science Society (2009)
5. Chrupala, G., Alishahi, A.: Online Entropy-based Model of Lexical Category Acquisition. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning (2010)
6. Cartwright, T., Brent, M.: Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. Cognition 63, pp. 121-170 (1997)
7. Brown, P., Mercer, R., Della Pietra, V., Lai, J.: Class-based n-gram models of natural language. Computational linguistics, 18(4), 467–479 (1992)
8. Clark, A.: Inducing syntactic categories by context distribution clustering. In Proceedings of the 2nd workshop on Learning Language in Logic and the 4th conference on Computational Natural Language Learning (pp. 91–94) (2000)
9. MacWhinney, B.: The CHILDES project: Tools for analyzing talk. Lawrence Erlbaum Associates Inc, US (2000)
10. Olguin, R., Tomasello, M.: Twenty-five-month-old children do not have a grammatical category of verb. Cognitive Development, 8(3), 245–272 (1993)
11. Kemp, N., Lieven, E., Tomasello, M.: Young children's knowledge of the determiner and adjective categories. Journal of Speech, Language, and Hearing Research, 48(3) (2005)
12. Mintz, T.: Frequent frames as a cue for grammatical categories in child directed speech. Cognition, 90(1), 91–117 (2003)
13. Parisien, C., Fazly, A., & Stevenson, S.: An incremental bayesian model for learning syntactic categories. In Proceedings of the Twelfth Conference on Computational Natural Language Learning (2008)
14. Rijsbergen, V.: Information Retrieval, 2nd edition. Dept. of Computer Science, University of Glasgow (1979)
15. Fung, B., Wang, K., Ester. M.: Hierarchical document clustering using frequent itemsets. In Proc. of the SIAM International Conference on Data Mining (2003)
16. Rosenberg, A., Hirschberg, J.: Vmeasure: A conditional entropy-based external cluster evaluation measure. In EMNLP (2007)
17. Meila, M.: Comparing clusterings – an information based distance. Journal of Multivariate Analysis, 98:873–895 (2007)
18. Frank, S., Goldwater, S., Keller F.: Using Sentence Type Information for Syntactic Category Acquisition. Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics, ACL 2010, pages 1–8 (2010)
19. Yu, S., Duan, H., Zhu, S., Swen, B., Chang, B.: Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation. Journal of Chinese Language and Computing, 13, 2,121-158 (2003)
20. Mei J., Zhu Y., Gao Y., Yin H.: TongYiCi CiLin, ShangHai DianShu ChuBanShe (1983)
21. Neal, R.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics, 9, 249–265 (2000)
22. Blei, D., Michael J.: Variational inference for Dirichlet process mixtures. Bayesian Analysis, 1(1):121{144, August (2005)
23. Vlachos, A., Ghahramani, Z., Korhonen, A.: Dirichlet process mixture models for verb clustering. In Proceedings of the ICML workshop on Prior Knowledge for Text and Language (2008)
24. Daumé III, H.: Fast search for dirichlet process mixture models. In: Meila, M., Shen, X. (eds.) Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 83–90 (2007)

Ontological-, linguistic- and interface-issues

Application of Classical Psychological Theory to Terminological Ontology Alignment

Fumiko Kano Glückstad

Copenhagen Business School
Dept. of International Language Studies and Computational Linguistics
Dalgas Have 15, DK-2000, Frederiksberg, Denmark
fkg.isv@cbs.dk

Abstract. Terminological Ontology [1] is a method used for knowledge sharing and domain-specific translation practices, and could potentially be suitable to apply and for simulating the cognitive theories/models explaining real-world inter-cultural communication scenarios. In this paper, we investigate - as a preliminary study - whether Tversky´s contrast model [2] is applicable to data-sets obtained from the Terminological Ontology method. The eventual purpose of this study is to propose an approach for identifying potential translation candidates by optimizing relevance between concepts in two remote languages such as European and Asian languages.

Keywords: Terminology, Ontology, Translation, Inter-Cultural Communication, Set-theory, Similarity, Generalization, Relevance

1 Introduction

The role of ontology in a multilingual setting is considered an emerging challenge to Semantic Web development. As a consequence, there are several major ongoing projects, such as the MONNET project on Multilingual Ontologies for Networked Knowledge [3] and the KYOTO project on Knowledge-Yielding Ontologies for Transition-Based Organization [4]. Though both projects deal with translation of terms from a Source Language (SL) to a Target Language (TL), they focus on linking lexical data through an interoperable common ontology rather than on optimizing relevance between concepts that are potentially measurable based on diverse models derived from cognitive theory.

Terminological Ontology (TO) is a domain-specific ontology used for knowledge sharing [1], which normally is applied in terminology work, cf. for example [5]. The unique characteristics of TO that differentiate it from other types of ontologies are feature specifications and subdivision criteria [6]. A feature specification consists of a feature dimension and its value. Hence, a representation of a whole concept is a feature structure, i.e. a set of feature specifications corresponding to the unique set of characteristics that constitutes that particular concept [1][6]. Terminological ontologists argue that concepts are defined in a language-dependent context, and therefore TO is language-dependent. TO is developed within a knowledge sharing

community, then dynamically updated and validated. If it is necessary to share knowledge with other communities, TOs developed in different communities should be compared, aligned and merged upon necessity. While the aforementioned two mainstream projects, MONNET and KYOTO, both deal with complex ontologies involving huge data-sets, TO usually handles smaller amounts of concepts.

Considering this, a point that should be emphasized in this work is that TO could potentially be a suitable tool to apply and for simulating cognitive theories/models explaining a real-world inter-cultural communication scenario. Thus, in the next chapter a problem existing in the real-world is explained by use of the Relevance Theory of Communication [7]. Next, the concept of similarity, which has a long history in psychological theory, is reviewed in Chapter 3. We consider how the theory proposed by Tversky [2] is applied in the context of the Relevance Theory of Communication. In Chapter 4, the empirical analysis is performed to assess the potential of applying the models based on [2] to TOs. Chapter 5 discusses findings and future work followed by conclusions in Chapter 6.

2 Real-World Problem

Imagine a situation where a non-native English speaking European and an Asian are debating in English about the issue of academic degree systems in their respective cultures. While a German might be explaining about the Doctor of Science (Habilitation) degree (the highest achievable academic degree in Germany after obtaining a Doctor of Philosophy degree), a Japanese might be having the highest possible academic degree in Japan in his mind which is a Doctor of Philosophy degree (also frequently referred to as Doctor of Science in Japan). This imagined conversation shows a typical scenery revealing a deep inherent misconception between the two communicating parties since each of them have their own conceptual - and correct - understanding of the highest obtainable academic degree in their respective cultures.

This example may further create problems for a translator who is going to translate academic titles into the language of the other party. When a translator translates the term of the German Doctor of Science Degree into Japanese, the first condition he/she has to fulfill is that his/her translation should convey the same meaning as the original German meaning. Gutt [8] explains that this requires *the receptors to familiarize themselves with the context envisaged for the original text.* Now the question is, when a Japanese receptor is not familiar with the German language and its academic culture, how should this particular German academic title then be translated into Japanese?

The proposal [8] of applying the Relevance Theory of Communication [7] is a key to address this issue. This theory focuses on how people share thoughts with one another and views communication as principally an inferential process. It means that the essential task of the communicator is to produce a *stimulus* from which an audience can infer what set of thoughts or assumptions the communicator intends to convey [8]. Hence the second condition the translator has to fulfill is that his/her translation should explicitly provide a set of assumptions that are adequately relevant to the audience. The issue here is how the translator should create the *stimulus* (that is *translation*) optimally relevant to the audience. Assuming that both German and Japanese have

their respective conceptual structures of the academic system rooted in their own cultures, translation candidates that have optimally relevant relationships identified from these two conceptual systems could avoid the gratuitous inferential processing effort on the audience´s part.

The optimization of the relevance between two concepts could be well explained by the cognitive theory, Tversky´s Set-theoretic model of similarity [2]. Thus the next Chapter reviews Tversky´s model and considers how this model could be used in the context of optimizing the relevance of communication.

3 Tversky´s Set-theoretic Model of Similarity

The concept of similarity has a long history within the area of psychological theory. Tversky´s view of similarity [2] is distinguished from the traditional theoretical analysis (c.f. [9]) on two points: 1) while the theoretical analysis of similarity relations has been dominated by the continuous metric space models, [2] argues that *the assessment of similarity between objects may be better described as a comparison of features rather than as the computation of metric distance between points;* and 2) although *similarity has been viewed by both philosophers and psychologists as a prime example of a symmetric relation*, the asymmetric similarity relation has been demonstrated in [2] based on several empirical evidences.

Based on these two points, [2] proposed a classic feature-set model of similarity as follows:

$$sim(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha \cdot f(A - B) + \beta \cdot f(B - A)} \quad (1)$$

Here, A and B are the feature sets of object *a* and object *b*. f denotes a measure over the feature sets. (A⋂B) represents the sets of features present in both A and B, (A-B) represents the sets of features present in A but not in B, and (B-A) represents the sets of features present in B but not in A. α and β are free parameters representing an asymmetric relationship between A and B. Since the similarity score in this equation is normalized, the obtained score lies between 0 and 1.

An interesting point is that the application of Tversky´s model requires *a limited list of relevant features and the representation of an object as a collection of features that is viewed as a product of a prior process of extraction and compilation* [2]. In fact, the principle of TO in a way follows rigid rules of categorization. This can systematically extract the collection of features based on the subdividing dimensions. Therefore, the hypothesis is that Tversky´s model is applicable to data-sets extracted from the terminological ontologies. Another important point in the context of the Relevance Theory of Communication is that translation should provide the set of assumptions that are adequately relevant to the audience, and the stimulus (that is translation) produced by the translator is such that it avoids gratuitous inferential processing effort on the audience´s part. Considering that *similarity serves as an organizing principle by which individuals classify objects, form concepts, and make generalizations* [2], the most

similar concept to a SL concept, which is identified in the audience´s culture through the feature matching, could be the set of assumptions which are adequately relevant to the audience. Thus, the second hypothesis is that the optimization of the relevance required in an inter-cultural communication can be achieved by aligning the ontological graphs (conceptual hierarchies) and feature specifications which constitute concepts in the two language-dependent terminological ontologies. In order to assess these hypotheses, terminological ontologies are developed from corpora describing real-world concepts in the two remote cultures. The similarity score of the selected concepts are computed by applying Tversky´s model [2] based on the collection of features extracted from these ontologies. This is dealt with in the next chapter.

4 Feature Matching Based on Tversky´s Model

4.1 Corpora

Texts describing the Japanese educational system have been identified from the "Multilingual Living Information[1]" site provided by the Council of Local Authorities for International Relations and from a pamphlet entitled "Higher Education in Japan[2]" published by the Japanese Ministry of Education, Culture, Sports, Science and Technology. For the Danish educational system, documents that are downloaded from the Eurydice web-site[3] published by the Education, Audiovisual and Culture Executive Agency under the EU commission have been used as text corpus. All these documents are officially published in English by reliable authorities of each country. Thus, all English translated terms and expressions in their original languages are considered as official terms. It means that it is feasible to identify terminological expressions in an original language from documents published by the respective authorities. This enables one to eventually identify translation equivalences linking between, in this case, Danish and Japanese. In this study, only the English documents describing language-dependent concepts in the two cultures are used as text corpora.

The Eurydice publishes documents describing the educational systems in a majority of the EU member countries both in English and in their native languages. It means that the same method can, in principle, be applied to other language combinations.

4.2 Ontology Construction

The terms and their definitions describing the educational systems in each country are manually identified from the respective English documents. Based on these terms and their definitions, terminological ontologies representing the educational system in each

1 http://www.clair.or.jp/tagengorev/en/j/index.html

2 http://www.mext.go.jp/english/koutou/detail/1287370.htm

3 http://eacea.ec.europa.eu/education/eurydice/index_en.php

of the two countries are developed using a Computer Aided Ontology Structuring prototype (CAOS) that is based on the TO principles defined in [1]. As described in Chapter 1, the uniqueness of TO is feature specifications and subdivision criteria [6]. A feature specification is presented as attribute-value pair - for example in Figure 1, [ENTRANCE REQUIREMENT: high school graduate]. Thus, a representation of a whole concept is a feature structure, i.e. a set of feature specifications corresponding to the unique set of characteristics that constitutes that particular concept [1]. In Figure 1, each box that represents a particular concept is divided into three layers: 1) top layer, lexical representation (term), 2) middle layer, dimension specifications, and 3) bottom layer, feature structure (set of feature specifications).

The use of feature specifications is subject to principles and constraints described in detail in [1]. Most importantly, a concept automatically inherits all feature specifications of its superordinate concepts. Secondly, polyhierarchy is allowed so that one concept may be related to two or more superordinate concepts. On the other hand, subdivision criteria that have been used for many years in terminology work are strictly implemented in TO by introducing dimensions and dimension specifications [1][6]. This enables the CAOS prototype to perform consistency checking which helps in constructing ontologies [1]. A dimension of a concept is an attribute occurring in a non-inherited feature specification of one or more of its subordinate concepts [1][6]. Values of the dimension allow a distinction among sub-concepts of the concept in question. In Figure1, the concept “academic degree” has the dimension [LENGTH OF EDUCATION] whose values are [2-3 years | minimum 4 years]. These dimension values distinguish the sub-concepts: “junior college degree” and “university degree”. This clarification makes it much easier to identify subdivision criteria and differentiating characteristics [6]. The same feature attribute can only occur on sister concepts and a given value can only appear on one of these sister concepts. In this way a concept must be distinguished from each of its nearest superordinate concepts as well as from each of its sister concepts by at least one feature specification [1][6].

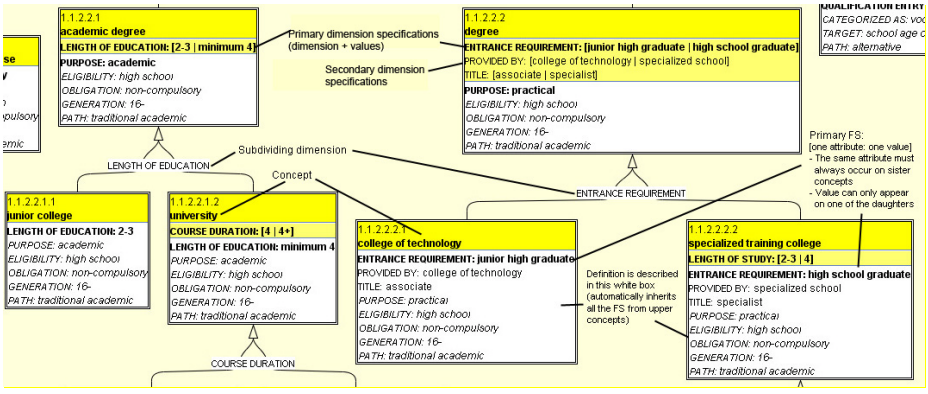


Fig. 1 Example of the Terminological Ontology.

