INTERNATIONAL WORKSHOP

MULTI-SOURCE, MULTILINGUAL INFORMATION EXTRACTION AND SUMMARIZATION

held in conjunction with the International Conference

RANLP - 2007, September 27-29, 2007, Borovets, Bulgaria

PROCEEDINGS

Edited by

Thierry Poibeau and Horacio Saggion

Borovets, Bulgaria

26th September 2007

International Workshop

MULTI-SOURCE, MULTILINGUAL INFORMATION EXTRACTION AND SUMMARIZATION

PROCEEDINGS

Borovets, Bulgaria
26th September 2007

ISBN 978-954-452-001-4

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

WORKSHOP ORGANISERS

Thierry Poibeau
(LIPN-CNRS, U. Paris 13 - France)

Horacio Saggion
(NLP Group, U. Sheffield - United Kingdom)

WORKSHOP PROGRAMME COMMITTEE

Sophia Ananiadou (U. Manchester, UK)
Roberto Basili (U. Roma Tor Vergata, Italy)
Kalina Bontcheva (U. Sheffield, UK)
Nathalie Colineau (CSIRO, Australia)
Nigel Collier (NII, Japan)
Hercules Dalianis (KTH/Stockholm University, Sweden)
Thierry Declerk (DFKI, Germany)
Brigitte Grau (LIMSI, France)
Kentaro Inui (NAIST, Japan)
Min-Yen Kan (National University of Singapore, Singapore)
Guy Lapalme (U. Montreal, Canada)
Diana Maynard (U. Sheffield, UK)
Jean-Luc Minel (CNRS-Modyco, France)
Constantin Orasan (University of Wolverhampton, UK)
Cecile Paris (CSIRO, Australia)
Agnes Sandor (Xerox XRCE, France)
Ralf Steinberger (European Commission - Joint Research Centre, Italy)
Stan Szpakowicz (University of Ottawa, Canada)
Lucy Vanderwende (Microsoft Research, USA)
Jose Luis Vicedo (Universidad de Alicante, Spain)
Roman Yangarber (University of Helsinki, Finland)
Liang Zhou (ISI, USA)
Michael Zock (LIF, France)

EDITORS' FOREWORD

Recent years have witnessed an explosion of information in textual form, making natural language processing technologies for information access particularly important for the information society. These technologies, however, face new challenges with the adoption of the Web 2.0 paradigm because of its inherent multi-source nature. These technologies have to deal no longer with isolated texts or single narratives but with large scale repositories, or sources -- in one or many languages -- containing a multiplicity of views, opinions, or commentaries on particular topics, entities or events. There is thus a need to adapt and/or develop new techniques to deal with these new phenomena.

Recognising similar information across different sources and/or in different languages is of paramount importance in this multi-source, multi-lingual context, in particular the ability to detect paraphrases in texts is relevant here. In *information extraction*, merging information from multiple sources can lead to increased accuracy relative to extraction from single sources. In *text summarization*, similar facts found across sources can inform sentence scoring algorithms. In *question answering*, the distribution of answers in similar contexts can inform answer ranking components. In occasions, it is not the similarity of information that matters, but its complementary nature. In a multi-lingual context, information extraction and text summarization can provide solutions for cross-lingual access: key pieces of information can be extracted from different texts in one or many languages, merged, and then conveyed in many natural languages in concise forms.

The objective of this *Multi-source Multilingual Information Extraction and Summarization* (MMIES) workshop is to bring together researchers and practitioners in information access technologies to discuss recent approaches to deal with multi-source and multi-lingual challenges.

Each paper submitted to the workshop was reviewed by three members of an international programme committee. The selection process resulted in this volume which contains six selected papers covering the following key topics:

- Mono-lingual and Multi-lingual Cross-document Coreference;
- Multi-source Information Extraction;
- Social Networks' Learning and Visualisation;
- Relation Extraction;
- Cross-lingual Information Extraction and Document Retrieval; and
- Ontology-based Information Extraction

We would like to thank the members of the International Programme Committee for their invaluable work. We are also grateful to Kiril Simov, Galia Angelova, and Ruslan Mitkov organizers of the conference Recent Advances in Natural Language Processing (RANLP 2007) for their support.

Our gratitude also goes to Bernardo Magnini for accepting to give an Invited Talk in our workshop.

September 2007.

Horacio Saggion and Thierry Poibeau

TABLE OF CONTENTS

Disambiguation of Standardized Personal Name Variants

Patricia Driscoll
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
pdriscoll@gmail.com

David Yarowsky
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
yarowsky@cs.jhu.edu

Abstract

A growing body of research addresses name disambiguation as part of coreference and entity resolution systems, but the systems do not robustly resolve the ambiguity introduced by standardized personal name variants, or nicknames. In many languages, personal name variants are governed by morphological and phonological constraints, providing a dataset rich in features which may be used to train and run matching systems. We present a supervised learning method to address the problem of standardized personal name variant matching in English. The system integrates information from multiple sources into a weighted voting model which significantly outperforms baseline methods.

Keywords

Nicknames, Name Disambiguation, Name-Matching, Personal Name Variants, Truncation, Morphology

1 Introduction

As data sources such as the Internet expand in size, the study of entity disambiguation, whose goal is to cluster large numbers of name mentions according to entity referents, has become increasingly important. As a crucial part of this process, personal name disambiguation aims to create linguistically motivated links between personal names using information about their structure and morphology which can be mined from multiple sources. Personal name disambiguation has been studied both for its place in the entity disambiguation process, where it can aid tasks like coreference resolution [1, 11, 13], and for the production of stand-alone tools and name disambiguation resources, such as proper name ontologies [9, 8], onomastica [15], and fuzzy name-matching tools [12] which accept candidate pairs as input. An important area of current research, name disambiguation systems have the capability to take into account social and occupational titles, honorifics, and variation in capitalization and punctuation [18]. Problems similar to personal name variant disambiguation, including name transliteration and cognate matching for common nouns, have also been studied in the context of machine translation [7, 10, 16].

Personal name variants, such as standardized nicknames of personal names, have been little-studied elements of the name disambiguation problem. Those systems that include standardized nicknames as equivalent to their corresponding full forms typically do so using a pre-packaged dataset such as a nickname pair list [6], or by simple string-matching methods which do not take into consideration the morphological relationship between standardized nicknames and their corresponding full forms, leaving the systems susceptible to error [12]. In this paper, we address the task of scoring arbitrary pairs of names and nicknames, creating a module to aid name disambiguation and overcome problems presented by language change, incomplete datasets, and scarce resources. By robustly extending the pool of potentially coreferent personal names, this work will enhance the recall of state-of-

the-art entity disambiguation systems.

2 Personal Name Variants

The terms 'nickname' and 'hypocoristic', are commonly used to refer to several distinct phenomena when describing personal names. One class of nicknames are pet names which are related to personal or relationship traits, but are generally linguistically unrelated to the full form name (Elvis Presley→The King). Non-standardized nicknames, while often related to long forms, are typically used to refer to one person in particular (Richard Nixon→Tricky Dick). Since links between full forms and familiar forms in these cases relate to entities rather than the names themselves, they are not able to be generalized for use in name-matching systems.

For the purposes of this paper, we will use 'nicknames' to refer to the set of standardized familiar form variants of personal names. Such familiar forms are linguistically linked to full forms, although links are governed by a combination of morphological and phonological constraints and convention that can range from highly regular (Christina→Chris) to relatively opaque (John→Jack). Name dictionaries linking these standardized familiar forms are not typically available electronically, and where available are often incomplete. Further complicating the picture is the idea that such familiar forms are somewhat productive, dynamic aspects of language for which it may be difficult to limit tasks to use of static resources.

Variation in nicknames is common, with sociological trends, idiosyncracies, and the desire to distinguish different individuals with the same name cited as some of the sources of variability [2, 17, 5]. Additionally, linguistic patterns for nickname formation are complex, governed by morphological and phonological constraints with diverse ordering conventions [17, 2]. Despite this variation, many languages do share common nickname pattern characteristics, the most commonly discussed of which are truncation, reduplication, and augmentation (See Table 1).

Because of the complexity of nicknaming patterns, handwritten rules for personal name variant matching are both time-consuming and incomplete. In this paper, we explore a variety of learning methods for building personal name variant resources and for doing matching tasks dynamically. Evaluation data for this domain is limited: name variant resources are scarce and incomplete in English and virtually nonexistent in many other languages. This data scarcity further attests to the utility of the automatic acquisition of usable resources.

Language	Full Form	Nickname
French	Paul	Paulo
	Jacques	Jacquot
	Raphaelle	Raphie
	Louis	Loulou
	Marguerite	Margot
Russian	Svetlana	Sveta
	Polina	Polya
	Nataliya	Natasha
	Yevengenya	Zhenya
Italian	Giovanna	Gianna
	Luigi	Gino
	Rosalia	Lietta
	Guiseppe	Pino

Table 1: *Nickname formation is highly structured, though each language has its own set of constraints.*

To obtain training and evaluation data, first names taken from 1990 U.S. Census data were used to query a nickname database at www.oxygen.com/babynamer, which yielded a set of 2543 name-nickname pairs using 907 of the census names. 1837 nicknames were represented in the data, which often included multiple nicknames for particular first names (Jennifer→ Jen, Jenny) as well as nicknames which were associated with multiple first names (Robert, Roberto →Bob). Although it contained many name variants, the resource was not exhaustive, thus presenting challenges for evaluation. One example of the lack of coverage occurred with spelling variants of full form names, in which nicknames were often linked to one form but not another.

An interesting property of the data was that more common first names (using probabilities given by census.gov) were likely to have more nicknames than their less common counterparts (Figure 1).

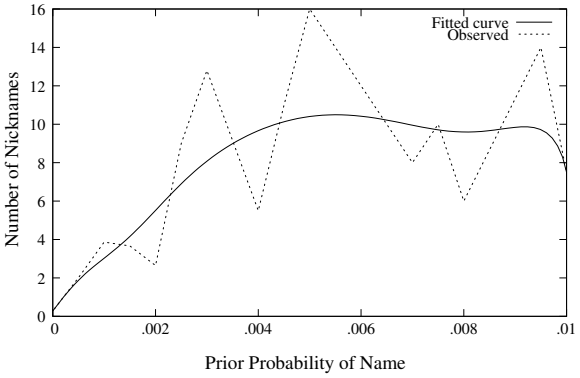

Fig. 1: *More common names have more nicknames.*

3 Detecting Name Variant Pairs

To link standardized personal name variants with corresponding full form names, several methods were chosen for strengths in accuracy, flexibility, and limited data and annotation requirements. Used both individually and in combination, the methods chosen are likely to work well in a variety of settings, including quick ramp-up for languages with little available data.

3.1 Web-Based Extraction

Although there are few languages for which entire name variant dictionaries are available for download from the Internet, the Internet can nonetheless be a reliable tool to use in the creation of such resources. Personal web pages, increasingly available in many languages, are rich in name information which can fill in gaps left by other systems, mediating recall limitations that may be addressed through the addition of alternate methods. In particular, web extraction often covers name pairs which are not related by simple morphological rules, providing knowledge about name variant matches that would be difficult to access otherwise. Findings of the web extraction method may also be used to give a boost to correct pairs which are found in the results of other methods, but are ranked below erroneous pairs.

There has been much recent work on web-based extraction, in which systems typically start with a hand-picked seed phrase [4] or seed instances [3]. For the web extraction component, we started from the seed phrase "My name is *full form* ... friends call me *variant*", issued this query to Internet search engine yahoo.com, and collected the first page of results for each seed nickname. This weakly supervised method ensures quick and easy extension to other languages. More strongly supervised methods with added alternate seed phrases and use of results beyond the one-page range would likely improve the performance of this component in English, both by eliminating erroneous hits and by expanding coverage.

Seed Nickname	Candidate Full Form
Katia	Katarina
Lynn	Madeline
Debbie	Phyllis
Lee	My

Table 2: *Web extraction component finds nicknames in a variety of contexts.*

Many of the recovered full form/variant pairs retrieved are complex but correct (Table 2). Because of the potential unreliability of this extraction method, candidate full forms that did not appear in the census data (Lee←My) were discarded. Another source of error was the fact that the "my name is" match immediately preceding the seed nickname on a particular web page was not always relevant: in some cases, long lists of online personal ads included phrases like "my name is" and "friends call me" as interchangeable (Debbie←Phyllis).

3.2 Morphological Analyzer

To exploit the feature-rich, highly constrained morphological derivation process involved in nickname formation, we used a toolkit as described in [19]. The toolkit uses the Word-Frame model which learns string transduction between inflected and root forms. It has previously been applied primarily to learning verb inflections, and its application to learning nickname inflection patterns is novel. The model itself is flexible, and can model arbitrary affixation morphological processes (prefixation, suffixation, and infixation). While it has no explicit support for processes like vowel harmony and partial word reduplication, [19] asserts that these processes can often be modeled satisfactorily by affixation rules.

Training data consisted of 1000 name-

Morphological Rule	Nickname	Full Form
IE → A	Elsie	Elsa
IE → ERTA	Albie	Alberta
EE → INA	Rosalee	Rosalina
EE → ENE	Charlee	Charlene

Table 3: *The morphological analyzer learns the common morphological nickname inflections from supervised training data.*

1. Exact nickname matches to the beginning or the end of full form.
2. Exact nicname matches anywhere in full form.
3. Lemmatized nickname matches at the beginning or the end of full form.
4. Lemmatized nickname matches anywhere in the full form.

Table 4: *The handwritten truncation rules give four levels of matches.*

nickname pairs. We additionally gave as input to the system a list of 10 suffixes, input which [19] has suggested can improve performance of the analyzer. Table 3 gives examples of some of the rules learned by the morphological analyzer.

3.3 Truncation Rules

In addition, a small set of handwritten truncation rules was used to supplement results from the weakly supervised components. The rules were developed using basic knowledge of left and right truncation with vowel-only augmentation, so as to exploit features requiring only limited knowledge and limited implementation time. Since truncation is a nickname formation mechanism found in many languages, similar rules might be written for other languages. As mentioned above, nickname formation is a complex phenomenon and attempts to provide hardcoded rules for all languages are too limited to capture all of the requisite variation.

These rules (Table 4) provide a rough cut at matching, and were used both to form a baseline and as a supplement to the system, since they provided information on the most likely guesses when highly regular matches existed.

Using the rules above, nicknames using basic truncation such as Elizabeth →Liz, Lizzie, or Beth would be recognized, while more complex forms (Elizabeth→Betsy) would not. The handwritten rules did little to constrain the truncation by pruning out unlikely examples, so Elizabeth would also be matched with candidate nicknames like Eli, Zoe, and Bea. Although highly regular, simple phonological changes were not modeled to limit size of rule set and allow for extensions to other languages.

4 Classifier Combination

Weights were trained using 145 pairs with names appearing in the top ten percent of census data. Since both the web extraction and morphological analysis components were generative and thus likely to be relatively sparse, we included a binary feature indicating whether the candidate appeared on each of these lists, regardless of score. Based on data indicating that number of nicknames varies according to full form prior (see Figure 1), these prior probabilities were also included as a candidate feature. Because Levenshtein distance, with a score cutoff of 4, gave no improved performance, it was not included in the final feature set.

The final weights chosen by the system were as follows:

Feature	Weight
Appears in the Web Extraction List	0.286
Handwritten Truncation Rule Score	0.214
Morphological Analyzer Score	0.143
Web Extraction Ranking	0.143
Full Form Prior Probability	0.143
Appears in Morphological Analyzer List	0.071

5 Experimental Results

Our initial data set of 2543 name-nickname pairs was split into three portions: test data, development data, and training for the morphological analyzer. Names from the top ten percent of the census data were used to create pairs for the test and development sets. Test

Nickname	**Confusors** (distance)	**True Full Form** (distance)
Pammie	Mammie (1) Tommie (2)	Pamela (3)
Jess	Bess (1) Jose (2)	Jessica (3)
Lenny	Benny (1) Wendy (2)	Leonard (6)

Table 5: *Levenshtein distance typically judges unrelated names to be closer than full form names.*

data using 278 pairs of these commonly occurring names and their nicknames was then chosen at random, and was expanded to include all correct names for nicknames seen in testing. Development data using 145 pairs was chosen at random from the remains of this set, and training data of 1000 pairs was selected using only nicknames not seen in test or development.

Baselines were a simple substring match ranked by proportion nickname comprised of full form, Levenshtein distance with a cutoff of 4, and the handwritten truncation rules described in Section 3. Table 5 gives examples which demonstrate why Levenshtein distance is a unsuitable approximation to the complex morphological processes involved in name variant formation. [1]

The dataset, although relatively formal and thorough, did not escape some of the issues inherent in the use of static datasets for standardized name variant matching. One particularly challenging obstacle was a seeming lack of full recall in the test set. A type of legitimate match which was often not given credit in the test set was seen with spelling variants of other full form names for which more extensive pair lists existed (e.g. Deborah → Debbie was in the test data, but Debra → Debbie was not). The inclusion of relatively obscure matches also made the problem of scoring a difficult one: if systems are expected to include matches such as Daisy ← Marguerite before being granted full credit, recall scores will be unrealistically low. To address these issues, components were built to produce ranked lists and scored by means of precision/recall curves. This was the most straightforward approach for integrating and evaluating information from a variety of heterogeneous sources, as some components produced reliable ranked lists, while others were more useful as isolated features into the system. Ongoing work is looking into combining these features into a supervised learning system such as a logistic regression model.

Table 6 shows examples of the kind of matches discovered by each of the system components. Figure 2 shows the performance of the components in isolation as well as the combined system performance. The combined system performance yields significantly superior performance to the baseline systems. It is able to combine the high precision/low recall performance of the web extraction component and the morphological analyzer with the high recall/low precision performance of the truncation rule component.

The weighted voting method, using features described above, outperformed all other methods as recall climbed above 15%. Web extraction and morphological analysis components were competitive at lower recall, indicating that these components were able to find certain pairs with good confidence. Truncation rules, substring, and Levenshtein methods were not strong on precision, although truncation rules remained steady as recall increased, indicating their strength as a component which selects many correct matches among noisy pair results.

6 Conclusions and Future Work

The results reported on matching standardized name variants with full form personal names are indicative of both capabilities in English and, because of the flexibility of the methods used, point to the likelihood of success using the methods described in languages with similar nicknaming conventions. In particular, the use of a trained morphological analyzer requires little data and supervision, and can serve as a stand-alone tool for name-

[1] Alternative noncontextual learned edit distance measures such as [14] would suffer from problems similar to those seen in traditional Levenshtein distance. Instead we look to the morphological analyzer presented in Section 3.2 as a linguistically motivated modeling tool.