By using the CAOS prototype that performs the consistency checking of the TO principles, the two educational ontologies are developed based on the terms and definitions manually extracted from the corpora.

4.3 Feature Matching Based on Tversky´s Contrast Model

The basic techniques generally used in the ontology matching are string-based (lexical) matching, graph-based (structural) matching, and feature-based matching [10]. Accordingly, string-based matching is manually performed as the first step. The Japanese and Danish educational system ontologies, respectively, contain 42 and 65 concepts consisting of terms (lexical representations) and their feature specifications. Among these, only two terms are completely matched. It indicates that graph-based matching based on the lexically matched nodes is not sufficient in this case. Therefore, feature-based matching is manually implemented in a top-down manner as the second step. The third highest dimension in the Japanese ontology and the highest dimension in the Danish ontology are "generation". Hence, the two ontologies are categorized into the following three blocks based on feature values of this generation dimension: "0-6 years old"; "6-15 years old"; and "16 years old and above."

Table 1. List of terms and feature sets.

Japanese concepts	
a:high school	A= {formal education, 16 years +, non-compulsory, lower secondary graduate}
b:general course	B= {formal education, 16 years +, non-compulsory, lower secondary graduate, general}
c:specialized training course	C= {formal education, 16 years +, non-compulsory, lower secondary graduate, specialized}
d:technical course	D= {formal education, 16 years +, non-compulsory, lower secondary graduate, specialized, technical}
e:business course	E= {formal education, 16 years +, non-compulsory, lower secondary graduate, specialized, business}
f: higher education	F= {formal education, 16 years +, non-compulsory, secondary graduate}
Danish concepts	
g:upper secondary education	G= {16 years +, lower secondary graduate}
h:general upper secondary education	H= {16 years +, lower secondary graduate, access to higher education}
i:gymnasium	I= {funded by state, 16 years +, lower secondary graduate, access to higher education}
j:business college	J= {self governing, 16 years +, lower secondary graduate, access to higher education}
k:HHX program	K= {self governing, 16 years +, lower secondary graduate, access to higher education, business}
l:HTX program	I= {self governing, 16 years +, lower secondary graduate, access to higher education, technical}
m:vocational or technical education	M={16 years +, lower secondary graduate, access to labor market}
n: tertiary education	N={secondary graduate, project and research}

In this study, the focus is on the block having feature value "16 years old and above." From this block, some of the sub-concepts and their feature values listed under the Japanese term, "non-compulsory education", and the Danish terms, "upper secondary education" and "tertiary education" are manually selected in Table 1 (due to the paper space, redundant data – e.g. concepts having similar constitution of feature sets – has intentionally been omitted). In order to apply Tversky´s model, synonymous feature expressions identified from the country specific corpora are approximately standardized by hand in Table 1. One thing to notice in Table 1 is that if a concept is categorized into several sub-concepts based on a dimension, an extra feature specification is added to each of them according to the principles of TO. Hence it is possible to observe the hierarchical structure from the feature values listed in Table 1.

Now the question is how to assign the asymmetric parameters in accordance to the translation direction. In [2], the direction of asymmetry is determined by the relative salience of the stimuli, in other words, the variant is more similar to the prototype than vice versa. Thus, *if sim(a,b) is interpreted as the degree to which a is similar to b, then a is the subject of the comparison and b is the referent. Hence the features of the subject are weighted more heavily than the features of the referent*. When considering a translation scenario, translators´ task is to identify a concept in audiences´ conceptual structure that is optimally relevant to the concept in the SL. It means that the stimulus selected by a translator should to the maximum extent be similar to a concept in the SL concept. Therefore, the features of a stimulus should be weighted more heavily than the ones of an SL concept in accordance to [2]. Hence, the asymmetric parameters are manually set as α=0.7 and β=0.3 in this empirical study. The result is shown in Table 2 and 3. In Table 2, the Danish concepts (*g-n*) are set as subject of the comparison and the Japanese (*a-f*) as referent. Opposite to this, the Japanese concepts (*a-f*) are set as subject of the comparison and the Danish (*g-n*) as referent in Table 3.

Table 2. Tversky´s similarity score: *a-f* (JP) as referent (SL), *g-n* (DK) as stimulus (TL).

SL	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*
a	**.77**	.61	.50	.50	.43	.43	.61	.00
b	**.69**	.56	.47	.47	.4	.4	.56	.00
c	**.69**	.56	.47	.47	.4	.4	.56	.00
d	**.63**	.51	.43	.43	.38	.57	.51	.00
e	**.63**	.51	.43	.43	.57	.38	.51	.00
f	**.38**	.30	.25	.25	.21	.21	.30	**.38**

Table 3. Tversky´s similarity score: *g-n* (DK) as referent (SL), *a-f* (JP) as stimulus (TL).

SL	*a*	*b*	*c*	*d*	*e*	*f*
g	**.58**	.49	.49	.41	.41	.29
h	**.54**	.45	.45	.39	.39	.27
i	**.50**	.43	.43	.37	.37	.25
j	**.50**	.43	.43	.37	.37	.25
k	.47	.4	.4	.35	**.53**	.23
l	.47	.4	.4	**.53**	.35	.23
m	**.54**	.45	.45	.39	.39	.27
n	.00	.00	.00	.00	.00	**.29**

Note1: Bold font is the highest score in each row

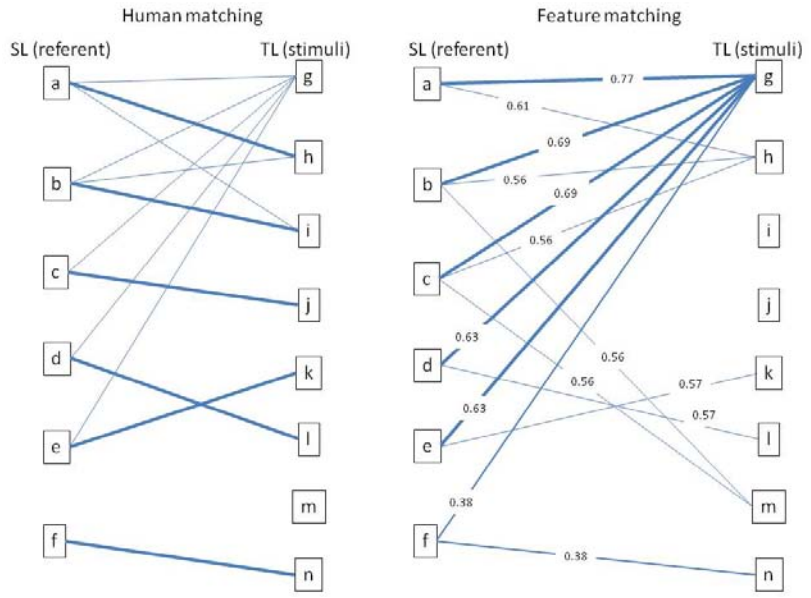


Fig. 2 Comparison with human matching: Danish terms as stimuli

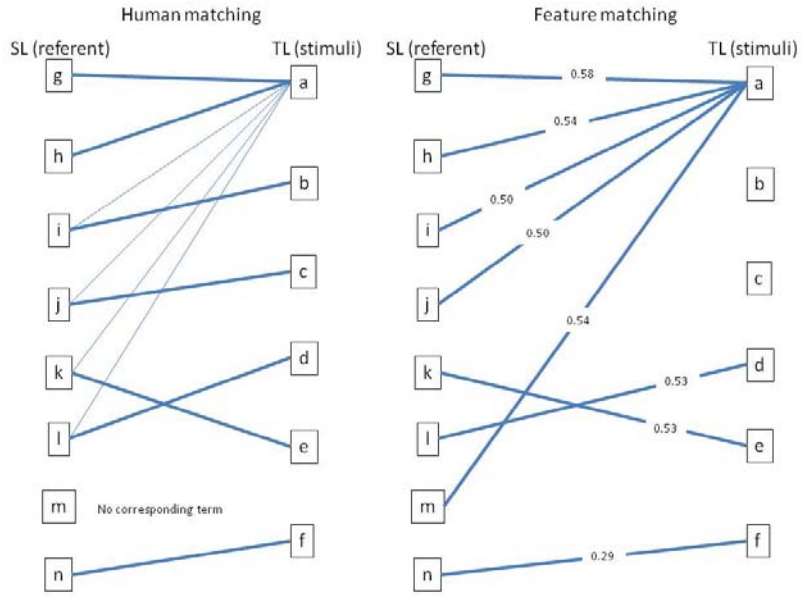


Fig. 3 Comparison with human matching: Japanese terms as stimuli

In Table 2 and 3, the scores with bold fonts are the highest scores in each row. From these tables, it can be interpreted that a concept with the highest score has the most optimal relevance to an SL concept.

The first notable point is that the application of the asymmetric parameters resulted in the asymmetric bidirectional relationships in most of the links between Danish and Japanese concepts. To be more precise, it is less optimal to use the Japanese term as stimulus, when conveying the original meaning of the Danish concept (e.g. the asymmetric score for term *c* as stimulus and *g* as referent: 0.49). On the other hand, it is more optimal to use the Danish term as stimulus, when conveying the original meaning of the Japanese concept (the asymmetric score for term *g* as stimulus and *c* as referent: 0.69).

Another point is that the majority of the identified optimal stimuli in Table 2 and 3 were the most general terms located at the highest hierarchy in the data-sets. From this viewpoint, it is difficult to assess whether the application of Tversky´s model to the terminological ontologies is successful. Hence the feature matching results are compared with the human matching results in Figure 2 and 3. In the human matching charts (the left side of the figures), the bold line indicates the ideal stimulus (optimal translation candidate) and the slim line indicates the acceptable stimuli (reasonably acceptable translation candidates) for each SL term. In the feature matching charts (the right side of the figures), the bold line indicates the optimal stimulus having the highest score for each SL term. The slim line indicates the stimuli that are not the highest score for a SL term, but having scores over 0.55.

5 Discussion

As described in Chapter 4, among all the concepts between the two ontologies, only two terms were completely matched in the string-based matching. This indicates that the English educational terminology used in respective knowledge sharing communities is immensely dissimilar, even though the educational concepts existing in the two countries are relatively similar. From this observation, it can be elaborated that it is very complicated to link concepts in two remote languages. This is because language resources having direct links between two remote languages are usually very limited, and therefore a pivot translation via English is often required both for dictionary-based human translations and for statistically-based machine translations. This also emphasizes the necessity for carefully analyzing how the meanings of a concept in one culture can be conveyed to a person in another culture through English as lingua franca.

From this viewpoint, the empirical analysis in Chapter 4 showed modest progress. The results listed in Table 2 and 3 as well as Figure 2 and 3 indicate that Tversky´s contrast model [2] is to a certain extent applicable to data-sets extracted from terminological ontologies. The application of the asymmetric parameters showed an interesting indication that it is less optimal to use the Japanese term, e.g. “c: high school-specialized course” as stimulus for a Japanese audience, when conveying the original meaning of the Danish concept e.g. “g: upper secondary education” (the asymmetric score is 0.49). On the other hand, it is more optimal to use the Danish term

e.g. “g: upper secondary education” as stimulus for a Danish audience, when conveying the original meaning of the Japanese concept e.g. “c: high school-specialized course” (the asymmetric score is 0.69). Even though concepts in two cultures are mapped to each other, it is not necessarily true that a translational equivalence holds in a bidirectional way, if the two concepts are not 100% identical. This result can be explained as follows: when considering equation (1) of Tversky´s model, it is obvious that if features for the parameter α (features that are in the feature set A but not in set B: (A-B)) increases, the similarity score severely decreases. The reason for the aforementioned result (the use of Japanese term as stimulus is less optimal) is that, when we categorize the two ontologies into three blocks, we select the “generation” dimension that is the third highest dimension in the Japanese educational system. Therefore, the feature values of the first- and second highest dimensions have been inherited to all the Japanese concepts listed in Table 1. This indicates that the asymmetric parameters of Tversky´s model, to a certain degree, reflect the hierarchical structure hidden behind the feature structures of the terminological ontologies.

A final point is that, in most cases, the identified optimal stimuli based on the similarity score in Table 2 and 3 is the most general term located at the highest hierarchy in the data-sets. According to the principle of TO, when a concept is subdivided into several sub-concepts based on a dimension, an extra feature is added to each sub-concept. Hence it is often the case that concepts having more features are more specific sub-concepts. It means that the lower a concept is located in the ontology, the more features the concept inherits from superordinate concepts. If dimensions and their values at the lower part of the two ontologies are not consistent, all the inherited features are simply acting as noise in the data-sets. The positive interpretation could be that Tversky´s model is applicable to identify corresponding pairs with less noise, in other words, pairs that optimally share common features with less noise. The negative interpretation could be that Tversky´s model has limitations in identifying corresponding pairs at the optimally specific level. Considering communication in the real world, it is not incorrect to say that the relevance required in the communication is achieved in this way, since people can usually achieve a mutual understanding much easier at a reasonably general level than at a very specific level. However, as Figure 2 and Figure 3 illustrate, the optimal translation candidates selected by human are the optimally specific concepts. Hence one of the challenges is to identify the reasonably specific terms from noisy data-sets. In order to achieve this, additional investigations (e.g. implementing the feature matching for the all concepts in the two ontologies) are required.

Another future challenge is to further investigate and compare this empirical study with data-sets obtained from terminological practices as well as from the translation practices in the real world. The data-sets used in this study are English documents published by the EU commission for the Danish educational system. The EU commission has used English terminology that is standardized based on the International Standard Classification of Education (ISCED) defined by the United Nations Educational, Scientific and Cultural Organization. Therefore, the documents describing the educational system for the majority of EU member countries are based on the standardized classification and English terminology. Hence, it may be much easier to align the educational system ontologies constructed from documents published by the EU member countries. On the other hand, the Japanese ontology does

not conform to the same standardized classification and terminology. By applying Tversky´s contrast model to the different data-set combinations, it may be possible to investigate the behavior of data in different scenarios. Extending from this viewpoint, the further development of the CAOS prototype for automating the knowledge extraction, ontology construction and update described in [11] could efficiently be synchronized in order to link language-specific concepts existing in different countries. To be more concrete, it might be effective to automatically extract knowledge from the domain-specific corpora based on pre-defined feature dimensions derived from a standardized classification and the terminology e.g. the ISCED classification.

Finally, Tversky´s contrast model has been extended by several researchers in different disciplines. Especially in the area of cognitive science, Tenebaum and Griffiths [12] proposed a framework that subsumes Tversky´s model of similarity by recasting Shepard´s universal law of generalization [10] in a more general Bayesian framework. Frank et al. [13] further extended this framework in [12] in order to model informative communication based on [7]. Hence, it is an obvious future challenge to apply these extended cognitive models to the aforementioned different combinations of data-sets.

6 Conclusion

In this paper, the applicability of Tversky´s contrast model derived from the cognitive theory [2] to data-sets extracted based on the Terminological Ontology method is investigated. The study indicates that the application of [2] to [1] could, to a certain extent, enable one to analyze not only the degree of relevance between concepts in two cultures, but also the degree of asymmetric relationship between the concepts. By extending [2] to e.g. [12][13], it may be feasible to investigate further how meanings of a concept in one culture can be effectively conveyed to a another culture through English as lingua franca. However, further investigations using data-sets obtained from terminological practices as well as from translation practices are needed in order to clarify the limitations pointed out in this study.

Acknowledgments. I would like to express my gratitude to my supervisors, Hanne Erdman Thomsen and Daniel Hardt for their valuable guidance and support on my Ph.D. project. I would also like to thank my colleagues in the psycholinguistic reading group at Copenhagen Business School, who provided me with a selection of highly relevant papers and constant inspiration in the area of intercultural communication and translation theories.

References

1. Madsen, B.N., Thomsen, H.E., Vikner, C.: Principles of a System for Terminological Concept Modelling. In: Proc. The 4th International Conference on Language Resources and Evaluation, ELRA, 15-19 (2004)
2. Tversky, A.: Features of Similarity. In: Psychological Review 84, 327-352 (1977)
3. Declerck, T., Krieger, H.U., Thomas, S.M., Buitelaar, P., O´Riain, S., Wunner, T., Maguet, G., McCrae, J., Spohr, D., Montiel-Ponsoda, E.: Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge Across Europe. In: Rooz, J., Ivanyos, J. Eds.: Internal Financial Control Assessment Applying Multilingual Ontology Framework, Budapest: HVG Press, 67-76 (2010)
4. Vossen, P., Agirre, E., Calzolari, N., Fellbaum, C., Hsieh, S., Huang, C., Isahara, H., Kanzaki, K., Marchetti, A., Monachini, M., Neri, F., Raffaelli, R., Rigau, G., Tescon, M.: KYOTO: A System for Mining, Structuring and Distributing Knowledge Across Languages and Cultures. In: Proc. The 6th International Conference on Language Resources and Evaluation, Morocco, 1462-1469 (2008)
5. ISO 704: Terminology Work – Principles and Methods. Genève: ISO (2000)
6. Madsen, B.N., Thomsen, H.E., Vikner, C.: Comparison of Principles Applying to Domain Specific Versus General Ontologies. In: Oltramari, A., Paggio, P., Gangemi, A., Pazienza, M.T., Calzolari, N., Pedersen, B.S., Simov, K. (eds.): OntoLex 2004: Ontologies and Lexical Ressources in Distributed Environments. ELRA, 90-95 (2004)
7. Sperber D., Wilson, D.: Relevance: Communication and Cognition. Blackwell, Oxford (1986)
8. Guut, E.A.: A Theoretical Account of Translation – without a translation theory. In: Target: International Journal of Translation Studies, 2/2, 135-164 (1990)
9. Shepard, R.N.: Toward a Universal Law of Generalization for Psychological Science. In: Science, New Series, Vol. 237, No. 4820, 1317-1323 (1987)
10. Euzenat J., Shvaiko, P.: Ontology Mapping. Springer-Verlag, Berlin Heidelberg (2007)
11. Madsen B.N., Thomsen H.E., Halskov J., Lassen T.: Automatic Ontology Construction for a National Term Bank. In: Proc. Terminology and Knowledge Engineering Conference 2010, Dublin, 502-533 (2010)
12. Tenenbaum, J.B., Griffiths, T.L.: Generalization, Similarity, and Bayesian Inference. In: Behavioral and Brain Sciences 24, 629-640 (2001)
13. Frank, M. C., Goodman, N. D., Lai, P., Tenenbaum, J. B.: Informative Communication in Word Production and Word Learning. In: Proc. The 31st Annual Meeting of the Cognitive Science Society (2009)

Functions of Explicit Negation in German News Texts

Alexandra Klein[1], Brigitte Krenn[1], and Harald Trost[2]

[1] Austrian Research Institute for Artificial Intelligence (OFAI),
Freyung 6/6, 1010 Vienna, Austria,
[2] Section for Artificial Intelligence,
Center for Med. Statistics, Informatics, and Intelligent Systems,
Medical University of Vienna,
Freyung 6/2, 1010 Vienna, Austria

Abstract. Usually, humans have no problem interpreting negation in text. For NLP systems, there are so far no standard solutions for handling negations. Many NLP systems do not model negation phenomena and encounter difficulties whenever facts are taken as given even when the statement is negated. Besides from negating objective facts, negation may fulfill other functions in natural language, such as being part of a rhetorical relation. Thus, an adequate interpretation of negations requires a mechanism for distinguishing the different functions of negation as otherwise, the content of a text cannot be analysed properly. While negation has been extensively studied in linguistics, it seems that there is no classification of negation functions which can be used for NLP. Thus, a corpus-based study was carried out on German newspaper texts in order to derive a taxonomy of negation functions from the point of view of NLP. Four main categories have been identified: negations related to statements, discourse markers, speaker attitudes and idiomatic expressions. All categories may occur in the context of temporal markers which assign only a transient validity to the negated expression. The categorization serves as a base for a pattern-based negation-processing module which identifies and classifies negations.

1 Introduction

Negation fulfills various functions in text. In the following paper, we will argue that NLP systems need to handle negation, e.g. by means of a pre-processing module: some instances of negation may be vital to representing the underlying meaning of a text while negation in other contexts may not contribute to the propositional content of a core statement. No classification seems to be available which can be used to determine the contribution of negated expression to textual meaning from the point of view of NLP. Thus, we have carried out an empirical study on the use of negation in German newspaper texts as a base for a negation-processing module for the domain. The study is motivated by be needs of NLP applications such as Information Extraction, Sentiment/Subjectivity Analysis and Opinion Mining, and it analyzes the lexicalizations, contexts and functions

of the observed instances of negation in terms of their relevance for further automatic processing. While the study has been carried out on German newspaper texts and while the realization of negation is language-specific, we expect that the analysis of negation function may have many language-independent and domain-independent aspects.

For illustrative purposes, we will give some English examples from newspaper texts for what we consider the main functions of negation. An obvious function is the reversal of the truth value in a statement about an objective situation. In these cases, the statement is explicitly marked as false considering a specific situation and contrary to the assumed expectations on the reader's part [1,2]. In the example sentence

Example 1. Nicklas Bendtner has not travelled to Malaysia for the first part of Arsenal's pre-season tour as he closes in on a move away from the club.[3]

there is the underlying fact of Nicklas Bendtner's traveling to Malaysia with Arsenal. The statement marks this as false, given the current situation, and refutes the assumed prior expectation on the reader's part that the soccer player will join the pre-season tour. If negation is ignored in this sentence, e.g. by a bag-of-words approach, the extracted facts consist of Nicklas Bendtner's traveling to Malaysia, but they do not contain the vital information that actually, Nicklas Bendtner did not go on this journey with his teammates. Thus, Information Extraction applications need to include negations in a representation of the content of a text if the negations are part of statements.

Negation does not only occur in core statements regarding the propositional content of a sentence; it may also indicate a speaker attitude. In the sentence

Example 2. Melanie Phillips is, as you might expect, not happy either at the idea of the likes of Coogan and Hugh Grant taking a moral stance.[4]

the core statement is the fact that Coogan and Hugh Grant take a moral stance. But this is an embedded statement, and the negation is part of the introductory segment of the sentence containing the information that Melanie Phillips is *not happy* about this idea. Sentiment/Subjectivity Analysis or Opinion Mining systems would have to take into account this information as it refers to the opinion or attitude of a person with respect to a situation.

Negation may also signal different rhetorical relations. In the sentence

Example 3. Other incidents breached his privacy but not the law.[5]

[3] guardian.co.uk, "Nicklas Bendtner misses Arsenal's Asian tour as transfer nears", Sunday 10 July 2011, `http://www.guardian.co.uk/football/2011/jul/10/nicklas-bendtner-arsenal-transfer`

[4] guardian.co.uk,"News of the World phone-hacking scandal part one", Monday 11 July, `http://www.guardian.co.uk/news/blog/2011/jul/11/news-world-hacking-scandal-live`

[5] guardian.co.uk, "News International papers targeted Gordon Brown", Monday 11 July 2011, `http://www.guardian.co.uk/media/2011/jul/11/phone-hacking-news-international-gordon-brown`

the negation *not* is part of a contrast relation: the privacy was violated, but the mentioned incidents were not illegal. In this case, an Information Extraction system has to consider and resolve the negation as it is important to note that the sentence essentially contains two statements: 1) the incidents breached his privacy, 2) the incidents did not breach the law.

There is also a rhetorical relation which contains negation as a discourse marker which contributes less to the propositional content of the core statement:

Example 4. According to the newspaper, the forces of Murdoch not only attempted to gain access to Brown's voicemail but also obtained private banking and medical information.[6]

In this example, both statements are positive: 1) the forces of Murdoch attempted to gain access to Brown's voicemail, 2) the forces of Murdoch also obtained private banking and medical information. The two statements are connected by means of *not only... but also*. An Information Extraction system should not consider the negation if it does not extract rhetorical relations, but instead it should focus on the two positive statements.

Negations may also be modified by temporal markers:

Example 5. News International has not yet responded to the claims.[7]

Here, the negation refers to a situation at a certain point in time, which may be transitory. Temporal modifications may influence all types of negations. Information Extraction and Opinion Mining systems need to take into account temporal modifiers which may indicate transient situations.

As the examples have shown, it is important to analyse which parts of a statement or its context are negated and what the intended effects are. For the extraction of relations, it is essential to consider negated relations between terms or concepts as negative expressions are often much more than mere function words from the point of view of systems representing or annotating textual content.

2 Handling Negation in NLP

Most of research concerning NLP and negation has been carried out in the domain of biomedical texts. Negation is important in this domain as it may indicate the absence of certain relevant symptoms, pre-conditions, or adverse reactions. This is indispensable information for electronic patient records, quality control and billing purposes. In approaches for extracting information from biomedical text, negation is handled by defining or learning patterns for combinations of negation markers and medical concepts, cf. e.g. [3,4]. Corpora of biomedical texts

[6] American Journalism Review, "The Escalating Murdoch Scandal", Monday 11 July 2011, `http://www.ajr.org/Article.asp?id=5109`

[7] "The list goes on: Brown allegedly hacked by Sunday Times", Monday 11 July 2011, `http://citywire.co.uk`

which have been annotated with negation information have become available for English [5].

In contrast to NLP systems for biomedical texts, research on Information Extraction in other domains has just started to examine the functions and effects of negation. [6,7] represent recent approaches to incorporating negation in NLP beyond the medical domain. The application is the detection of conflicts and inconsistencies among pieces of information for automatic question answering.

Another field of research where negation has to be considered is Sentiment/Subjectivity Analysis and Opinion Mining; for an overview, cf. [8]. In order to represent negative contexts of opinion words, words or phrases are marked as negative, and they are added to the features which are used to determine the positive, negative or neutral polarity of an expression [9,10]. [11] distinguish between prior and contextual polarity of expression which may be reversed by negation. [12] describe a system for determining the scope of negation using Conditional Random Field (CRFs) models which are trained on the output of a dependency parer.

To our knowledge, no classifications of negation functions are available which can be adopted in NLP systems. [1] has carried out an empirical study on corpora of written and spoken English, and she has analysed the use of negation. She has derived a classification where she distinguishes between the two main functions REJECTION and DENIAL. REJECTION concerns suggestions while DENIAL concerns propositions. This classification comes from a corpus- or discourse-linguistics perspective; it lacks concern for the requirements of NLP applications. In particular, it does not differentiate between negated propositional content, negation in speaker attitudes, rhetorical relations and idiomatic expressions and the contribution of negation to textual meaning as analyzed by NLP systems.

3 Empirical study: Negation in a corpus of German news texts

Corpora of German newspaper texts serve as base for our analysis of negation. While there are many conventions regarding the style of newspaper articles, there is a large variety of expressions compared to other domains, and newspaper articles cover a wide range of topics, some describing temporal developments. Furthermore, large corpora are available – not only for English, but also for other languages. We therefore think that newspaper articles are a suitable base for a classification of negation function beyond more limited domains such as biomedical texts.

We consider only instances of explicit negation and only lexically realized negation. We did not take into account morphological negation, e.g. encoded in the prefix *un-*, as in approaches which are based on bag-of-words, a word with a negative prefix becomes a feature, while lexical negation has to be analysed in terms of its context. Thus, we define a negation element for our purposes as a lexeme which carries no meaning beyond its negative function, e.g. *no, not*, and

which belongs to a closed-class category. We do not take into account any other lexemes, e.g. *deny* which often signals negation in the biomedical domain:

Example 6. The patient denies any chest pain.

As a lexeme, *deny* carries a positive and a negative meaning. The act of stating something is positive, while the assertion that something is not true is negative. *deny* combines both aspects. The rationale for the exclusion of any lexemes which are not only negation 'function words' lies in the fact that, from the point of view of NLP, lexemes with positive/negative meaning tend to be domain-dependent, productive due to word formation processes, and fuzzy or non-binary in polarity assignment (e.g. *hardly*).

The NEGRA corpus [13] Version 2 is a syntactically annotated German corpus of newspaper texts. In order to derive a negation subcorpus, all sentences containing one or more negation elements were extracted from the first 4000 sentences of the NEGRA corpus. 626 (15.7%) of the 4000 sentences contain 705 negation elements, with *weder...noch*, the German equivalent to *neither...nor*, being counted as one negation element. Most sentences contain only one negation element, but there were sentences with up to four negation elements.

nicht ('not') is the most frequent negation element, which is not surprising, as it is the main negation lexeme for sentence negation in German, similar to the English *not*. Figure 1 shows a list of all negation elements which were found in the subcorpus with their frequency of occurrence.