Nickname	Full Form	Top Component Choices		
		Truncation Rules	Morphological Analyzer	Web Extraction
Steve	**Stephen** **Steven**	1. Stevie 2. **Steven**	1. Steve 2. Stevie	**1. Stephen** **2. Steven**
Vikie	**Victoria**	1. Vikki 2. Viki	**1. Victoria** 2. Viki	*none*
Chas	**Charles** **Chastity**	1. **Chastity**	*none*	**Charles**

Table 6: *Each of the different components captures an important aspect to the nickname process: the truncation rule component allows high recall, the morphological analysis component allows more flexible matches than simple truncation rules, and the web extraction component is a high precision matcher.*

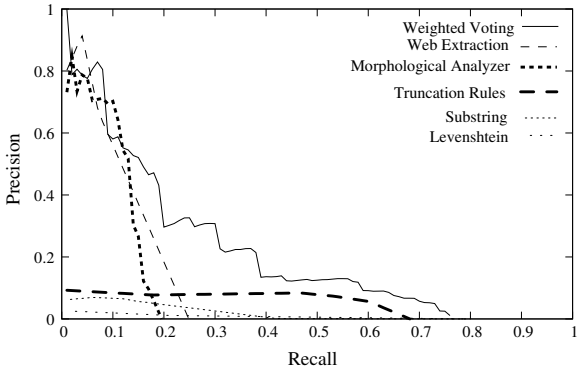

Fig. 2: *The combined system is able to take advantage of the benefits of each of the components: the high precision of the web extractor and the morphological analyzer, and the high recall of the truncation rules.*

matching with limited seed data. Web extraction methods can fill in the gaps where linguistically-motivated methods leave off, and future work which expands the seed phrases used will likely improve on the success shown here.

Because extensive name variant resources are not widely available, expensive to compile, and susceptible to change, the methods described above provide the opportunity to create such resources with limited data, supervision, and implementation time. The methods presented are dynamic and flexible and can be re-trained when data changes, and are rank-based, avoiding recall drop-offs likely in static dictionaries. Improvements in entity disambiguation via higher recall will be achieved by this more dynamic and flexible approach.

We have presented here a general paradigm for learning name variant models. Future work will explore alternative methods within this framework for learning variants, such as SVM classifiers and weighted Levenshtein distance.

In this paper, we do not explore the extension of the methods presented to other types of name variants, such as personal name spelling variants. Such work would likely have the creation of personal name equivalence-classes as a goal, which could then extend nickname results to all members of target classes.

While important, work in languages other than English has been inhibited by the difficulty of test data collection, which speaks further to the need for building reliable systems in these languages. While nickname formation in languages other than English has seen a limited amount of attention in the theoretical linguistics community, the extension of the full set of methods applied in this paper to a larger set of languages is a promising area of further research.

References

[1] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In C. Boitet and P. Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 79–85, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[2] L. Benua. Identity effects in morphological truncation. In J. Beckman, S. Urbanczyk, and L. Walsh, editors, *University of Massachusetts Occasional Pa-*

pers in Linguistics. University of Massachusetts, 1995.

[3] S. Brin. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, pages 172–183, 1998.

[4] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Web-scale information extraction in knowitall. In *WWW*, 2004.

[5] J. Ito and A. Mester. Sympathy theory and german truncations. In V. Miglio and B. Moreen, editors, *University of Maryland Working Papers in Linguistics*. University of Maryland, 1997.

[6] Z. Kazi and Y. Ravin. Who's who? identifying concepts and entities across multiple documents. In *International Conference on System Sciences*, 2000.

[7] K. Knight and J. Graehl. Machine transliteration. *Computational Linguistics*, 24(4):599-612, 1998.

[8] C. Krstev, V. Dusko, D. Maurel, and M. Tran. Multilingual ontology of proper names. In *Language and Technology Conference*, 2005.

[9] G. S. Mann. Building a proper noun ontology for question answering. In *Proceedings of SemaNet02: Building and Using Semantic Networks*, pages 16–22, 2002.

[10] G. S. Mann and D. Yarowsky. Multipath translation lexicon induction via bridge languages. In *NAACL*, 2001.

[11] G. S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the Conference on Natural Language Learning*, pages 33–40, 2003.

[12] G. Navarro, R. Baeze-Yates, and J. Arcoverde. Matchsimile: a flexible approximate matching tool for searching proper names. *JASIST*, 54(1), 2003.

[13] Y. Ravin and Z. Kazi. Is hillary rodham clinton the president? disambiguating names across documents. In *Proceedings of the ACL '99 Workshop on Coreference and its Applications*, 1999.

[14] E. Ristad and P. Yianilos. Learning string edit distance. *IEEE Trans. PAMI*, 20(5):522-532, 1998.

[15] S. Sheremetyeva, J. Cowie, S. Nirenburg, and R. Zajac. Multilingual onomasticon as a multipurpose nlp resource. In *LREC*, 1998.

[16] T. Sherif and G. Kondrak. Substring-based transliteration. In *ACL*, 2007.

[17] N. Topintzi. Prosodic patterns and the minimal word in the domain of greek truncated nicknames. In *International Conference of Greek Linguistics*, 2003.

[18] N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of proper names in text. In *Proceedings of Fifth Conference on Applied Natural Language Processing*, pages 202–208, 1997.

[19] R. Wicentowski. Multilingual noise-robust supervised morphological analysis using the wordframe model. In *ACL SIGPHON*, 2004.

Ontological Integration of Information Extracted from Multiple Sources

Adam Funk, Diana Maynard, Horacio Saggion and Kalina Bontcheva
Department of Computer Science
University of Sheffield
Regent Court
211 Portobello Street
Sheffield, S1 4DP, U.K.
{a.funk,d.maynard,h.saggion,k.bontcheva}@dcs.shef.ac.uk

Abstract

We describe here an ontologically based approach to multi-source, multilingual information extraction. Structured, semi-structured and unstructured documents of various types are annotated using a range of hand-crafted and machine-learning information extraction processes; the resulting annotations are used as statements to update a knowledge base for business intelligence. Our approach in particular uses domain-oriented ontologies that extend the *de facto* Proton standard to ensure compatibility between the extracted data so that they can be integrated into a consistent, precise set of results.

1 Introduction

Multi-source information extraction typically deals with the specific problem of cross-document coreferencing, i.e. determining which named entities in a set of documents have the same referents; as Bagga and Baldwin [2] point out, this problem differs significantly from coreference identification within individual documents, where we can expect more consistency and a smaller potential domain. Approaches to this coreferencing problem include vector space modelling on document contexts [2, 3], adaptation of a Context Thesaurus originally developed for query refinement in information retrieval [18], and shallow syntactic analysis of multi-word terms [10].

Systems aimed at the business domain include JV-FASTUS [1], which carried out shallow text analysis with results that were interesting but (naturally for MUC) based on template-completion without reference to a domain-related ontology, and the MBOI tool [8] for discovering information about business opportunites on the internet, which however requires specific semi-structured data sources.

A semantically enhanced system is h-TechSight [14, 15], which uses information extraction and retrieval with an ontology to monitor markets and detect trends and changes, e.g. for business intelligence about competitors' products in company reports and news articles or for employers and applicants to watch the employment market. Unlike the system we will present here, however, the ontology is quite small with a few fixed concepts.

As Maynard et al. [16] point out, however, existing systems that aim to extract information for business intelligence do not deal sufficiently with unstructured text input. We therefore aim to develop and combine tools for various input types so that we produce coherent, consistent output.

2 Background

In the MUSING[1] project, we wish to provide a new generation of versatile yet integrated tools for business intelligence using semantically enhanced information extraction and reasoning for three application areas:

- financial risk management, especially credit risk management concerning small and medium enterprises (SMEs);
- internationalization, i.e. identifying, capturing, representing and localizing knowledge in the context of global competition; and
- operational risk measurement for IT systems.

We are interested in making the best use of declarative and statistical information extraction techniques on a variety of documents with different degrees and types of structure and mixtures of

[1] http://www.musing.eu/

numeric and textual content, such as companies' web pages, articles from the financial press, government documents and corporate financial reports. We have therefore designed a high-level approach to multi-source information extraction, based on integrating the results of various information extraction tasks using semantic knowledge.

3 Methodology

This section describes our ontologically-based approach to the problem of integrating information extraction from diverse sources using various information extraction techniques.

3.1 Input

We wish to extract information from a variety of document types which present different problems and characteristics for information extraction.

News articles consist mainly of free natural language text, with some metadata from the provider's database as well as XML or HTML annotation. Companies' web pages (particularly the index, "contact us" and "about us" pages of each site examined) similarly consist of free text with varying degrees of HTML annotation, some of which (such as headings and URLs) can be particularly useful for information extraction.

Wikipedia[2] articles are also mostly free text, although parallel articles often have parallel structure and tabular data in regular formats (for example, each article about a country or region usually contains a fairly standardised table with figures for population, surface area, etc., and similar headings and natural-language expressions recur). The CIA World Factbook[3] has a much more consistent and therefore easy analysable format (but does not cover many regions within countries). Government documents also contain a wide variety of numeric and textual information in semi-structured and unstructured forms.

Balance sheets and other financial reports are now structured fairly consistently according to international accounting standards and can also be written in the emerging XBRL[4]. [9] However it is also useful to take advantage of NLP techniques to analyse the information in the free-text notes to these reports, which may significantly affect the interpretation of the easily analysable numeric parts.

It is also worth noting that to meet the needs of modern business intelligence we wish to take advantage of sources in various languages.

3.2 Extraction techniques

Our information extraction applications are based primarily on GATE[5], which provides a development environment, an architecture, a library of robust, adaptable tools for natural language processing (including machine-learning), and facilities for manual annotation of documents, and which is well-suited for multilingual information extraction. [4, 6]

These applications fall into two categories: *declarative* and *machine-learning*.

3.2.1 Declarative tools

Our declarative or hand-crafted applications are generally derived from GATE's standard information extraction system, ANNIE [13], which provides standard NLP tools (tokeniser, sentence splitter, POS (part-of-speech) tagger, lemmatiser) as well as some gazetteers and JAPE[6] grammars for general-purpose information extraction. ANNIE already has very good performance (F-measure 92.9%) for traditional information extraction on general news texts [16] and is therefore a good base to build on.

For this purpose, we add gazetteers of key words and phrases found in the documents and JAPE grammars to detect patterns and annotate the information desired by our representative users in the project. Table 1 lists a small sample of the datatypes requested for commercially evaluating different countries and regions. We are developing several applications along these lines to deal with the various input document types.