Negation element	**Count**	**%**
nicht ('not')	468	66.7%
kein ('no')	103	14.7%
ohne ('without')	53	7.5%
nichts ('nothing')	23	3.3%
nie ('never')	21	3.0%
weder...noch ('neither...nor')	10	1.4%
niemand ('nobody')	8	1.1%
keinerlei ('no...at all/no...whatsoever')	3	0.4%
niemandem ('nobody'), dat)	3	0.4%
nirgendwo ('nowhere')	2	0.3%
keineswegs ('by no means')	2	0.3%
niemanden ('nobody', acc)	2	0.3%
keins ('none')	1	0.1%
nicht- ('non-')	1	0.1%
nirgends ('nowhere')	1	0.1%
niemals ('never')	1	0.1%

Fig. 1. Negation elements

In most cases, the grammatical negation scope consists of a verb phrase, a noun phrase or an adjectival phrase. For our approach, we have examined

negation context instead of grammatical scope. On the one hand, we wanted our negation-processing module to be robust, which precludes relying on deep parsing. On the other hand, the whole negated statement which is important for relation extraction and ontology building may consist of more than the verb phrase. Therefore, we have focussed on negation context in terms of the core statements which contain negations as well as the statements which are part of a rhetorical relation modified by negation.

4 Classifying the Function of Negation

For a bottom-up classification of negation functions, we have started from the sentence structure. Analyzing the negation subcorpus, it becomes apparent that negation appears in four types of sentence structure: Statement, Rhetorical Relation, Speaker Attitude and Negative Polarity Item/Idiomatic Expression. Figure 2 shows how the 705 lexically realized negation elements appear as part of a statement, a rhetorical relation, a speaker attitude, and a NPI, based on the manual annotation.

Type	**Count**	**%**
Statement	580	82.3
Rhetorical relation	90	12.8
Speaker attitude	29	4.1
NPI/Idiomatic expression	6	0.9

Fig. 2. Contexts and functions of negation

In the following, we will briefly describe the four types of functions and illustrate them with examples from the corpus. All negation functions may appear in the context of temporal information. Temporal modifiers were marked in the corpus.

4.1 Statement

For the purpose of identifying the scope and function of negation, we classify a statement as propositional content which may be represented in terms of a predicate-argument structure. A statement may be lexicalized as a clause or a noun phrase. Negation in single statements may negate the predicate (a verb in a verb phrase, a nominalization in a NP) or one or more of its attributes.

The following sentence from the corpus (sentence number 35 in the NEGRA corpus), which is taken from a review of a concert, gives an example of a negated statement:

Example 7. Selbst die flotteren Passagen werden nie ausgelassen und fröhlich. (35)
('Even the livelier passages never become jolly and cheerful.')

4.2 Rhetorical relation

Negations may be part of rhetorical relations. Within rhetorical relations, negations may contribute to the propositional content of a statement, or they may mark the relation without negating a statement. Relations between statements establish textual coherence. Rhetorical Structure Theory (RST) [14] and Segmented Discourse Representation Theory (SDRT) [15] rely on varying inventories of discourse relations in text analysis and generation. In some cases, a subset of relations is selected which is most relevant in a specific application. Rhetorical relations connect sequences of statements; they may be implicit, or signaled by discourse markers.

For Information Extraction, it is important to distinguish the contribution (with respect to propositional content) and function of negation. All instances of explicit negation in the newspaper subcorpus were examined in order to determine if they occur as part of a rhetorical relation. The inventory of rhetorical relations was taken from the SDRT relations defined in [15]. [16] presents an approach for using linguistic cues to classify rhetorical relations from a subset of the SDRT relation inventory in [15]. In the newspaper negation subcorpus, three different types of rhetorical relations were associated with negations:

- **CONTRAST:**
  nicht...sondern (*not...but*), *nicht...aber* (*not...however*) signal a CONTRAST relation. The first part of the sequence describes a statement (referring to an expectation) which is marked as false. In the second part, it is replaced by information which describes the actual situation. Both segments usually share some common features in form and content. In CONTRAST contexts, negation is important as it explicitly rejects an expectation.
  Example from the corpus:
  nicht das Wort, sondern die Tat ist wichtig (2978)
  ('not words, but actions are important')

- **CONTINUATION:**
  Patterns of the type *nicht nur...sondern* (*not only...but also*), signal a CONTINUATION relation. The second part of the sequence, which may contain more unexpected information than the first part, is added by means of negating restrictive adverbs such as *nur* or *allein* ('only', 'simply'). In CONTINUATION contexts, the negation only has a rhetorical function. Its contribution to the propositional content can be neglected. For the purposes of analysis in an NLP system, the negation and restriction can be removed, and both parts can be treated as related facts with equal contributions.
  Example from the corpus:
  Nicht nur ein Tanz, sondern ein Gefühl (3437)
  ('not only a dance, but a feeling')

- **CONDITION:**
  wenn nicht... (*if not...*) or *wenn...nicht* (*if...not*) specify a CONDITION for

a specific situation. Negation explicitly contributes to the truth conditions which must be given for one statement in order for a second statement to be applicable.
Example from the corpus:
Allerdings soll dies nur möglich sein, wenn sich Kunst nicht in den Dienst politischer Forschrittskonzepte stellt. (1810)
('however this should only be possible if art does not adhere to political concepts of progress')

If negation is part of a rhetorical relation, negation marks a sequence of statements and establishes a relation between them in form and content.

4.3 Speaker attitude

Newspaper commentary is a genre which contains a complex mixture of facts and opinions [18,17]. In the subcorpus, speaker attitudes were expressed mostly by verbs and adjectives. If a negation occurred, it reversed the meaning of the verb or adjective, e.g.

Für nicht empfehlenswert hält er Fußball [...] (2224)
('He does not recommend (playing) soccer')

[1] notes in her empirical study that negation tends to occur often with mental verbs such as *think*.

4.4 NPIs/Idiomatic expressions

While the previously described categories for categorizing negations employ functional criteria, lexicalized expressions containing negation elements were assigned their own category since these expressions have to be represented as a whole in order to grasp textual meaning. These cases are important for Information Extraction systems as bag-of-words approaches may come up with distorted results if idiomatic expression are taken literally. Lexical items may prefer or establish negative contexts. In some cases, the resulting expressions are idiomatic expressions, e.g. *He didn't budge.* – *Er zuckte nicht mit der Wimper.* Other examples not being idiomatic expressions in the strict sense are lexemes which are likely to occur within negative constructions, e.g. *any* in English.

In the subcorpus, only six – different – instances of idiomatic expressions containing negation occurred, cf. Figure 3.

For German, a list of NPIs is available which has been extracted from large corpora using co-occurrence information [19,20]. This list is helpful for automatically detecting negation when it is part of an NPI. Since collocations are rigid structures with a lexicalized meaning, negation in NPIs should be marked so that it will not distort the result during relation extraction.

Expression	Translation	No.
nicht Kinder von Traurigkeit sein	'to know how to enjoy oneself'	(235)
kein Geringerer als	'none other than'	(1084)
nicht hinter dem Berg halten	'to be unhesitating'	(1678)
nicht schlecht staunen	'to be astonished'	(3528)
nicht verhehlen	'not to conceal'	(3616)
über Geld nicht reden sondern es haben	'not to talk about money but to have it'	(3744)

Fig. 3. NPIs/Idiomatic expressions

5 Automatic pattern-based analysis of negations

The analysis of negation in the NEGRA corpus, which was described in the previous sections, has resulted in a gold standard as well as patterns for negation recognition and classification. In order to examine whether the classification which was derived from the NEGRA corpus can be applied to other news texts, a second corpus was built which consists of articles from the Austrian daily newspaper *Der Standard*. 65 articles published on the same day in various sections of the newspaper were selected. The corpus was processed by a part-of-speech tagger [21], which also assigned sentence boundaries. In contrast to the situation in the NEGRA corpus, both part-of-speech tags and sentence boundaries in the *Standard* corpus were not corrected manually. The sentences were divided into segments according to rules. The aim was to create segments consisting of a one statement each. The patterns for negation identification and classification were formulated as regular expressions, and they were applied to the *Standard* corpus. The results of the negation identification and classification were then compared to a manual annotation.

According to the assignment of sentence boundaries, the corpus contains 1611 sentences; 169 sentences (i.e. 10.5 % of the 1611 sentences) contain at least one negation element. The 1611 sentences in the *Standard* corpus were automatically separated into segments, yielding 4952 segments. 171 segments (i.e. 3.45% of the segments) contain at least one negation element. In the 171 segments, 138 cases of negated statements, 15 cases of rhetorical relations (9 for the CONDITION relation, 4 for the CONTRAST relation and 1 for the CONTINUATION relation), 17 cases of speaker attitudes and only 1 case of an NPI were assigned. 14 temporal markers were identified by the system.

All negation elements were identified correctly, which comes as no surprise as we have concentrated on explicit lexical negation, with a closed class of lexical items. Of the 169 sentences containing at least one negation element, 112 (66.3 %) were segmented correctly, i.e. in a way that the negated statement was segmented. In the remaining 37 cases, a deeper syntactic analysis is needed. Since in contrast to the NEGRA corpus, the *Standard* corpus is raw and was not cleaned manually, many segmentations were assigned where punctuation was not disambiguated or where part-of-speech tags seem to suggest a segment boundary.

13 segmentation problems are due to coordination as it if often difficult to determine wheter coordination connects different statements or whether it connects different elements within one statement, e.g.

Example 8. Man soll ja nicht überkritisch oder vorauseilend böse sein [...]
('one should not be overcritical and anticipartorily mean [...]')

where the negation scope extends to both adjectives, but the segmentation separates them at the coordination *and* versus

Example 9. Wenn nichts passiert und die Regierung auf das Erreichen des Nulldefizits wartet [...]
('if nothing happens and the government awaits reaching zero deficit [...]')

where the coordination separates two statements.

The *Standard* negation corpus contains only one NPI, and it was not found by the system as it is not contained in the NPI list for German [20]. This may be due to the fact that it is an expression which is common in Austrian German, but less common in Standard German (*Schmied vs. Schmiedl* 'expert vs. non expert').

Of the 17 instances of speaker attitudes were assigned, none can be considered an actual speaker attitude. The patterns only apply to cases where the speaker attitude concerned an embedded sentence with the actual statement describing an objective situation. 12 of the classifications are due to errors of the part-of-speech tagger. Thus, it turned out that the assignment of speaker attitudes did not yield any correct assignments without taking into account the actual opinion words; the integration of lexical indicators for speaker attitudes seems necessary.

For rhetorical functions, however, all assignments in the *Standard* negation corpus are correct. In sum, all 15 instances of rhetorical functions were found, namely 9 for the CONDITION relation, 4 for the CONTRAST relation and 1 for the CONTINUATION relation.

Of the 14 temporal modifiers identified in the *Standard* negation corpus, 4 indicate 'not yet' and 10 indicate 'not anymore'. 2 of these assignments turned out to be false positives. There are no false negatives. The errors are due to lexical ambiguities.

6 Conclusion

Unstructured texts are layered structures which convey information about facts, (e.g. causal or temporal) relationships and attitudes to readers. Each negation element occurring in a text plays a role in presenting this information by emphasizing that a certain assumed fact is not valid in a given situation. Human readers or listeners are usually able to interpret this information given the context. They develop a representation of the positive facts in a statement and reverse the polarity of the negated aspects of the statement. Apart from systems in the biomedical domain, where negation tends to be comparatively straightforward, and opinion mining approaches, where negation directly contributes to

assigned sentiment polarities, NLP systems mostly rely solely on the positive facts, without considering negated contexts. Since the contribution of negated information to the propositional content depends on the function of negation in a specific context, different types of negations need to be distinguished. In order to come up with a classification of negation function, an empirical study was carried out; German newspaper texts were selected as a sample domain.

A negation subcorpus was derived by selecting sentences which contain explicit lexical negation from a sample of the NEGRA corpus. A classification of the functions of the observed negations and their contexts was derived, based on the perspective of NLP applications such as Information Extraction and Opinion Mining. Negation elements, scopes and functions were annotated in the subcorpus.

As a next step, the negation phenomena which were encountered in the NEGRA subcorpus were transformed into regular expressions for a pattern-based negation-processing approach. A small raw corpus of texts from the Austrian newspaper *Der Standard* was part-of-speech tagged, and the patterns were applied to the test corpus. The results are encouraging: it turns out that even this simple approach performs reasonably well identifying factual negation, rhetorical relations and temporal modifiers. For the classification of negation in the context of speaker attitudes, more lexical resources are needed. The segmentation of negated statements still needs to be improved.

From the point of view of linguistic resources, the empirical study has resulted in an annotated negation subcorpus of German newspaper texts which may prove useful in various applications. Furthermore, a classification of negation functions was derived which describes the negation phenomena in the German newspaper corpus but which also can be applied to other languages and domains. From there, patterns describing functions of negations in German were created which will be extended and integrated into a negation-filtering module which can be used by NLP systems for German.

Acknowledgment

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology.

References

1. Tottie, G.: (1991). *Negation in English Speech and Writing: A Study in Variation*. San Diego: Academic Press, 1991.
2. Horn, L. R.: *A Natural History of Negation*. The David Hume Series: Philosophy and Cognitive Science Reissues. CSLI Publications, 2001.
3. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., Buchanan, B. G.: A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, **34**:301–310, 2001.
4. Morante, R.: Descriptive Analysis of Negation Cues in Biomedical Texts. *Proc. of LREC'10*. 2010.

5. Vincze, V., Szarvas, G., Farkas, R., Mòra, G., Csirik, J.: The BioScope Corpus: Annotation for Negation, Uncertainty and their Scope in Biomedical Yexts. *BMC Bioinformatics*, 9(Suppl 11):S9, 2008.
6. Harabagiu, S., Hickl, A., Lacatusu, F.: Negation, Contrast and Contradiction in Text Processing. *Proc. of AAAI06*, 2006.
7. de Marneffe, M.-C., Rafferty, A., Manning, C. D.: Finding Contradictions in Text. *Proc. of ACL/HTL*, 2008.
8. Wiegand, M. , Balahur, A., Roth, B., Klakow, D., Montoyo, A.:. A Survey on the Role of Negation in Sentiment Analysis. *Proc. of the Workshop on Negation and Speculation in Natural Language Processing*, Uppsala, Sweden, 60–68, 2010.
9. Das, S. R., Chon, M. Y.: Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53:1375–1388, 2007.
10. Na, J.-Ch., Sui, H., Khoo, Ch., Chan, S., Zhou, Y.: Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews. In *Conference of the International Society for Knowledge Organisation (ISKO)*, pp. 49–54, 2004.
11. Wilson, Th., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 347–354, 2005.
12. Councill, I., McDonald, R., Velikovich, L.: What's great and what's not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 51–59, Uppsala, Sweden, July 2010. University of Antwerp.
13. Skut, W., Krenn, B., Brants, T., Uszkoreit, H.: An Annotation Scheme for Free Word Order Languages. *Proc. of ANLP-97*, Washington, DC., 1997.
14. Mann, W. C., Thompson S.: Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text*, 8(3):243–281, 1988.
15. Asher, N. and Lascarides, A.: *Logics of Conversation*, Cambridge University Press, 2003.
16. Sporleder, C., Lascarides, A.: Exploiting Linguistic Cues to Classify Rhetorical Relations. In *Proceedings of RANLP-05*, pages 532–539, Borovets, Bulgaria, 2005.
17. Asher, N,, Benamara, F., Mathieu, Y. Y.: Distilling Opinion in Discourse: A Preliminary Study. In *Proceedings of COLING*, volume Companion Volume: Posters, pages 7–10, Manchester, UK, 2008.
18. Stede, M.: Surfaces and Depths in Text Understanding: The Case of Newspaper Commentary. In *Proc. of the HLT/NAACL Workshop on Text Meaning*, Edmonton/AL, 2003.
19. Lichte, T. and Soehn, J.-Ph.: The Retrieval and Classification of Negative Polarity Items using Statistical Profiles. In *Roots: Linguistics in Search of its Evidential Base*. Mouton de Gruyter, 2007.
20. Soehn, J.-Ph., Liu, M., Trawinski, B., Iordachioaia, G.: Positive und Negative Polaritätselemente als lexikalische Einheiten mit Distributionsidiosynkrasien. In *Proceedings of the Europhras*, 2008.
21. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, September, 1994.

{AskMe*}: Reducing the costs of adoption, portability and learning process in a natural language interface to query databases

Miguel Llopis[1] and Antonio Ferrández[2]

Dept. Languages and Information Systems, University of Alicante
[1] *mll9@alu.ua.es,*
[2] *antonio@dlsi.ua.es*

Abstract. This paper describes the underlying design concepts and techniques that are being used in {AskMe*}; a database-independent template-driven natural language interface to query databases aid by domain-specific query authoring services, enabling low adoption and portability costs and a fast learning curve for end users.

Keywords: Natural language interface, intermediate representation languages, relational database, web ontology language, ontology extraction, concept hierarchy, query-authoring services.

1 Introduction

There have been many different attempts to improve the way users communicate with computers, by means of a language understandable for both parties. Among other disciplines in Computer Science, database access is a very important area. In this scope, there exists a standardized, industry-wide accepted language, known as Structured Query Language (SQL). While SQL is universally used in the context of database access, it is yet a language for specialized knowledge, Information Technology-focused, users [1].

In order to make databases accessible for non-technical users, there have been many attempts to build a rich tooling environment around databases that allow users to query databases using their own natural language, and therefore, keeping them from learning a technical or machine-friendly language. These interfaces are commonly known as Natural Language Interfaces to query databases (NLIDB). Natural language interfaces can be either textual or based on user interface (UI) menus. However, each of these two main approaches presents some problems, as we will explain later.

This paper describes the design of {AskMe*}, a system that allows non-technical users to query databases by means of a natural language, rich text-based approach, combined with a set of authoring services that will help the user formulating queries, as well as help the system to parse and translate these queries in a more accurate and efficient way. The system is reconfigured automatically any time it is connected to a

new database. This way, the portability and configuration costs are minimized and, by means of the query-authoring services, the user learning process is dramatically accelerated as well.

2 State of the art in NLIDBs

Due to the steady growth of the amount of information and services available on the Web, or within intranets, the need for intuitive and effective user interfaces to access data becomes more and more crucial. NLIDBs essentially take as input a question expressed in natural language and return an answer from a given knowledge base [2]. The use of NLIDBs has been a field of research since the 1960s; see for instance LUNAR [15]. In these five decades of NLIDB research, a lot has been said and therefore an exhaustive description of each approach, and the underlying sub-problems, is out of the scope of this article. However, we will briefly enumerate some of the main approaches to the design of an NLIDB that relate to the design decisions made in {AskMe*} and describe a few existing systems developed recently from which {AskMe*} has taken some ideas and is now trying to improve, by opening new design questions that will help building more user-friendly systems and reducing system deployment costs to new domains or databases.

2.1 Textual NLIDBs vs. Graphical NLIDBs

One of the biggest questions in the space of NLIDBs through decades has been the disjunctive of choosing either a textual or a graphical user interface to build the system [13].

In one hand, textual interfaces are great mechanisms for the purpose of writing queries in a natural language because users are not required to learn new artificial languages in order to be able to communicate with the system. Natural languages are very useful in some cases where the question can be easily expressed with words, but not with a UI menu-based solution [11].

In contrast, there are still some open issues for textual natural language interfaces [11, 12]. For instance, the fact that the linguistic coverage of the system is not obvious to the users, making them confused and not aware of what kinds of language structures are supported by the system and which are not.

Another problem of these systems is the overlap of linguistic failures and conceptual failures [13], a user can easily think that a question has been rejected by the system because it has been ill-formed from the linguistic standpoint, while the actual reason is that the concepts behind the formulated question do not correspond to the underlying database conceptual model (for instance, a user can query the system for “the birth date of a customer” which may not be reflected on the system, as opposed to “the birth date of an employee” which may be in the system). However, there is not any mechanism by which the system can warn the user about this issue until the query has been fully processed.

This problem could be solved by using a UI menu-based solution in which, once the user has selected the entity “customer” the system does not provide an option for

"birth date". While this approach solves the problem exposed here, UI menu-based solutions usually lack of flexibility and the expressivity power is drastically reduced, compared to textual natural language interface systems, as the user's expressivity power is pretty much restricted to what the UI design allows, rather than providing the expression power of the user's own natural language words [13].

As interesting as this question of textual vs. graphical NLIDBs is, there exists also a third option that stands in the middle of both approaches and consists on having textual NLIDBs that leverage the power of rich text: bold, italics, font colors, squiggles, text completions, etc. To our best knowledge, this option has not been intensively explored yet in the field of NLIDBs, while, however, it is very common in domain-specific languages from other contexts, such as programming languages Interactive Development Environments (IDEs). We decided to leverage these useful design concepts in {AskMe*}. Interestingly, the combination of textual NLIDBs with rich query-authoring services, as we call this set of features (syntax coloring, error squiggles, completions, etc.), provides a substantial improvement in the user experience when writing queries, especially in regards to query accuracy, as we will describe later in the evaluation section.

2.2 Portability of NLIDBs

The problem of portability of NLIDBs is, from our perspective, one of the most critical ones to be solved. By itself, the cost of developing an NLIDB can be very high, and in most of the approaches taken for creating NLIDBs, the resulting systems are tightly coupled to the underlying databases.

In the last few years, there have been interesting approaches to the design of NLIDBs that are database-independent [e.g. 3, 4], in the sense that they can cope effectively with queries targeting different domains without requiring substantial reconfiguration efforts. One of the best examples of this approach is PRECISE [3]. This system combines the latest advances in statistical parsers with a new concept of semantic tractability. This approach allows PRECISE to easily become highly reconfigurable. In addition, this was one of the first NLIDB systems that used the parser as a plug-in, so it could be changed with relative ease in order to leverage newest advantages in the parsers' space.

An interesting advantage of adapting the parsing process to each of the knowledge domains that the system connects to is that analyzing an input question in NLIDB systems is often based on a part-of-speech (POS) tagging, followed by a syntactic analysis (partial or full) and finally, a more or less precise semantic interpretation. Although there are broadly accepted techniques for POS tagging (e.g. [7, 8, 9]) and syntactic analysis (e.g. [10, 8]), techniques for semantic parsing are still very diverse and ad hoc. In an open-domain situation, where the user can ask questions on any topic, this task is often very difficult and relies mainly on lexical semantics only. However, when the domain is limited (as is the case of an NLIDB), the interpretation of a question becomes easier as the space of possible meanings is smaller, and specific templates can be used [14]. It has been demonstrated [16] that meta-knowledge of the database, namely the schema of the database, can be used as an additional resource to better interpret the question in a limited domain.

Another interesting existing solution, based on the creation of a new NLIDB every time that the system is connected to a new database, is the one developed for the CINDI virtual library [4], which is based in the use of semantic templates. The input sentences are syntactically parsed using the Link Grammar Parser [17], and semantically parsed through the use of domain-specific templates. The system is composed of a pre-processor and a run-time module. The pre-processor builds a conceptual knowledge base from the database schema using WordNet [15]. This knowledge base is then used at run time to semantically parse the input and create the corresponding SQL query. The system is meant to be domain independent and has been tested with the CINDI database that contains information on a virtual library. The improvements which this research work provides in regards to the portability problem space are described in the next sections.

2.4 Contribution of our approach compared to previous related work

The main improvements of our proposal compared to other existing systems are the significant reduction of costs: implementation and reconfiguration costs are optimized due to the dynamic nature of the system; learning costs for end users are greatly reduced as well thanks to the use of query-authoring services.

We propose the combination of textual NLIDBs with rich query-authoring services (syntax coloring, error squiggles, tooltips, etc.). This provides a substantial improvement in the user experience when writing queries, especially in regards to query accuracy in order to solve both linguistic failures and conceptual failures, which could not be fully solved by the use of menu-based user interfaces either. The use of query-authoring services helps to reinforce the conceptual center of the dialog between the user and the NLIDB around the domain entities in focus.

In order to achieve this, we combine domain-specific information, captured in concept-hierarchy ontologies any time the system is connected to a new database. The system automatically generates the syntactic and semantic parsing templates and the rest of components needed in order to provide query-authoring services. In addition, the system is fully reconfigurable without the need of any specialized knowledge. This is a significant improvement compared to the existing portable solutions mentioned above, because it makes the entire reconfiguration process fully transparent to the end users as opposed to having to perform some reconfiguration steps for entity mapping, disambiguation, etc. This is a substantial improvement not only by the amount of extra work that we save in the reconfiguration steps, but also because it enables a system to be automatically managed without the need of an advanced user to participate in the process. This represents a step towards the democratization of NLIDBs, as users fitting a non-technical profile will be able to use the system on their own, throughout the entire system lifecycle, from the very early steps of adoption and deployment of the system towards a real-world production environment to the management, reconfiguration and diagnosis steps across multiple domains, for which the system is able to adapt itself automatically. In this sense, also, the role of query-authoring services is fundamental because they enable to perform the very little manual reconfigurations needed, if any, driven by intuitive real-time hints in the query authoring process.

3 Overview of {AskMe*}

{AskMe*} is a database-independent NLIDB and uses a template-based approach for the dynamic generation of the lexer, syntactic and semantic parsers. Figure 1 shows the different modules of the system.

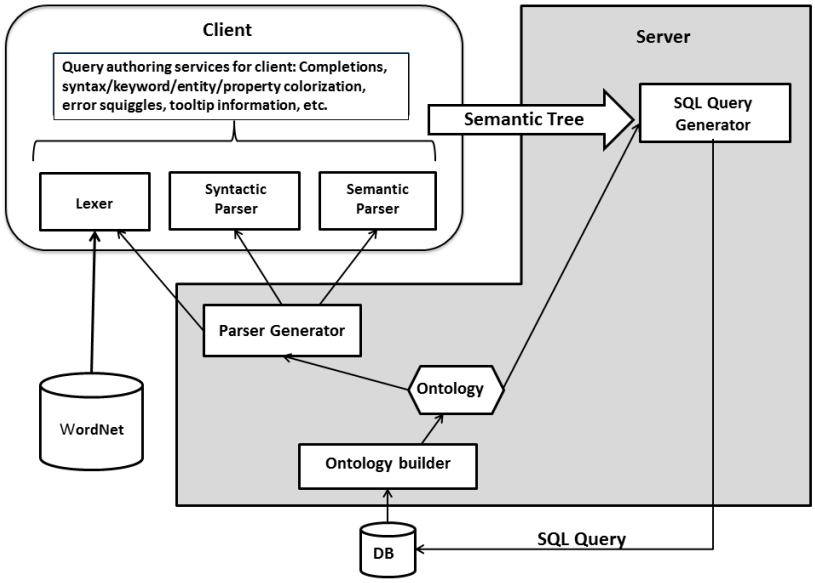


Figure 1: {AskMe*}'s high level architecture

An exhaustive description of every component of the system is out of the scope for this article, instead we will focus on describing the most relevant techniques that enable the proposed improvements of the system compared to other state-of-the-art systems: dynamic generation of the system and query-authoring services.

3.1 Dynamic ontology generation

Every time the system is connected to a new database, a search for its corresponding ontology in the ontology repository is performed in the server. This repository is a dictionary in which each entry contains information about the database (server name, DB name) and ontology name. If the ontology already exists, it is loaded into the system so the rest of modules that depend on this ontology can be subsequently reloaded as well. If it does not exist, the process to build a new ontology is started.
We are using OWL for representing ontologies. In order to build the ontology for each database, and keep the system within a manageable range of data volume, only the minimal information needed from the database is stored into the ontology. Concretely, entity names, properties and value types are mapped from the database into the ontology, while the actual data is not.

Database component	OWL component
Table/Entity	Class
Column	Functional Property
Column Metadata: - Data type. - Mandatory/Non-nullable. - Nullable.	OWL Property restriction: - AllValuesFrom restriction. - Cardinality() restriction. - MaxCardinality() restriction.

Table 1: Database - Ontology mapping

During our investigation prior to the design of the system, we explored different alternatives and mechanisms to build the ontology following the mapping described in Table 1. From all of them, we found that the concept-hierarchy ontology construction described in [9] was the best choice for our system; because it was really optimized for minimum size and maximum performance.

3.2 Dynamic parser generation and query-authoring services

As most of the state of the art NLIDB systems [2, 3, 4, 5, 11, 13], {AskMe*} uses an staged parsing for transforming from a natural language query into a T-SQL query that the underlying database engine can understand. This parsing is performed at three different levels: lexical, syntactic and semantic. Once these three steps are completed, we will have a unique in-memory parsing tree. This tree will be then transformed into a T-SQL query, using the mapping against database schema objects.
After the ontology has been built, the rest of components that depend on this domain-specific knowledge have to be built accordingly as well. In order to do that, we are using T4 templates [18] which will allow to inject the specific parts into each declarative piece (lexicon, syntactic grammar, semantic grammar, etc.). In the next subsections, we describe the dynamic generation process and query-authoring services offered by each of these components.