Tabular data can be analysed with gazetteers and JAPE grammars designed to identify the row and column headings and boundaries and to annotate the statistics accordingly. JAPE rules can also take advantage of a document's "original markups" such as HTML or XML tags, and therefore treat headings differently from paragraphs, for example.

3.2.2 Machine learning

Users are also manually annotating documents using GATE's Ontology-based Corpus Annotation

[2] `http://www.wikipedia.org/`

[3] `https://www.cia.gov/library/publications/the-world-factbook/index.html`

[4] eXtensible Business Reporting Language `http://www.xbrl.org/`

[5] `http://gate.ac.uk/`

[6] JAPE is an engine for pattern-matching over annotations on GATE documents and adding further annotations or executing arbitrary Java code. [5]

Class	Short name	Full name
LabourAvailabilityIndicator	EMP	Employment rate
	WAGE	Minimum wage
MarketSizeIndicator	RUR	Rural population (%)
	LRT	Literacy rate total (%)
	DENS	Population density
ResourceIndicator	FOREST	Forest area (%)
	RFOREST	Reserved forest (sq km)
	AGRIC-LAND	Agricultural land (%)

Table 1: *Example datatypes for region evaluation*

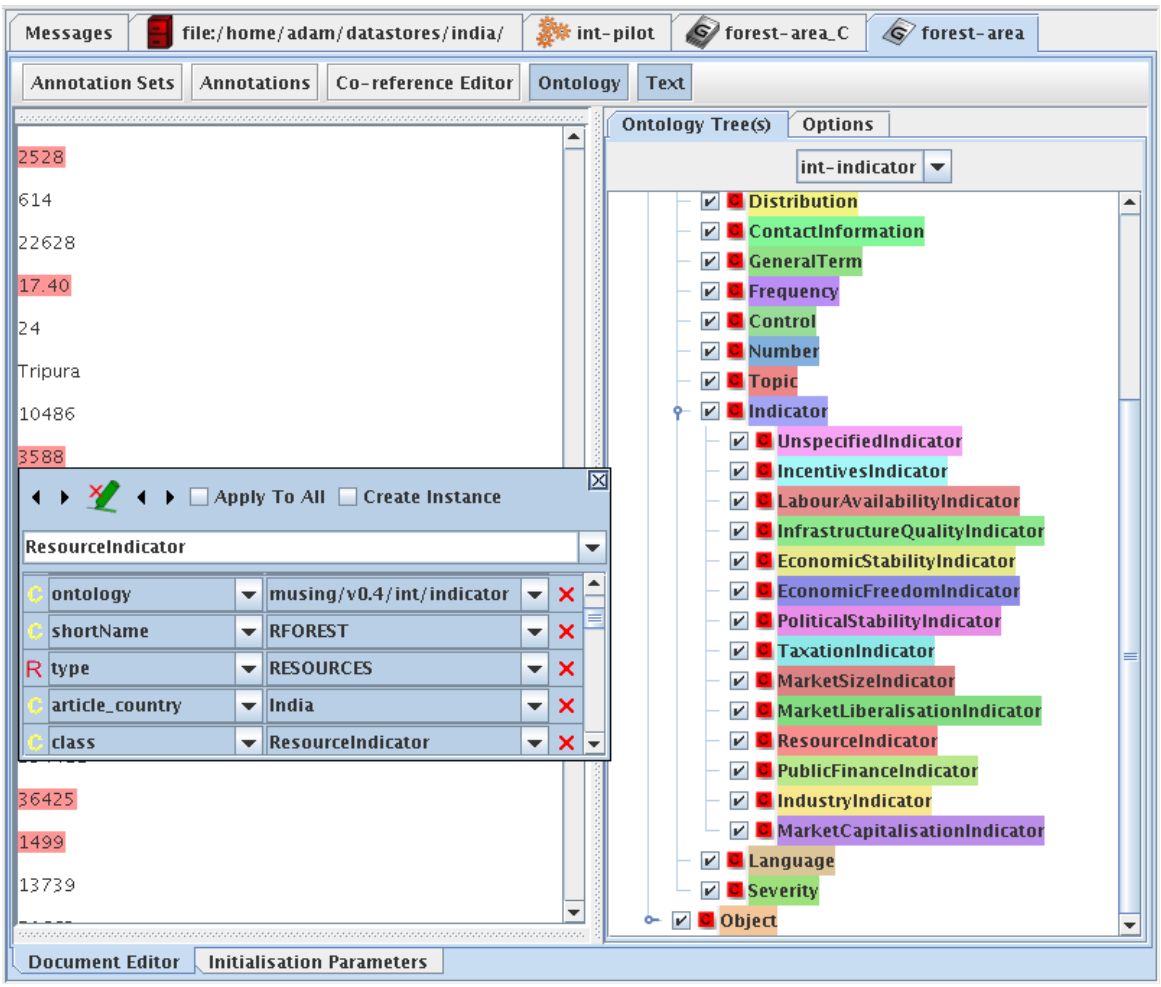

Figure 1: *Ontological annotation in GATE*

Tool, OCAT, as illustrated in Figure 1.[7] Although GATE's annotation model (which is based on the TIPSTER model) allows each annotation to contain a map of arbitrary feature-value pairs, using the OCAT extension constrains all new manual annotations to have the same type (usually `Mention`) and a `class` feature whos values must be selected from the active ontology. The annotators can do this work using the normal, locally installed GATE GUI or an easy remote service based on JavaWebStart[8].

These documents (at this point, especially the news articles from the financial press) are being used to train machine-learning applications for the project's information extraction tasks.

In particular, the web service allows us to serve documents that have already been automatically annotated so that the human annotators can correct them by adding, removing or changing annotations. These documents with manually improved annotations can be fed back into an information extraction system designed for progressive improvements in machine learning that takes advantage of the ontological structure of the annotations. [11, 12]

3.2.3 XBRL mapping

Financial information written in XBRL (which was briefly introduced in Section 3.1) is already machine-readable, and MUSING's ontologists are working on mappings between XBRL and our domain ontologies to ensure that such data can be annotated very precisely with full integrability into our system. Related information supplied with XBRL data (such as free text associated with a financial statement) can be annotated with the other techniques (declarative information extraction and machine learning) using the same ontologies for consistency (although at a performance level appropriate for analysis of natural language). [7]

3.3 Integration

Instead of concentrating on the traditional, low-level multi-source information extraction tasks such as cross-document coreferencing, we are interested here in the high-level task of refining and growing a knowledge base in a consistent manner.

For this purpose, ontology experts at DERI Innsbruck[9] have developed and continue to refine, based on information provided by other MUSING partners, a set of domain ontologies for business intelligence that extend the Proton[10] ontology.

To be precise, the MUSING ontologies contain `owl:imports` statements that refer to Proton's *System*, *Top* and *Knowledge Management* modules (using the *Upper* module would adversely affect decidability) so that our domain-specific extensions consist of subclasses and instances of Proton classes, as well as instances of our classes; this extension of a well-known *de facto* standard in the semantic web field could facilitate interoperability with other parties' tools in the long term.

We therefore ensure that all the automatic annotation carried out by both types of components described in Section 3.2 makes good use of this semantic enhancement: specifically, we design our declarative components so that every annotation contains `ontology` and `class` features whose values point to a particular MUSING ontology and to one of its classes, respectively; and we ensure that our training data (manually annotated documents) and machine-learning components also respect this requirement.

We can process a diverse range of input documents through appropriate information extraction engines in a many-to-many relationship; documents can be analysed with several techniques to capture a wider range of information.

Figure 2 shows a schematic of the integration of these processes into a multi-source information-extraction *system*, in which the common set of domain ontologies (tied together as extensions of Proton) serves to unify the engines (as indicated by the dashed arrows) so that the resulting annotated documents are semantically coherent and the information can be consistently added to an ontological knowledge base.

The different applications running in parallel can also be evaluated to compare their performance as part of the evaluation of our system and to help us refine it.

In effect, the constraints on annotation according to the domain ontologies act as an "information funnel" to ensure consistency and compatibility of the extracted information going into the knowledge base.

3.4 Coreferencing

Our approach also deals with the classical problem of cross-document coreferencing, but takes advantage of the semantically enhanced annotation in order to treat it as an *ontology population* problem. For example, Figure 3 shows three texts

[7] The right-hand pane shows ontology classes colour-coded to match the corresponding annotations in the left-hand pane.