3.2.1 Lexer

The first one of the key modules for the parsing process of {AskMe*} is the lexer. In order to build the lexicon, we combine the set of nouns derived from the domain knowledge contained in the database, namely the entity and property names, with a general-knowledge vocabulary terms, mostly verbs, adjectives and adverbs. We are retrieving these general-knowledge vocabulary terms from WordNet, a large lexical database of English. This database classifies nouns, verbs, adjectives and adverbs into sets of cognitive synonyms. Thanks to these cognitive synonyms sets, we are also able to complement the existing set of domain-specific nouns (entities and properties from the domain ontology) with an important amount of synonyms, into the system lexicon. This aspect is very important, as it will allow the lexer to automatically accept terms that, even when they are not the exact noun used in the underlying database schema, represent the same concept for the user. For example, the database may contain a property called “Telephone” for the entity “Customer”; while the user probably refers to it simply as “Phone”. The lexer is able to recognize “phone” as a valid term as well.

Besides the dynamic lexicon generation process, an interesting feature from the lexer to be mentioned is the lexical error detection capability. Once the system has been configured, a user can start typing in a query and it will be processed by the lexer at first. Every time that a white space has been added to the buffer, the lexer analyzes the term that goes immediately before this white space and decides whether it is valid from the lexical perspective or not. If the term does not appear in the lexicon, the lexer will tag it as an invalid lexical item. This tag information is automatically retrieved by the query-authoring services component, that will underline the invalid term with red squiggles in the query bar, making it evident to the user that the underlined part is wrong in his query, even before he finishes writing it, and displaying the completions suggestions for this term in the sentence, as part of the error message.
The other query-authoring service offered by {AskMe*} at lexer level is the completion suggestions mechanism, which offers in a dropdown pop-up menu, which appears below the word that the user is currently typing, the set of suggested words that contain the portion typed by the user as a fragment. This helps the user to remember the exact word that he is trying to write, and also to autocomplete it, making him write queries faster.

3.2.2 Syntactic parser

We are using a custom syntactic parser implementation that leverages the Link Grammar Parser for the core syntactic parsing operations. The parser capabilities are exposed via an Application Programming Interface (API) to make it easy to incorporate the parser into third-party systems. In the case of {AskMe*}, we are leveraging this API in order to customize the set of functionalities that Link Grammar Parser performs, we provide our own domain-specific dictionary of words constructed from the domain ontology (as previously described), and implement concurrency mechanisms to enable a more interactive communication with the query-authoring services system. These concurrency mechanisms implemented on top of the Link Grammar Parser API are based on event notifications for all the syntactic parser events: every time the parser processes and tags a fragment of the input query an event is generated, containing information about the syntactic classification for each token. This is a key component for driving the syntactic query-authoring service that {AskMe*} implements: syntactic error squiggles. These squiggles warn the user about syntactic errors in a query, even before the query authoring has been fully completed.

3.2.3 Semantic Parser

The third parsing step performed to an input query is the semantic parsing. In {AskMe*}, given its dynamic domain-specific knowledge acquisition nature, it may be feasible to find that a certain query is valid according to the lexical and syntactic analysis, but does not represent a concept that fits into the current domain. For example, the query “Name and date of the customers from the country where most orders were made in 2010” could be lexically and syntactically valid, all the terms in the sentence may be present in the dynamic lexicon, and the syntactic construction

and order of words match one of the valid categories of phrases in the Link Grammar Parser. However, as you will notice, the concept of Date may not exist for the entity Customer. This is definitely an error in the input query, a semantic error.
In order to detect this kind of errors, the semantic parsing step is applied to the input query. The semantic parser is guided by the use of semantic templates which are filled with the concepts captured in the domain ontology. The set of rules that are modeled by these dynamically-generated semantic templates are:

- **Entity-Property correspondence**: This rule enforces that all the requested properties for an entity in a query are indeed part of the current domain schema.
- **Cross-entities relationships**: This rule is applied to queries that contain multiple sub-phrases, and its purpose is to enforce that there exists a foreign-key relationship in the database schema between the entities in the query.

Schema Relationships	Input Query	Result	Reason
Customers-Orders and Orders-Products	Products from customers whose last name is Llopis.	Fail	There is not an existing relationship between products and customers in this domain.
Customers-Orders and Orders-Products	Products that were ordered by more than 100 customers in 2010.	Success	Products are related to orders, and every order references a customer.

Table 2: Examples of semantic rules behavior

- **Entities' default attributes**: There are cases in which the query is valid from a lexical, syntactic and semantic analysis, but it does not specify which attributes must be present in the result. For instance, the query that we previously considered: "Products that were ordered by more than 100 customers in 2010", does not specify which product properties we are interested in. This semantic rule does not invalidate a given input query, but rather imposes that the resulting T-SQL query must return all the product attributes that are not-null in the database schema, such as the Product ID, Product Name, Price, etc. This information, as we explained previously, was captured in the domain ontology as an OWL cardinality metadata attribute.

You can see examples of these rules in Table 2. In the case that one or more of these semantic requirements are not met by the input query, the semantic analysis would report errors. These errors are notified to the system in the form of events. The query-authoring services component is subscribed to these semantic events, in the same way as it is to the lexical and syntactic ones, and would therefore notify the user in a visual way about the issue, by highlighting the portions of the input query that cause the inconsistency. When the user hovers with the mouse over these highlighted regions, a tooltip containing a description of the inconsistency comes up. This description is also template-based, see Table 3.

Inconsistency Type	Error Description Message
Entity – Property mismatch	"EntityA" does not contain a property called "PropertyA". *(Where EntityA and PropertyA are the values in a query)*
Missing relationship	"EntityA" and "EntityB" are not related to each other.

Table 3: Examples of template-based semantic error messages

4 Evaluation

In order to evaluate the effectiveness of our approach, we are applying three different metrics. The first one is to evaluate the accuracy of our query interpretation process in a concrete domain. For that purpose, we evaluated our system using data from the Air Travel Information (ATIS) domain [20]. The selection of ATIS was motivated by three concerns. First, a large corpus of ATIS sentences already exists and is readily available. Second, ATIS provides an existing evaluation methodology, complete with independent training and test corpora, and scoring programs. Finally, evaluating on a common corpus makes it easy to compare the performance of the system with those based on different approaches. Our experiments utilized the 448 context independent questions in the ATIS "Scoring Set A", which is one of the sets of questions of the ATIS benchmark, generally the most commonly used for the evaluation of other systems, and the one that let us compare with most of them. {AskMe*} produced an accuracy rate of 94.8%. Table 5 contains a comparison of the results obtained by {AskMe*} in the ATIS benchmark, to other state-of-the art systems.

HEY [27]	SRI [26]	PRECISE [3]	{AskMe*}	MIT [25]	AT&T [24]
92.5	93	94	**94.8**	95.5	96.2

Table 4: Accuracy comparison using ATIS between various NLIDB systems

The second metric that we are using is measuring how the use of query-authoring services improves the overall usability of the system, by enabling early detection of query errors. In order to do that, we asked a set of ten users to write a number of queries in a given domain. These users were completely new to the system and they did not have any previous knowledge about the underlying domain. We gave them an initial description of the database, without schema representation or concrete entity/property names, and let them query the system in an exploratory way. During this process, users are very likely to introduce mistakes in most of the queries they come up with for the first time. We captured traces for all of these queries and recorded in which stage of the parsing process they were raised.

Our results indicate that, from the set of fifty input queries per user, almost 90% of them contained errors, from which roughly the 80% of these wrong queries could be detected before they were translated into SQL and, therefore, before they were being executed against the database. This fact results in significant improvements in terms of latency time for wrong queries, since thanks to the query-authoring services that {AskMe*} implements, they are locally detected by the system instead of being translated into T-SQL and executed against the database.

The results of this experiment show that while an important amount of errors (23%) are due to lexical errors (usually things like typos), and 26% correspond to syntactic errors (mostly ill-formed sentences in the English language), most of the errors are due to semantic errors (51%). In order to help minimizing the probability of having lexical errors in a query, the system provides auto-completion for entities and properties, and also auto-correction of typos based on distance-editing algorithms.

Regarding semantic errors, Figure 2 shows their distribution classified by the main semantic rules that {AskMe*} implements. In terms of semantic error distribution classified by the main semantic rules that {AskMe*} implements, this evaluation determines that 51% of them fall in the rule of entity-property mismatch, thus being the most common semantic error, 41% of errors correspond to queries trying to refer to a missing relationship that does not exist in the domain and the remaining 8% represents semantic errors due to the query specifying invalid values in property conditions.

Our third metric focuses on evaluating the portability of the system. For this purpose, we have created a script that simulates the user actions through the visual interface. In this test, the system will be connected to three different databases that we have previously configured: ATIS, AdventureWorksDB [22] and Northwind [23]. For each of these database connections, a custom benchmark made up of fifty different queries is executed against the system, asserting that the query-authoring services work as expected and that the resulting T-SQL query is generated as expected as well. Finally, the test also evaluates the behavior when the system is connected to a database that had been already connected before, checking that the ontology generation process is not kicked-off again, but rather the existing ontology for that source is pulled back from the store and brought into the current connection context. The results of this experiment indicate that there is not any lose in accuracy after a reconnection to a different database, and the results are the same as if the system was only connect to a single database for its lifetime. This means, the same results as in the first and second metrics apply to the scenario of multiple database reconnections.

5 Conclusion

{AskMe*} is an adaptive natural language interface and environment system to query arbitrary databases. Internally, the system leverages an ontology-based approach in which a new ontology is auto-generated every time the system is connected to a different database. Once this ontology has been generated, the rest of the system – domain-specific grammar, query-authoring services, etc. –reconfigures itself based on the set of language terms and relationships contained in the ontology. This automatic reconfiguration enables an effective lexical, syntactic and semantic validation of an input query, which will result in a higher accuracy of the system. The evaluation process showed how, despite the system is not specific to any concrete domain, the result of 94.8% of accuracy against the ATIS benchmark is relatively good compared to other existing state-of-the-art systems, both domain-dependent and independent.

Furthermore, this approach enables full portability of the system without any reconfiguration steps needed for the system to successfully execute queries against any new database. Extra mapping reconfigurations, user preferred ways to refer to elements of the domain-model, can be done through easy UI gestures such as right-clicking elements (i.e. words) of a given query. We believe that the simplification of the reconfiguration process when connecting to new database schemas is a very important step towards the democratization of NLIDBs in real world setup, as it enables non-technical users to be able to fully control the system through its entire lifecycle.

In addition, it enables the construction of a customized textual query environment in which a set of query-authoring services can be provided to the user, to help authoring and disambiguating queries. These query-authoring services play a fundamental role in systems usability, making it possible to early detect query errors, as demonstrated in the evaluation section, where we observed that around the 80% of the queries that contained errors could be detected before they were actually translated into T-SQL, resulting in a more efficient, lower-latency, user-interactive system. The classification of these errors based on the parsing stage in which they are detected, as showed in the evaluation, gives us the possibility to selectively focus on improving the quality and functionality of query-authoring services at each stage of the parsing process, in order to maximize the investment in relation to the gain of the overall user experience.

Based on our very positive evaluation results for early error detection, thanks to the use of query-authoring services, as future work, we are trying to maximize this benefit by experimenting with new query-authoring services and improving the existing ones.

Anaphora resolution is an active research field in the space of NLIDBs; this capability enables users to have the possibility to dramatically abbreviate the number of words to be written when asking different questions about different aspects of the same entity, which will result, again, in another important usability shift for {AskMe*} [21].

6 References

1. S. Abiteboul, V. Hull, R. Viannu, Foundations of Database Systems, Addison Wesley, 1995
2. P. Cimiano, P. Haase, J. Heizmann, Porting natural language interfaces between domains: an experimental user study with the Orakel system, in: Intelligent User Interfaces, ACM, 2007.
3. A. Popescu, A. Armanasu, O. Etzioni, D. Ko, A. Yates, PRECISE on ATIS: Semantic Tractability and experimental results, 2004.
4. N. Stratica, L. Kosseim, B. C. Desai, Using Semantic Templates for a natural language interface to the CINDI virtual library, 2004.
5. M. Minock, C-PHRASE: A system for building robust natural language interfaces to databases, Data Knowl. Eng., 2009.
6. E. Brill, Transformation based error driven learning and natural language processing: A case study in part of speech tagging, Computational Linguistics 21 (4) (1995) 543–565.
7. D. Jurafsky, J. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition and Computational Linguistics, Prentice Hall, 2000.
8. C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

9. H. Santoso, S. Haw, Ziyad.T. Abdul-Mehdi, Ontology Extraction from Relational Database: Concept Hierarchy as Background Knowledge, Knowledge-Based Systems (2010).
10. D.D. Sleator, D. Temperley, Parsing English with a link grammar, in: Proceedings of the Third International Workshop on Parsing Technologies, 1993.
11. A. Aho, J. Ullman, The Theory of Parsing, Translation and Compiling, vol. 1, Prentice-Hall, Englewood Cliffs, NJ, 1972.
12. I. Androutsopoulos, G.D. Ritchie, Database interfaces, in: R. Dale, H. Moisl, H. Somers (Eds.), Handbook of Natural Language Processing, Marcel Dekker Inc., 2000, pp. 209–240.
13. L. Androutsopoulos, Natural Language Interfaces to Databases – An introduction, Journal of Natural Language Engineering (1995).
14. M. Watson, NLBean(tm) version 4: a natural language interface to databases, Available from: www.markwatson.com
15. G. Miller, WordNet: a lexical database for English, Communications of the ACM 38 (1) (1995) 39–41.
16. R. Bartolini, C. Caracciolo, E. Giovanetti, A. Lenci, S. Marchi, V. Pirrelli, C. Renso, L. Spinsanti, Creation and use of Lexicons and Ontologies for NL Interfaces to Databases, International Conference on Language Resources and Evaluation (2006).
17. Link Grammar Parser, http://www.link.cs.cmu.edu/link
18. T4 templates, http://msdn.microsoft.com/en-us/library/bb126445.aspx
19. T-SQL value formats, http://msdn.microsoft.com/en-us/library/bb630352.aspx
20. Bates, M., Boisen, S., and Makhoul, J., Developing an Evaluation Methodology for Spoken Language Systems, Speech and Natural Language Workshop, Hidden Valley, Pennsylvania, 102-108.
21. Vicedo, J. L., Ferrandez A., "Importance of Pronominal Anaphora resolution in Question Answering systems", Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL'2000, (2000).
22. AdventureWorks DB, http://msdn.microsoft.com/en-us/library/ms124659.aspx.
23. Northwind, http://msdn.microsoft.com/en-us/library/aa276825(SQL.80).aspx.
24. Hindle, D., "An analogical parser for restricted domains", 1992.
25. Zue, V., Glass, J., et al, "The MIT ATIS system", 1992.
26. Moore, R. C., Appelt, D. E., et al, "SRI's experience with the ATIS evaluation", 1990.
27. He and S. Young, "A data-driven spoken language understanding system", IEEE Workshop on Automatic Speech Recognition and Understanding, 2003.