[8] `http://java.sun.com/products/javawebstart/`

[9] `http://www.deri.at/`

[10] `http://proton.semanticweb.org`

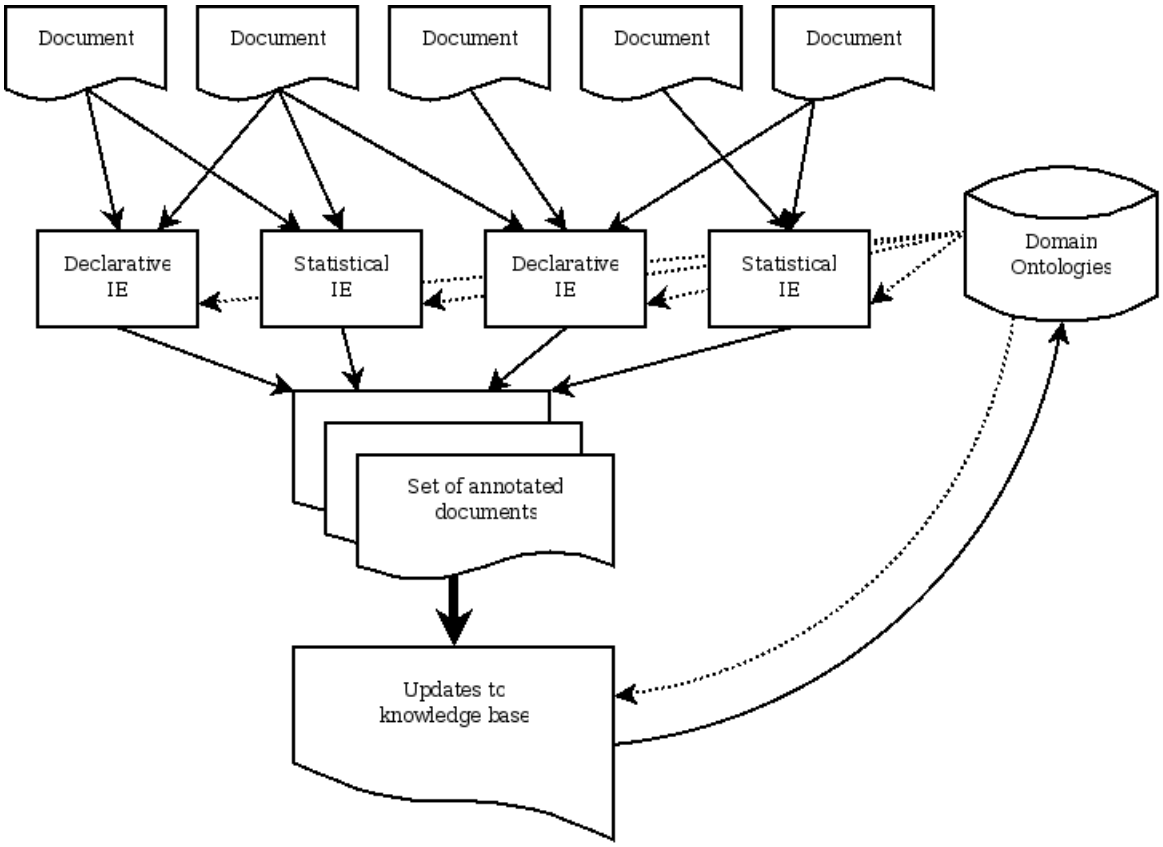

Figure 2: *Ontological integration of extracted information*

that use different expressions to refer to the same company (Alcoa Inc.), and we wish to link the separate and often complementary pieces of information (address, chairman, financial announcements) together for better intelligence.

We treat each named entity as a possible ontology instance and retrieve candidate instances (of companies, for example) from the relevant domain ontology, along with known features of those instances from the knowledge base. We will then employ a rule-based system defined by domain experts to compute similarity scores between the possible instance and the candidates, in order either to dereference the named entity or to add a new instance to the ontology (which uses the KIM OWLIM [17] semantic repository).

3.5 Refinement

Both the declarative and machine-learning information extraction tools will be continually refined through a feedback loop in which human annotators correct the automatic annotations on selected documents using AnnotatorGUI [reference], a web application tool recently developed by the NeOn project.[11]

This application is deployed as a JavaWebStart service (briefly discussed above in Section 3.2.2) which runs and loads GATE documents and ontologies from a server at the University of Sheffield and save the modified documents back to the same server, where they can later be manually inspected to refine the declarative information extraction tools or automatically fed back into a machine-learning loop. The correct annotations (according to the human annotator) and the previous automatic annotations can be stored in distinct annotation sets in the same document, so that the automatic ones can be scored using GATE's AnnotationDiff tool.

New documents can also be introduced into the loop in order to test and improve the versatility of existing IE tools and to determine if new ones are required to enlarge the scope of the integrated system.

4 Discussion and future work

We have described the design of a coherent system of information extraction for business intelligence, which uses expertly designed domain ontologies that extend a *de facto* standard (Proton) for the semantic web in order to integrate the output of a variety of separate information extraction tools that process a variety of document types ranging from unstructured text to highly structured information. This system also allows superficial redundancy in that each document could be processed using multiple tools in order to improve the overall precision (i.e. to reduce the quantity of "missed" information).

We have implemented much of this system but have not yet annotated enough data to carry out reliable quantitative evaluation[12], which will be a focus for our work in the new future—not just to validate our work but also to continue to carry it out, since such evaluation is an important part of the feedback loop described in Section 3.5.

Acknowledgements

This research is supported by the EU grant IST-2004-027097 for the MUSING (Multi-industry, semantic-based next generation business intelligence) project and benefits from work supported by EU grant IST-2005-027595 for the NeOn (Networked ontologies) project.

References

[1] D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, and M. Tyson. Description of the JV-FASTUS system as used for MUC-5. In *Proceedings of the Fourth Message Understanding Conference MUC-5*, pages 221–235. Morgan Kaufmann, California, 1993.

[2] A. Bagga and B. Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79–85, 1998.

[3] A. Bagga and A. W. Biermann. A methodology for cross-document coreference. In *Proceedings of the Fifth Joint Conference on Information Sciences (JCIS 2000)*, pages 207–210, 2000.

[4] K. Bontcheva, D.Maynard, V. Tablan, and H. Cunningham. GATE: A Unicode-based infrastructure supporting multilingual information extraction. In *Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgaria, 2003.

[5] H. Cunningham, D. Maynard, and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, Nov. 2000.

[6] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[11] `http://www.neon-project.org/`

[12] This will be based on precision and recall over annotations, as is usual in this type of work, with appropriate scoring for partial matches not only for overlapping annotation wspans but also for subsumption of annotation classes.

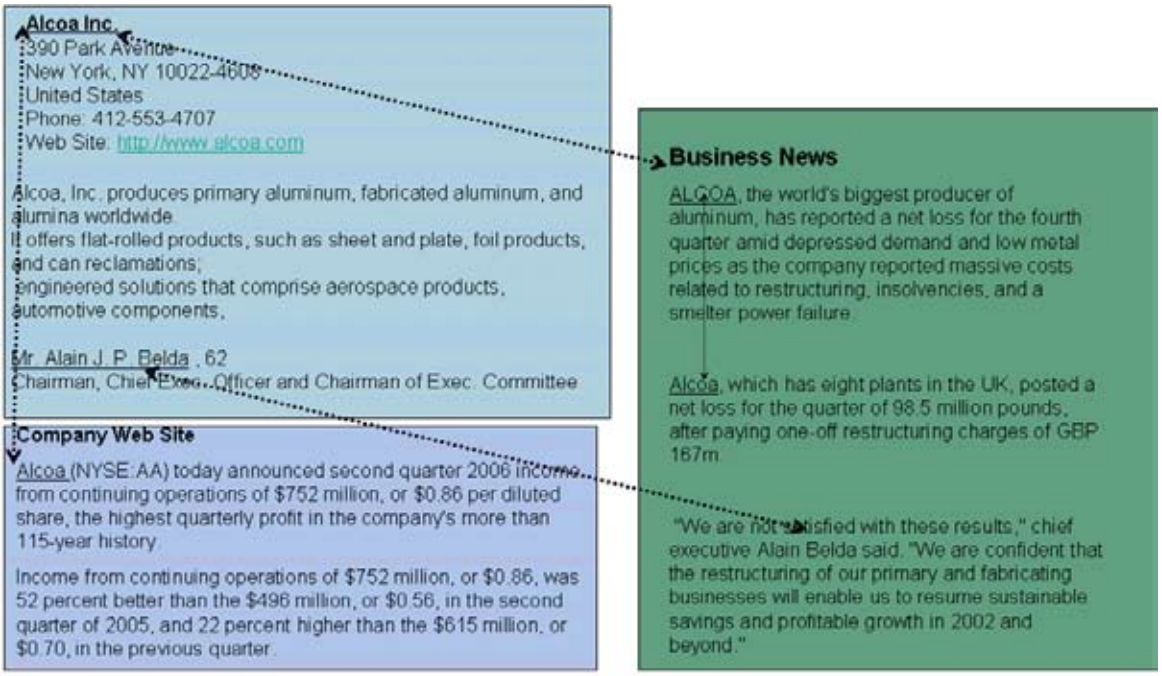

Figure 3: *Related information from multiple sources*

[7] T. Declerck and H. Krieger. Translating XBRL into Description Logic: an approach using Protege, Sesame and OWL. In *Proceedings of Business Information Systems (BIS)*, Klagenfurt, Germany, 2006.

[8] J.-Y. N. F. Paradis and A. Tajarobi. Discovery of business opportunities on the internet with information extraction. In *Workshop on Multi-Agent Information Retrieval and Recommender Systems (IJCAI)*, pages 47–54, Edinburgh, Scotland, 2005.

[9] F. Fornasari, A. Tommasi, C. Zavattari, R. Gagliardi, T. Declerck, and M. Nannipieri. Xbrl web-based business intelligence services. In P. Cunningham and M. Cunningham, editors, *Innovation and the Knowledge Economy: Issues, Applications, Case Studies. Proceedings of eChallenge 2005*. IOS Press, 2005.

[10] Z. Kazi and Y. Ravin. Who's who? Identifying concepts and entities across multiple documents. In *Proceedings of the* 33^{rd} *Hawaii International Conference on System Sciences*, volume 3, Hawaii, USA, Jan 2000.

[11] Y. Li, K. Bontcheva, and H. Cunningham. Perceptron-like learning for ontology based information extraction. Technical report, University of Sheffield, Sheffield, UK, 2006.

[12] Y. Li, K. Bontcheva, and H. Cunningham. Hierarchical, Perceptron-like Learning for Ontology Based Information Extraction. In *16th International World Wide Web Conference (WWW2007)*, pages 777–786, 2007.

[13] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks. Named Entity Recognition from Diverse Text Types. In *Recent Advances in Natural Language Processing 2001 Conference*, pages 257–274, Tzigov Chark, Bulgaria, 2001.

[14] D. Maynard, H. Cunningham, A. Kourakis, and A. Kokossis. Ontology-Based Information Extraction in hTechSight. In *First European Semantic Web Symposium (ESWS 2004)*, Heraklion, Crete, 2004.

[15] D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop "End User Apects of the Semantic Web")*, Heraklion, Crete, 2005.

[16] D. Maynard, H. Saggion, M. Yankova, K. Bontcheva, and W. Peters. Natural Language Technology for Information Integration in Business Intelligence. In *10th International Conference on Business Information Systems (BIS-07)*, Poznan, Poland, 2007.

[17] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM – Semantic Annotation Platform. *Natural Language Engineering*, 2004.

[18] Y. Ravin and Z. Kazi. Is Hillary Rodham Clinton the president? disambiguating names across documents. In *Proceedings of the ACL 1999 Workshop on Coreference and its Applications*, Jun 1999.

Using Information Extraction to Improve Cross-lingual Document Retrieval

Dilek Hakkani-Tür
ICSI
Berkeley, CA, 94704, USA
dilek@ICSI.Berkeley.EDU

Heng Ji Ralph Grishman
Department of Computer Science, New York University
New York, NY, 10003, USA
{hengji, grishman}@cs.nyu.edu

Abstract

We present a filtering mechanism using two cross-lingual information extraction (CLIE) systems for improving document relevance of cross-lingual information retrieval (CLIR) for queries conforming to predefined templates. Experiments on retrieving Chinese documents in response to English GALE[1] arrest queries show that this approach can obtain a 12.7% absolute improvement in relevance (representing a 24.8% relative error reduction) for the top 25 retrieved documents. We also demonstrate that Chinese IE can provide a valuable supplement to English IE to enhance retrieval performance.

Keywords

Cross-lingual, Information Extraction, Document Retrieval

1. Introduction

A shrinking fraction of the world's Web pages are written in English, and so the ability to access pages across a range of languages is becoming increasingly important for many applications. This need is being addressed in part by the research on Cross-Lingual Information Retrieval (CLIR), which, given an arbitrary query stated in one language, seeks to retrieve relevant documents in one or more foreign languages. CLIR combines two difficult problems, document retrieval and term or text translation, and these have limited the performance of CLIR systems in general.

For some applications, however, we are able to identify particular types of queries which are of primary interest to a community of users. In such circumstances we can optimize the document retrieval function for these queries, while still using general CLIR mechanisms for other queries. In this paper we study the benefits of this optimization.

Specifically, we conducted this study in the context of the "distillation evaluation" of the GALE program. As we will describe in more detail below, the evaluation task involves answering queries with respect to a multilingual collection, where the queries must conform to one of a set of query templates. Some of these templates allow for very general queries, while others involve specific types of relations or events. We examine one of the more specific query templates in detail, focusing on the problem of retrieving relevant Chinese documents in response to the (English) query. We compare several retrieval strategies, using as a baseline keyword-based retrieval on the English (machine) translations of the documents, and then adding filters based on the output of two information extraction systems, one operating on the Chinese source, the other on the translated documents. We show in particular the benefits (and some of the limitations) of using source-language information extraction for document filtering.

The rest of this paper is structured as follows. Section 2 describes our main research task and experimental setting. Section 3 briefly describes the previous efforts made by researchers of using sophisticated linguistic analysis to enhance IR performance. Section 4 describes the motivation for our approach. Section 5 presents an overview of our system architecture and strategies for using IE to filter out the irrelevant documents. Section 6 presents the experimental results. Section 7 compares our approach with two possible alternative approaches and Section 8 then concludes the paper and sketches our future work.

2. Task and Terminology

Our approach has been evaluated in the framework of the U.S. Government's DARPA GALE program. One of the GALE evaluations (the *distillation* task) involves responding to queries based on a set of question templates (17 templates in GALE 2007). For the experiments presented here, we used the *arrest* class of questions (template number 15, according to the GALE program). These take the form: "***Describe arrests of persons from X and give their role in the organization***", in which X is an organization name such as "Peruvian government", "Shining Path", "WorldCom", "US Federal Bureau of Investigation", "Enron Corporation", "Jemaah Islamiyah", "ETA", "al Qaeda" and "the PLF".

We use the University of Massachusetts INDRI IR system[2] as a major component to return the top N (N<=50 in this paper) relevant documents in response to a query. We then use a statistical approach to extract sentences from these documents. Our goal in this paper is to improve the precision of identifying Chinese documents relevant to these questions (we don't consider the sentence extraction phase in this paper).

[1] Global Autonomous Language Exploitation

[2] http://www.lemurproject.org/indri/

We make use of two cross-lingual IE systems developed around the ACE[3] evaluations to reach these goals. ACE defines 8 types of events, with 33 subtypes, including *Arrest-Jail* events which cover arrests, capture events, extraditions, jailing events, incarceration, etc.

3. Related Work

The attempt to marry natural language processing (NLP) techniques with large-scale IR is not new, but effective integration of the two remains an open research question. Researchers have experimented with syntactically derived word pairs (Strzalkowski et al., 1996; Zhai et al., 1996; Arampatzis et al., 1998), case frames (Croft and Lewis, 1987), paraphrases (Duclaye and Yvon, 2003), part-of-speech tagging (Eichmann 2003), name tagging (Eichmann 2003; Harabagiu et al., 2005; Katz and Lin, 2003), reference resolution (Harabagiu et al., 2005), parsing (Smeaton et al., 1994; Harabagiu et al., 2005) and syntactic relation patterns (Shen et al., 2005) as units of indexing. However, none of these experiments resulted in a dramatic improvement in precision or recall, and sometimes even resulted in degraded performance. Note however these efforts aim to handle arbitrary retrieval queries, whereas we can take advantage of specific types of query templates. In that regard, our work is more similar to the document filtering experiments which are sometimes used to assess information extraction (Yangarber et al., 2000).

Our spirit of using IE results as a post-processor for IR is closest to (Schiffman et al., 2007). But we have a different focus, on the problem of CLIR, whereas they emphasized the extraction of sentences from English texts only. They use IR at the first stage to return a small number of relevant documents; IE results are then used to select highly relevant words to revise the query for a second retrieval pass. But in our cross-lingual environment, the candidate documents returned by IR engine are much more noisy, therefore we use IR in high recall mode to return a large collection, then use extracted events to filter the irrelevant documents to improve precision. Our work can be considered as an application of the filtering approach in (Hakkani-Tür et al., 2007) in a cross-lingual environment.

4. Motivation

Despite the intuition that linguistically sophisticated techniques should be beneficial to IR, real gains in performance have yet to be demonstrated empirically in a reliable manner. We believe that the key to effective application of NLP technology is to selectively employ it in situations where we can expect to improve performance, without abandoning simple techniques. We felt that these template-based queries offer such an opportunity.

The setting of template-based CLIR has encouraged the development of CLIE systems, but it also raises the issue of the best approach for CLIE performance. We can employ the following two cross-lingual IE (CLIE) pipelines to process Chinese documents:

MT_English IE: Translate Chinese texts into English, and then run English IE on the translated texts.

Chinese IE_MT: Run Chinese IE on the Chinese texts, and then use MT word alignments to translate (project) extracted information into English.

In the mono-lingual environment English IE systems generally perform better than Chinese IE, so it's natural to apply English IE. However, Chinese is linguistically very different from English and statistical MT performance for Chinese-English is still quite poor. Therefore the process of applying English IE on MT output becomes much more lossy than on English documents.

In order to quantify the information lost by MT, we count the number of "arrest" event trigger words (A trigger is the word that most clearly expresses the event occurrence) which have associated arguments, detected from 19 ACE Chinese texts, and their overlap with the true events. The results are shown in Figure 1.

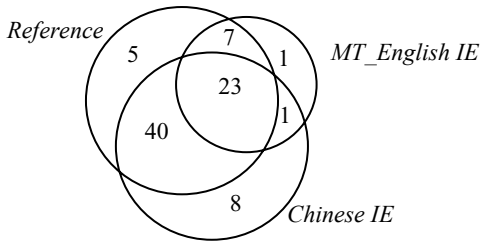

Figure 1. Number of event triggers with arguments

We can see that applying English IE on MT provides high precision (only 2 of 32 extracted events are not in reference), but only covers a small fraction (30/74) of the key events. On the other hand, by bypassing MT, Chinese IE can be a useful supplemental procedure, recovering 40 correct events missed by English IE, although it has lower accuracy (9 wrong events were generated at the same time).

In the following we use two translated example texts, to give more intuition about how these relevant events were missed by MT.

Example 1:

[***Query***] *Describe arrests of persons from [al Qaeda] and give their role in the organization.*

[***Chinese Text***] *[迪拜]昨天阿拉伯新闻频道报道说盖达组织第二首脑, 欧萨玛的得力助手扎瓦里已经在伊朗落网。*

[***Text Machine Translation***] *DUBAI Arab news channel, reported yesterday that the terrorist organization Al-Qaida*

[3] http://www.itl.nist.gov/iad/894.01/tests/ace/

second, Osama's right-hand man, Abdurrahman Wahid in Iran.

[***Text Reference Translation***] *(DUBAI) Arab News Channel reported that Zawahiri, the second most important person of Al-Qaeda and Osama's right-hand man, was arrested in Iran.*

Example 2:

[***Query***] *Describe arrests of persons from [the PLF] and give their role in the organization.*

[***Chinese Text***] *据报道，“巴勒斯坦解放阵线”领导人阿巴斯在巴格达南郊的住所内被美军逮捕。*

[***Text Machine Translation***] *According to reports, PLF leader Abbas in Baghdad outside the residence of the US military.*

[***Text Reference Translation***] *According to reports, the leader of the PLF, Abbas was arrested by the US military in his residence in the outskirts of Baghdad.*

Although the organization names “al Qaeda” and “PLF” in these queries are correctly translated, the event trigger words representing “arrest” are missing. In example 1, the “arrest” trigger “落网 (fall into meshwork)” is fairly rare and metaphorical; in example 2, MT met difficulty probably because of the re-ordering of phrases. In these cases, applying IE directly on source (Chinese) texts could help.