Source Language Generation from Pictures for Machine Translation on Mobile Devices

Andrew Finch[1], Wei Song[2], Kumiko Tanana-Ishii[2], and Eiichiro Sumita[1]

[1] NICT, Keihanna Science City, Kyoto, 619-0289, Japan.
{andrew.finch,eiichiro.sumita}@nict.go.jp
http://kccc.nict.go.jp
[2] University of Tokyo, Tokyo, 101-0021, Japan.
{song@cl.ci.,kumiko@}i.u-tokyo.ac.jp
http://www.cl.ci.i.u-tokyo.ac.jp

Abstract. This paper proposes a new method for the generation of natural language from a sequence of pictures and shows its application to machine translation within the framework of the picoTrans system. The picoTrans system is an icon-driven user interface for machine translation that facilitates cross-lingual communication through two heterogeneous channels of communication simultaneously. The first channel being the usual automatic natural language translation method; the second channel being a sequence of pictures that both parties understand which conveys structured semantic information in parallel with the first channel. Users are able to communicate using this device both by using it as a picture book and also by using it as a machine translation device. By pointing at pictures alone, basic expressions can often be communicated, eliminating the need for machine translation altogether, and even with machine translation, the picture sequence provides a useful second opinion on the translation that helps to mitigate machine translation errors. There are limits, however to the expressiveness of a sequence of pictures compared to the expressiveness of natural language. This paper looks at two methods by which syntactic information can be added into a sequence of pictures: a hidden n-gram model and monotonic transduction using a phrase-based statistical machine translation system. This additional information is added automatically, but the system allows the user to interact to refine the generated language. We evaluate both methods on the task of source sentence generation in Japanese using automatic machine translation evaluation metrics, and find the statistical machine translation method to be the more effective technique.

Keywords: user interface, machine translation, mobile devices

1 Introduction

Recently there has been a huge increase in the demand for machine translation (MT) services, as the translation quality for many language pairs has improved to levels that are of practical use. One common platform for applications of machine translation is mobile devices, since they can be used wherever they are

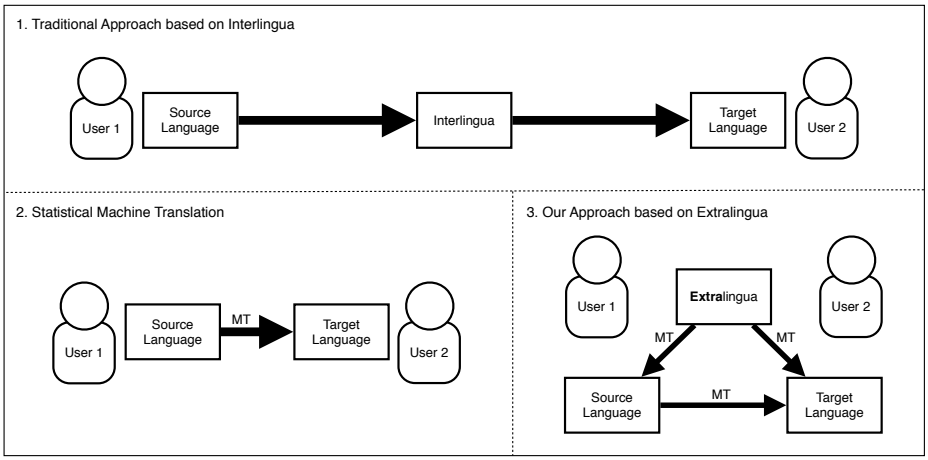


Fig. 1. Various Translation Channels Among Communicators Using Machine Translation

needed. The popularity of MT systems has cast light on the marginal problems of MT other than its translation quality. A major issue facing real world MT applications is the sheer variety of input that user could feed into an MT system. For example the input may contain mistakes, irregular or ungrammatical sentences, abbreviated words, smileys and contractions of words (such as writing "u" instead of "you"). These phenomena will degrade the performance of an MT system, but are essentially issues external to the core problem the MT system has to tackle, which is already a hard problem.

One way to address these issues would be to collect a large corpus consisting of irregular usages together with the corresponding correct usages, and learn to regularize the language in a supervised fashion. However, this approach is obviously limited, since schemes through which real users irregularize can depend on the user's communication mode or code, context, and even on his/her wit.

The picoTrans system [1] offers a simple solution: adopting an *extralingua* to assist MT. We have chosen the term *extralingua* deliberately due to its relationship to the term *interlingua*, and we expand on this later in the paper. Figure 1, shows some methods proposed for accomplishing translation. In part (1) of the figure, translation is performed through an interlingua, an intermediate language placed in between the source and the target language. When the interlingua is a natural language, the communication channel can be a concatenation of two MT systems: the first from source to interlingua, the second from interlingua to target [2]. Part (2) of the figure shows a process of direct translation from source to target. This can be achieved using widely-studied, state-of-the-art statistical machine translation systems.

In contrast, our approach (3) in Figure 1 uses an *extralingua*, which is exposed to both communicators. Both users are able to interact with the extralingua, assisted by three MT systems: the first between the extralingua to the source

language, the second between the source language and the target language, and the third between the extralingua and the target language. The reader might wonder why MT is needed at all if such an extralingua exists. This is in fact the point: the communicators lack in a common language through which they can communicate, and so far we have only considered ways to bridge this gap by using just a single MT channel. However, under many circumstances, the communicators do have other means for communication, such as images, signs and actions and will often use them when other means fail. This other mode of communication can be adopted in parallel independently of the MT channel, but our idea is to investigate the tight coupling of a second communication channel directly into a machine translation system.

The basic premise of our user interface, that sequences of images can convey a meaningful amount of information is directly supported by the findings of a number of studies. In [3], the effectiveness of using pictures to communicate simple sentences across language barriers is assessed. Using human adequacy scores as a measure, they found that around 76% of the information could be transferred using only a pictorial representation. In language generation [4] explore the possibility of communication by means of concepts. In assistive communication, the Talking Mats project [5], has developed a communication framework consisting of sets of pictures attached to mats to enable people with communication difficulties to communicate. There are other related systems and ideas based on the principle of using pictorial communication as a linguistic aid [6, 7]. In [8], a method for transforming icons into speech is proposed to aid people with Cerebral Palsy in communication.

In research into collaborative translation by monolingual users, [9] propose an iterative translation scheme where users search for images or weblinks that can be used to annotate sections of text to make its meaning more explicit to another user who does not share the same language. In other related work, [10], demonstrate the usefulness of a text-to-picture transduction process (essentially the converse of our icon-to-text generation process) as a way of automatically expressing the gist of some text in the form of images.

In our approach, we adopt icon sequences as the primary mode of input. There are multiple advantages to doing this. First and above all is to improve the quality of communication between users. Adopting an extralingua allows the users to communicate via two heterogeneous channels. Since we cannot expect MT output to be perfect, having a second independent mode of communication to reinforce or contradict will lead to a greater mutual understanding. Secondly, we believe that by constraining the user's input it should be possible to improve the MT quality since the input becomes regularized as a consequence, reducing variance in the input sequence and also decreasing the number of unexpected entries. This idea is supported by a study showing that normalizing language using a paraphraser can lead to improvements in translation quality [11].

A simple example of using a picture-book to communicate is illustrated in Figure 2. Suppose a user wished to translate the expression 'I want to go to the restaurant'; with a picture book, the user might point at 2 pictures: 'I want to to go to ∼', and 'restaurant'. A similar scenario for the picoTrans system is shown in Figure 3.

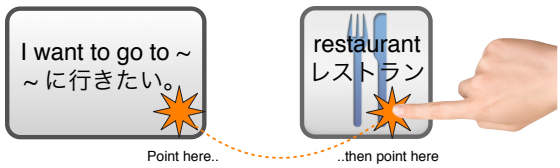


Fig. 2. The process of communication by using a picture-based translation aid for an English person. The order in which the pictures are pointed at is important, and thus if a Japanese person wished to convey the same information, the pointing order would be reversed (a Japanese user would first point to the 'restaurant' icon and then to the 'I want to go to' icon). This directly reflects differences in word order between the languages.

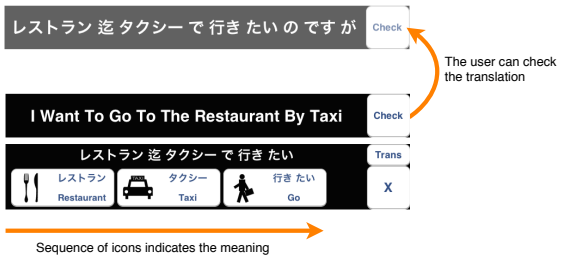


Fig. 3. The process of communication by using the picoTrans system. The sentence is a little more complex than that in Figure 2 to illustrate the additional expressive power of our approach. The picoTrans system displays the icon sequence, together with a translation of the user's intended meaning. This translation can be checked in the target language by referring to the icon sequence, and in the source language by referring to a back-translation.

In the next Section we describe our prototype system picoTrans; for a more complete description of the interface and operation of this system the reader is referred to [1]. The following section addresses the issue of natural language generation from an icon sequence, and describes the two approaches we have studied. Then we present an extension of the experiments reported in [1] that measure the expressiveness of the icon-based input approach on a limited domain, and also present our evaluation of the source generation techniques. Finally we conclude and offer some avenues for further research.

2 The picoTrans System

Picture-based translation-aids have been used in paper book forms and are currently integrated into hand-held devices but remain uncombined with machine translation systems. Briefly, in our proposed system picoTrans, the user taps picture icons appearing on the touch-screen, just like in a picture-based translation-aid. The system automatically generates the possible sentences from those selected icons, and feeds them to the machine translation in order that it

can display the translated result. Unlike a picture book, the sequence of icons is maintained on the display for the users to see, and interact with if necessary. When the input is complete, the system generates the full sentence in the source language automatically, which is then translated by the machine translation software and displayed on the screen together with the icon sequence. The user interaction is made through an interface which is currently implemented as a prototype working on the Apple iPad mobile tablet, although we believe our interface is applicable to smaller devices with touch screens such as mobile phones.

When communicating through our user interface, the user may combine the pictures in considerably more combinations than is possible with a picture book designed with combinations from only within the same page spread of the book in mind, making the application more expressive than a book. The machine translation system can contribute a detailed and precise translation which is supported by the picture-based mode which not only provides a rapid method to communicate basic concepts but also gives a 'second opinion' on the machine transition output that catches machine translation errors and allows the users to retry the sentence, avoiding misunderstandings.

Note how such a system is advantageous when applied to MT on mobile devices; the user input on these mobile devices can be cumbersome in the case of textual input [12], or errorful in the case of speech input [13]. As a consequence some users prefer to use simpler, more dependable methods of cross-lingual communication such as the picture book translation aids which are becoming increasingly popular both in paper form, and also in the form of electronic translation aid applications.

There are various applications available for hand-held devices in terms of either picture-based or machine translation, but none of them adopt both. In the former area, PictTrans [14] only shows picture icons, Yubisashi [15] (meaning *finger-pointing*) plays a spoken audio sound when tapping the icons, but these systems do nothing in terms of language generation which is delegated to the human users. Conversely, there are a substantial number of MT systems proposed for hand-held devices, for example the texTra [16] text translation system and the voiceTra [17] speech translation system, but as far as we are aware, none of them adopt an icon-driven user input system.

2.1 User Interface

A diagram of the full user interface for the picoTrans prototype system is shown in Figure 4. In brief, we allow the user to input what they wish to express as a sequence of bi-lingually annotated icons. This is in essence the same idea as the picture-book. Users can switch the user interface into their own language by pressing the User Interface Language Toggle Button (⑫ in Figure 4). The translation process proceeds as follows:

(1) The user selects a category for the concept they wish to express ⑪
(2) The user selects a sub-category ⑦
(3) The user chooses an icon ⑨, which is appended to the icon sequence ⑥

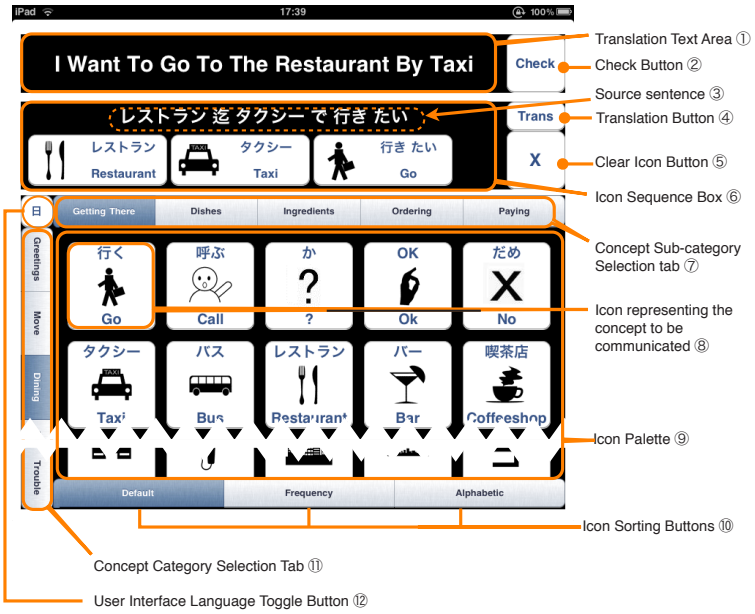


Fig. 4. The annotated user interface for the picoTrans system.

a) Go to (1) to select another icon for the sequence
b) The icon sequence is complete. Continue to step (4)

(4) The user interacts with the system to refine the source sentence (described in Section 3.4)
(5) The user clicks the 'Trans' button ④
(6) The translation appears in the translation text area ①

Once translation has completed, pressing the Check Button ② shows the back-translation of the target sentence into the source language for the user to verify the translation. Pressing the button again replaces the back-translation with the translation. The use of back-translation is somewhat controversial since the translation quality can be low and errors may confuse users, but in our system this is mitigated by the high translation quality of our restricted-domain system.