To sum up, combining both CLIE pipelines could allow us to incorporate information from a much wider knowledge base, spanning both the original and the translated documents. In the following section we will describe the algorithms to capture these intuitions.

5. System Architecture

5.1 System Overview

The overall system pipeline is presented in Figure 2. We split document retrieval into a two-stage process. The first stage simply applies cross-lingual IR, without any IE knowledge, and initially retrieves the top N (N<=50) documents for each query. Then we use the events detected from CLIE as additional constraints to determine whether a document is relevant or not. In the following we shall present the system components and detailed filtering algorithm.

5.2 Machine Translation

Both CLIE systems need machine translation to translate Chinese documents (for English IE) or project the extraction results from Chinese IE into English. We use the RWTH Aachen Chinese-to-English machine translation system (Zens and Ney, 2004) for these purposes. It’s a statistical, phrase-based machine translation system which memorizes all phrasal translations that have been observed in the training corpus. It computes the best translation using a weighted log-linear combination of various statistical models: an n-gram language model, a phrase translation model and a word-based lexicon model. The latter two models are used in source-to-target and target-to-source directions. Additionally, it uses a word penalty and a phrase penalty.

The model scaling factors are optimized on the development corpus with respect to the BLEU score similar to (Och 2003). Almost all bilingual corpora provided by LDC were used for training, which account for about 200 million running words in each language. Language modeling used the English part of the bilingual training corpus and in addition some parts of the English GigaWord corpus. The total language model training data consists of about 600 million running words.

This MT system produces a translation for each source document, and also the *word-to-word* mapping derived from phrase-based alignment.

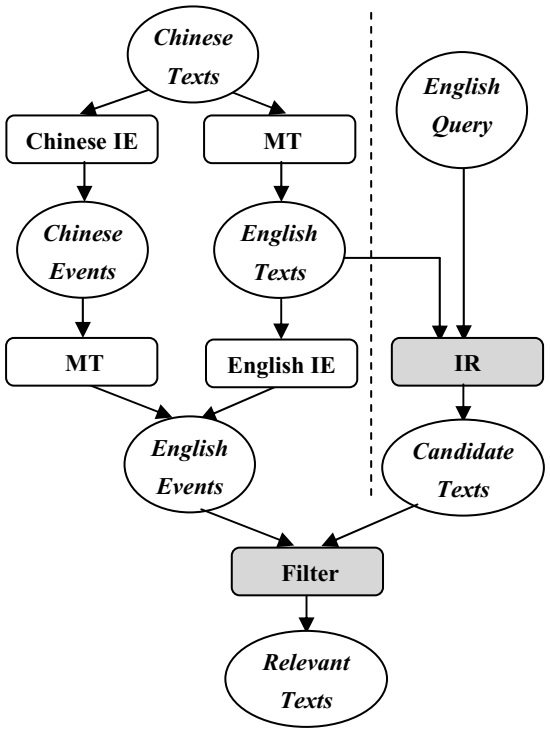

Figure 2. System Architecture

5.3 Baseline Cross-lingual IR

For an English query, we use the INDRI retrieval system (Strohman et al., 2005) to identify the top N translated documents, ranked by IR score. INDRI combines language modeling and inference network approaches in an architecture designed for large-scale applications. INDRI expands the query by keyword parsing, operator conversion, and date/numeric word processing. Like most other IR systems, INDRI is task independent and does not perform any deep analysis that particularly addresses the event information. However, since the query templates are known in advance, it has been possible to hand-craft queries for each template, incorporating additional keywords into the query (for example, adding “apprehend” and other related terms to the arrest query).

5.4 IE

We present the training and test procedures for the two IE systems as follows. Both of them are trigger-based and use pattern matching.

The English IE system combines pattern matching with statistical models (Grishman et al., 2005). For every instance of an event in the ACE training corpus, we construct two types of patterns representing the connection between the trigger word and the event arguments, and record the type and subtype of the associated event. One pattern is the sequence of constituent heads separating the trigger and arguments. The other pattern is the predicate-argument structure (Meyers et al., 2001) connecting the trigger to all the event arguments. For each argument, we record its ACE type and subtype and its head. In addition, we train a set of MaxEnt classifiers to distinguish events from non-events, to classify events by type and subtype, to distinguish arguments from non-arguments, and to classify arguments by argument role. In the test procedure, each document is scanned for instances of triggers from the training corpus. When an instance is found, we first try to match the environment of the trigger against the set of patterns associated with that trigger. This pattern-matching process, if successful, will assign some of the phrases in the sentence as arguments of a potential event. The argument classifier is applied to the remaining roles in the sentence; for any passing that classifier, we use the role classifier to assign a role to the mention. Finally, once all arguments have been assigned, we apply the event classifier to the potential event; if the result is successful, we report this as an event instance.

The Chinese IE system is based on patterns semi-automatically extracted from the ACE training corpus. For each event instance we replace the trigger word by its event type and subtype, and each argument by its entity type. Then we apply the Purdue University POS tagger (Huang et al., 2007) and chunker (Harper et al., 2005) on each event training instance. Then we edit the patterns by hand, replacing tokens by their POS tag, chunk type, or a wild card or deleting them entirely if they are not relevant to detecting the event type. Some patterns are collapsed, and some patterns which appear too specific or too general are deleted. To ensure that patterns are not over-generalized by the hand editing, the training corpus is split in two and patterns derived from one half are, after hand editing, applied to the other half to review their accuracy in event prediction. In the test procedure, each document is annotated with POS tagging and chunking, and then scanned against the patterns derived from the training corpus. Unlike the English IE system, currently we don't have statistical models following pattern matching.

5.5 Filtering Approach

In the GALE task we are given templates for queries in advance. When a template-based query is submitted to this IR engine, one challenge is determining the number of relevant documents that can be used for template-based question answering from the information retrieval output. This is problematic since it is hard to know the optimal value that holds for all queries. Sometimes a query has only one relevant document in the huge document repository and sometimes thousands. If the sentence extraction system processes a larger number of returned documents, this is expected to result in a higher number of false alarms unless document level processing is available.

One solution might be getting fewer documents from IR but this may result in poor recall. Alternatively one could exploit document and argument scores returned by INDRI. Here, the argument score reflects matches between the document and the value of the slots in the queries. However the document and argument scores usually have different dynamic ranges depending on the query and it is not easy to perform thresholding that works optimally for all queries using them.

Therefore, we use an intermediate processing stage between the IR engine and the sentence extraction module, to filter out irrelevant documents. The basic idea is as follows: Since the distillation query templates are known beforehand, it is sometimes possible to associate expected document contents with one or several types of ACE event annotations. For example for:

> **Query template 15:** *Describe arrests of persons from [organization] and give their role in the organization.*

the relevant document must have the ACE event of subtype *Arrest-Jail*. Since the CLIE systems provide such annotations from both source language and translated documents, the post-processing stage needs to check only whether the event mentioned in the query appears in the documents returned by IR. A more extensive version can also require that the organization name specified in the query be present in the returned document. However, in this work, we only considered filtering according to event types and subtypes.

6. Experimental Results

In this section we shall present the results of applying this method to improve GALE cross-lingual document retrieval.

6.1 Data

We evaluated our approach by using 12 example queries of GALE template 15 to retrieve documents from the English machine translation of the TDT5 Chinese corpus consisting of 56,485 newswire texts from four different news agencies. For each query the baseline IR system returns the top N (N<=50) documents, in total 421 documents.

Then these documents were manually labeled as relevant or not-relevant by one of the authors to construct the reference set for the evaluation. The decisions were made against the original Chinese documents, using the

rules in the GALE annotation guidelines[4]. For example, a document describing "the history of the Muslim Brotherhood" (e.g. "The Muslim Brotherhood is viewed as the second strongest political force in Egypt") is not relevant to the query for X= "the Muslim Brotherhood". We did not judge against MT output because when the translation quality is poor that procedure tends to be too subjective and MT system-specific. In total 130 documents were judged as relevant.

6.2 Evaluation Metric

6.2.1 Precision, Relative-Recall and F-Measure

We use the traditional IR metrics, *precision* and *relative-recall,* to evaluate our approach. As we have noted, the INDRI engine was set to return a maximum of 50 documents per query, and alternative results were obtained by filtering the set returned by INDRI. Consequently, for evaluation purposes our 'relevant document set' **R** was the subset of these 50 (or fewer) documents per query judged to be relevant. The baseline system therefore had a maximum relative recall of 100%.

If the system (with filtering) returns document set **D**, then precision and relative-recall can be defined as follows:

$$Precision = \frac{\#(\mathbf{D} \cap \mathbf{R})}{\#\mathbf{D}}$$

$$Relative\_Recall = \frac{\#(\mathbf{D} \cap \mathbf{R})}{\#\mathbf{R}}$$

F-measure combines these two metrics:

$$F\text{-}Measure = \frac{2*Precision * Relative\_Recall}{Precision + Relative\_Recall}$$

6.2.2 Search Length i

We also use the Search Length i Measure (Cooper, 1968) to measure user effort in terms of the number of non-relevant documents that a user must examine before finding i relevant documents. In our study we set i = 5.

6.3 Overall Performance

Figure 3 presents the overall precision and relative recall for the baseline and after adding cross-lingual IE to filter the irrelevant documents. The numbers on the curve show the F-measure (%) scores when N (the number of documents returned by INDRI) = 10, 20, 30, 40 and 50.

We can see that using English IE on MT output does indeed help achieve significant improvements in precision, but also relatively large loss in recall. Then by adding Chinese IE at the end, the system dramatically boosts recall, with a small loss in precision. As N increases, the overall improvement in F-measure increases from 4.8% to 18.8%.