3 Source Sentence Generation

3.1 Language model approach

In previous work, conducted on Japanese input [1], the icon sequence was transformed into natural language by using a simple language-model-based approach to restore the missing function words. This approach was well-suited to the Japanese language which is quite regular in form, uses particles adjacent to content words to indicate their function, and contains no determiners. Bi-grams

containing pairs of content words with function words attached to either left or right were extracted from training data, and these bigrams were inserted in place of their corresponding content words in the generation process. The model proposed in [1] used a 5-gram language model to score the set of hypotheses resulting from all possible such substitutions, and selected the hypotheses with highest language model score as the best candidate. A beam search was employed to reduce the search space to a manageable size. While this approach proved very effective for Japanese where the simple addition of particles can, for many simple sentences, produce a good Japanese sentence from a sequence of content words representing the icons being input, we believe in order to generate other less suitable natural languages from sequences of icons, a more general approach is necessary. For this we turn to statistical machine translation.

3.2 Machine Translation

The task of transforming our icon sequence into the full source sentence is quite similar to the task of transliteration generation which can be performed using a phrase-based statistical machine translation system (SMT) using a monotonic constraint on the word re-ordering process [18, 19]. We adopt a similar approach, but use a Bayesian co-segmentation technique (explained in the next section) to derive the phrase table for the SMT system.

In order to train our SMT system, we generate a training corpus by means of word deletion. In our experiments we used Japanese as the source language and we analyzed our corpus using the publicly available MeCab [20] morphological analysis tool. A set of POS tags representing the classes of content words that would be represented by icons in our system (for example nouns, verbs, adjectives etc.) was compiled by hand, and the remaining classes of words (particles, dependent nouns and auxiliary verbs) were deleted from the source sentence. Furthermore, all inflected lexemes were reduced to their lemmata. The result of this process was a bilingual corpus, consisting of a sequence of content words (in lemma form) on the source side representing the icon sequence, and the full source sentence on the target side (see Figure 5).

For our experiments we used a phrase-based machine translation decoder closely related to the MOSES [21] decoder, integrating our models within a log-linear framework [22]. Phrase-pair discovery and extraction was performed using a Bayesian bilingual aligner [23]. A 5-gram language model built with Knesser-Ney smoothing was used. The systems were trained in a standard manner, using a minimum error-rate training (MERT) procedure [24] with respect to the BLEU score [25] on the held-out development data to optimize the log-linear model weights.

The machine translation systems were trained on approximately 700,000 bilingual sentence pairs comprised of the types of expressions typically found in travel phrase books. This is a very limited domain, and the sentences in this domain tend to be very short (on average 7-words in the English side of the corpus), making them very easy to translate. The machine translation system is a state-of-the-art system, and as a consequence of limiting the application to short sentences in a restricted domain it is capable of high quality translation.

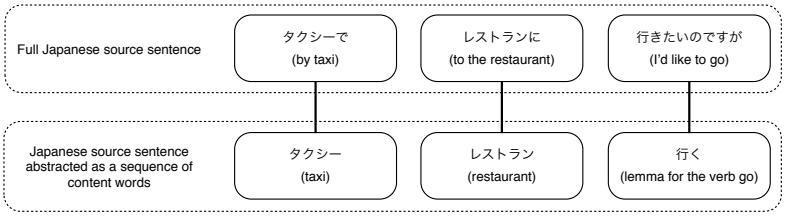


Fig. 5. Co-segmentation of a sequence of content words representing icons in our system, with the corresponding full Japanese sentence that they represent.

3.3 Bayesian Co-segmentation

At the core of all phrase-based statistical machine translation systems is the phrase-table. This table is the basic set of building blocks that are used to construct the translation.

The creation of a phrase-table during a typical training procedure for a phrase-based SMT system consists of word alignment step, often using GIZA++ [26], followed by a phrase-pair extraction step using heuristics (for example *grow-diag-final-and* from the MOSES [21] toolkit). This approach works very well in practice, but is asymmetrical with respect to source and target and is based on maximum likelihood methods that tend to over-fit the data.

The model we use for co-segmentation is based on a Dirichlet process model, similar to approach of [23]. We use a Bayesian approach here not only because results show that this approach is more effective on monotonically alignable sequences than using GIZA++/MOSES heuristics [23], but also because it results in a single self-consistent bilingual segmentation of the corpus. We believe this consistency is a very desirable characteristic for building our models since our system generates natural language simply by composing these phrase-pairs. This co-segmentation process is illustrated in Figure 5.

3.4 User Interaction

The output from both methods of source language generation are search graphs that represent the process by which the source sentence was constructed. This graph is provided by the MT system or n-gram model to the user interface client, which is able to use the information in the graph to guide the user to a satisfactory outcome without the need for continuous re-decoding of the input during the interaction process. This interaction process has not been developed very far in this work and remains an interesting area for future study. In related work [27], it has been demonstrated that an MT search graph can be used effectively in an interactive manner to assist human translation.

In our system, following the generation process from the icon sequence, the user is presented with the most probable hypothesis for the full source sentence given the input sequence. Should this sentence not convey the user's intended meaning, the user is able to interact with the icon sequence in order to refine the generated sentence. The user may tap on any icon in the sequence of icons

displayed on the interface. The user interface will consult the search graph and present the user with an n-best list of partial translation hypotheses up to and including the translation of the icon that was selected. At present we do not allow direct text entry into the system, although we do appreciate this would be possible and perhaps necessary in a real-world system, as our research is primarily concerned with exploring the possibilities arising from an icon-based approach to user input. The price to be paid by restricting the input in this manner is expressiveness, and we therefore examine our system empirically with this in mind in the following section, which is an extension of the experiments reported in [1].

4 Evaluation

4.1 Expressive Power

One of our main concerns about icon-driven user input was its expressive power within the domain, since sentences need to be expressed by only using icons that are available on the device. We therefore conducted an evaluation of the system to determine the proportion of in-domain sentences it was capable of representing. To do this we took a sample of 100 sentences from a set of held-out data drawn from the same sample as the training corpus, and determined whether it was possible to generate a semantically equivalent form of each sentence using the icon-driven interface and its source sentence generation process. The current version of the prototype has not been developed sufficiently to include sets of icons to deal with numerical expressions (prices, phone numbers, dates and times etc.), so we excluded sentences containing them from our evaluation set (the evaluation set size was 100 sentences after the exclusion of sentences containing numerical expressions). Handling numerical expressions is relatively straightforward however, and we do not foresee any difficulty in adding this functionality into our system in the future. The set of icons used in the evaluation corresponded to the most 2010 frequent content words in the English side of the training corpus, that is content words that occurred more than 28 times in the corpus. Thus value was chosen such that the number of icons in the user interface was around 2000, a rough estimate of the number of icons necessary to build a useful real-world application. We found that we were able to generate semantically equivalent sentences for 74% of the sentences in our evaluation data, this is shown in Figure 6 together with statistics (based on a 30-sentence random sample from the 100 evaluation sentences) for cases where fewer icons were used. We feel this is a high level of coverage given the simplifications that have been made to the user interface. A comparison of the two methods without human correction of the output are shown in Figure 7. We found that the MT method gave a higher level of coverage.

4.2 Quality of Source Sentence Generation

We measured the quality of the source language generation component of our system using version 13a of the NIST mteval scoring script in terms of the BLEU

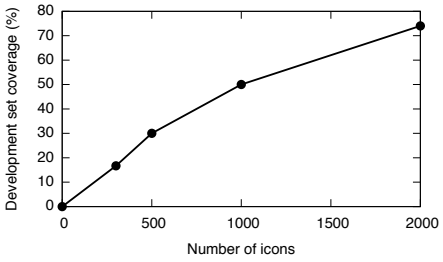


Fig. 6. The coverage of unseen data with icon set size, with human interaction.

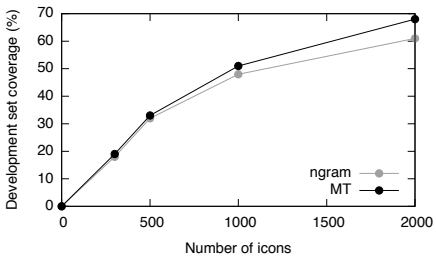


Fig. 7. The coverage of unseen data with icon set size, without human interaction.

score [25], and the NIST score [28], both common methods for measuring machine translation performance based on n-gram precision. We filtered sentences containing numerical expressions from our evaluation data; the initial set of 510 Japanese sentence was reduced to 455 sentences after filtering. We fed these sentences through the MeCab morphological analyzer and removed the words that will not be associated with icons in our systems. Our experimental results are shown in Table 1. The scores for the hidden n-gram (lemmatized) and SMT generation were derived from exactly the same input. The score of the hidden n-gram system is low in this case because it is unable to generate the inflected forms. Therefore, in a second experiment we allow the hidden n-gram model to generate from the correct surface forms of the inflected words. This gives the model an unfair advantage over the SMT generation model which needs to predict the inflection. Nonetheless, the source sentences generated by the SMT process score higher in terms of both of the evaluation metrics used. This combined with the fact that we expect it to be generally applicable to all languages, makes it unreservedly the better generation technique for our purposes.

Model	BLEU	NIST
Hidden n-gram (lemmatized)	0.27	4.16
Hidden n-gram (surface form)	0.66	8.60
SMT Generation	0.76	8.89

Table 1. Source sentence generation quality using an SMT approach relative to a hidden-ngram method to restore missing particles.

5 Conclusion

In this paper we have proposed a method of natural language generation from icon sequences based on a phase-based statistical machine translation system coupled with a Bayesian co-segmentation scheme that is used to perform icon-sequence to word-sequence alignment in order to train the translation model. We

have evaluated this method against an n-gram particle-insertion model used previously to generate Japanese natural language from icon sequences. Our results show that the SMT system is able to outperform the n-gram model in terms of both its coverage of the language and also in terms of the quality of the language generated measured using automatic machine translation metrics.

The picoTrans prototype opens up a wide range of possible directions for future research. In future work we plan to investigate the use of more informative abstract representations and the effects of tightly coupling them into the translation process, and also advance our system from a user interface perspective. We would also like to look at the possible ways in which users of the system might interact, with each other bilingually through the system, and also with the system itself monolingually. We plan to enhance the icon selection component of the interface with a more sophisticated algorithm able to dynamically predict an optimal ordering of the icons presented to the user based on their likelihood of being the next choice given the current context of the dialog, geographical location, and the user's history of icon choices. Finally, we believe the icon sequence to natural language generation techniques we are developing may find application outside the domain of machine translation, and this is something we wish to explore in the future, for example in the field of assistive communication aids.

References

1. Song, W., Finch, A.M., Tanaka-Ishii, K., Sumita, E.: picotrans: an icon-driven user interface for machine translation on mobile devices. In: Proceedings of the 16th international conference on Intelligent user interfaces. IUI '11, New York, NY, USA, ACM (2011) 23–32
2. Paul, M., Yamamoto, H., Sumita, E., Nakamura, S.: On the importance of pivot language selection for statistical machine translation. In: HLT-NAACL (Short Papers). (2009) 221–224
3. Mihalcea, R., Leong, C.W.: Toward communicating simple sentences using pictorial representations. Machine Translation **22** (2008) 153–173
4. Zock, M., Sabatier, P., Jakubiec, L.: Message composition based on concepts and goals. International Journal of Speech Technology **11** (2008) 181–193
5. Murphy, J., Cameron, L.: The effectiveness of talking mats with people with intellectual disability. British Journal of Learning Disabilities **36** (2008) 232–241
6. Ader, M., Blache, P., Rauzy, S.: Overcoming communication difficulties: a platform for alternative communication. 4mes Journes Internationales 'L'Interface des Mondes Rels et Virtuels (2008)
7. Power, R., Power, R., Scott, D., Scott, D., Evans, R., Evans, R.: What you see is what you meant: direct knowledge editing with natural language feedback (1998)
8. Vaillant, P., Checler, M.: Intelligent voice prosthesis: Converting icons into natural language sentences. Computing Research Repository **abs/cmp-lg** (1995)
9. Hu, C., Bederson, B.B., Resnik, P.: Translation by iterative collaboration between monolingual users. In: Proceedings of Graphics Interface 2010. GI '10, Toronto, Ont., Canada, Canada, Canadian Information Processing Society (2010) 39–46
10. Zhu, X., Goldberg, A.B., Eldawy, M., Dyer, C.R., Strock, B.: A text-to-picture synthesis system for augmenting communication. In Proceedings of the 22nd International conference on Artificial intelligence **2** (2007) 1590–1595

11. Watanabe, T., Shimohata, M., Sumita, E.: Statistical machine translation on paraphrased corpora. In: Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands (2002)
12. MacKenzie, S., Tanaka-Ishii, K., eds.: Text Entry Systems — Accessibility, Mobility, Universality. Morgan Kaufmann (2007)
13. Suhm, B.: Empirical evaluation of interactive multimodal error correction. In: in IEEE Workshop on Speech recognition and understanding, IEEE, IEEE (1997) 583–590
14. picTrans: A simple picture-based translation system. 7Zillion (2010) http://www.7zillion.com/iPhone/PicTrans/.
15. Yubisashi: Yubisashi. Information Center Publishing (2010) Available in many languages, found at `http://www.yubisashi.com/free/t/iphone/`, visited in 2010, August.
16. TexTra: (Text Translator by NICT). NICT (2010) http://mastar.jp/translation/textra-en.html.
17. VoiceTra: (Voice Translator by NICT). NICT (2010) http://mastar.jp/translation/voicetra-en.html.
18. Finch, A., Sumita, E.: Phrase-based machine transliteration. In: Proc. 3rd International Joint Conference on NLP. Volume 1., Hyderabad, India (2008)
19. Rama, T., Gali, K.: Modeling machine transliteration as a phrase based statistical machine translation problem. In: NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, Morristown, NJ, USA, Association for Computational Linguistics (2009) 124–127
20. Kudo, T.: MeCab. [Online] Available: http://mecab.sourceforge.net/ (2008)
21. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowa, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: ACL 2007: proceedings of demo and poster sessions, Prague, Czeck Republic (2007) 177–180
22. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). (2002) 295–302
23. Finch, A., Sumita, E.: A bayesian model of bilingual segmentation for transliteration. In: In Proceedings of the IWSLT, Paris, France (2010)
24. Och, F.J.: Minimum error rate training for statistical machine translation. In: Proceedings of the ACL. (2003)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2001) 311–318
26. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29** (2003) 19–51
27. Koehn, P.: A web-based interactive computer aided translation tool. In: Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore, Association for Computational Linguistics (2009) 17–20
28. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Proceedings of the HLT Conference, San Diego, California (2002)