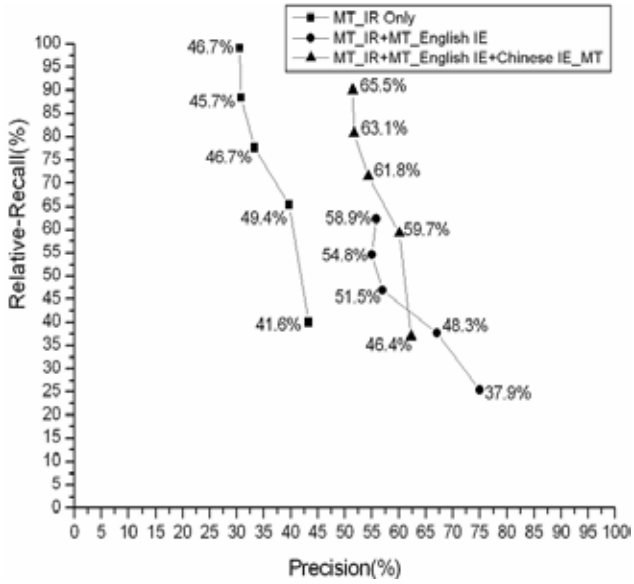

Figure 3. Overall Precision/Relative-Recall

6.4 Performance Breakdown

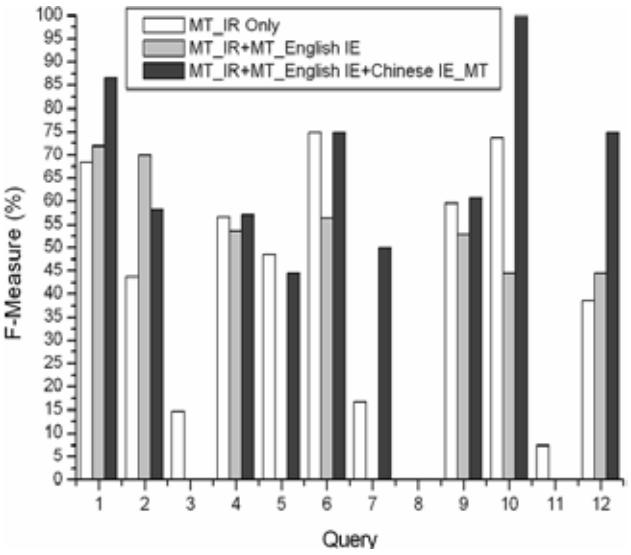

Figure 4. F-Measure for Each Query When N=25

We then performed experiments to evaluate the impact of IE on each query, with a fixed N=25[5]. Figure 4 shows the results. Overall the F-measure increases from 48.8% to 50.5% by using English IE on MT, and increases to 61.5% by further adding Chinese IE.

Except for the small losses for queries 3, 5 and 11, IE produces clear improvements for all the other nine queries, with 1.5%-36.3% gain in F-measure. By adding Chinese IE, the F-measures for all the queries except for

[4] http://projects.ldc.upenn.edu/gale/Distillation/DistillationTrainingDataGuidelinesV2.3.pdf

[5] We assume N=25 reflects the number of texts a user would look at for a document retrieval task.

query 2 are significantly boosted. In the contrast, the performance by adding only English IE is much less consistent.

In order to check how robust our approach is, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on relevance F-measures for these 12 queries, for all the retrieved documents (N<=50). The results show that we can reject the hypothesis that the improvements using IE were random at a 94.6% confidence level, and reject the hypothesis that the improvements by adding Chinese IE were random at a 94.5% confidence level.

6.5 User Effort Measure

For the 8 queries which retrieved at least 5 relevant documents, we measure their search length with i=5 as defined in section 6.2.2. In general, our approach using IE knowledge significantly outperforms the baseline, reducing the mean of search length from 1.625 to 0.375.

6.6 Error Analysis

Getting a reasonably complete set of patterns (linguistic expressions) for an event type is inherently difficult. Gaps in the pattern sets used by IE lead to the filtering out of relevant documents:

Example 3:

[***Query***] *Describe arrests of persons from [Hezbollah] and give their role in the organization.*

[***Chinese Text***] *2000 年 10 月以来, 真主党已有 4 名成员被以军扣押。*

[***Text Reference Translation***]: *Since October of 2000, four Hezbollah members have been (detained and seized) by the Israeli army.*

The arrest event was missed because the trigger compound word "扣押 (detain and seize)" doesn't appear in the ACE training data.

Other errors reflect the drawbacks of not using richer CLIE outputs:

Example 4:

[***Query***] *Describe arrests of persons from [Peruvian government] and give their role in the organization*

[***Chinese text***]*玻利维亚警方 30 日宣布在该国中部缴获 186.3 公斤纯可卡因，并逮捕 5 名贩毒嫌疑人。据警员路易斯介绍， 5 名贩毒嫌疑人经常在玻利维亚收集毒品，并定期运往秘鲁,然后再贩卖给美国和欧洲。*

[***Text Reference Translation***] *On 30th Bolivian Police announced that 186.3 kilogram of pure cocain were captured in the middle of the country, and five suspects of selling drugs were arrested. According to the policeman Louis ,these suspects often collected drugs in Boliva, and regularly shipped to Peru, and then sold to America and Europe..*

The document is not relevant to the query because "Peru" is not an event argument of "arrest". But our filter mistakenly confirmed it as a relevant text without using event argument information. Taking advantage of this information (which is generated by IE) could filter out some irrelevant documents, but (due to errors in IE argument output) could also block some relevant documents.

7. Discussion

7.1 Comparison with Query Expansion

One alternative solution to reduce query/document mismatch is to add a more extensive list of "arrest" trigger words to the query. But when such trigger words are used without context, they may be highly ambiguous, leading to many false hits. For example, "pursue" can mean "to follow in an effort to overtake or capture" to indicate an arrest event: "US-Britain special troop pursued Bin Laden"; but it also can mean "to strive to gain or accomplish" such as "pursue lofty political goals".

Furthermore, the Chinese ACE 2005 training corpus includes 43 different "arrest" trigger words, some of which appear in different forms, noun and verb, sometimes singular and plural, yielding many different trigger translations for a given event type (e.g. armies attack, bombs explode…can all indicate 'attack').

7.2 Comparison with Query Translation

The work described here complements the cross-lingual question answering (CLQA) research such as (Mitamura et al., 2006). They presented an English-to-Chinese QA system that translates the query from English to Chinese and then searches the translated query among Chinese documents. Combining this approach into our framework can magnify the gains possible with IE for cross-lingual IR.

However, as Mitamura et al. mentioned, the English to Chinese translation accuracy is low because of word sense ambiguities as well as regional language differences for Chinese. In our task the main difficulty of applying this approach is to properly translate English names into Chinese. In particular, many names in our task are written in abbreviations, such as "ETA" and "the PLF", which will be difficult to translate/expand into Chinese. The regional language difference problem also commonly exists in English-to-Chinese name translation. For example, "al Qaeda" is translated into "基地组织" (based on meaning) in mainland China but "阿尔盖达" (based on its pronunciation) in Taiwan, and "卡伊达" in Singapore (based on part of its pronunciation). This could immediately lead to failure in document retrieval using a query-translation approach.

8. Conclusion and Future Work

We identified the linguistic phenomena that cross-lingual IR has difficulty with, even with predefined query templates, and demonstrated that IE enables us to successfully handle these difficulties by effectively exploiting events that can be reliably extracted from texts and matching queries with documents at the event level, thereby significantly improving precision of document retrieval with little loss in recall.

The experiments suggest that some further gains can be achieved through employing richer IE results such as

event arguments, combining this with additional shallow linguistic features such as the positions of events, titles, document structure, document topic and name concurrence. We are also interested in applying the filtering approach to the snippet (sentence and sub-sentence) retrieval system described in (Hakkani-Tür and Tur, 2007).

9. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency under Contract No. HR0011-06-C-0023, and the National Science Foundation under Grant IIS-00325657. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government. The authors would like to thank Gokhan Tür, Mary Harper and Satoshi Sekine for their valuable help and comments on this work.

10. References

[1] A. Arampatzis, Th.P. van der Weide, C.H.A.Koster, and P. van Bommel. 1998. Phrase-based Information Retrieval. *Information Processing and Management*. 34(6):693-707, December

[2] W. S. Cooper. 1968. Expected search length: A Single Measure of Retrieval Effectiveness based on the Weak Ordering Action of retrieval systems. *Journal of American Society of Information Science*, 19(1), 30-41.

[3] B. Croft and D. Lewis. 1987. An Approach to Natural Language Processing for Document Retrieval. *In ACM SIGIR 1987*.

[4] Florence Duclaye and Francois Yvon. 2003. Learning Paraphrases to Improve a Question-Answering System. *In ACE2003 workshop on natural language processing for question answering*.

[5] David Eichmann. 2003. Issues in Extraction and Categorization for Question Answering. *In AAAI Spring Symposium on New Directions in Question Answering*, Stanford, CA, US.

[6] Ralph Grishman, David Westbrook and Adam Meyers. 2005. NYU's English ACE 2005 System Description. *In ACE 2005 Evaluation Workshop*. Washington, US.

[7] Sanda Harabagiu, Dan Moldovan, Christine Clark, Mitchell Bowden, Andrew Hickl and Patrick Wang. 2005. Employing Two Question Answering Systems in TREC-2005. *In TREC 2005*.

[8] Mary Harper, Bonnie Dorr, John Hale, Brian Roark, Izhak Shafran, Matthew Lease, Yang Liu, Mattew Snover, Lisa Yung, Anna Krasnyanskaya and Robin Stewart. 2005. *Parsing and Spoken Structural Event Detection*. Technical Report, The John-Hopkins University, 2005 Summer Research Workshop.

[9] Zhongqiang Huang, Mary Harper and Wen Wang. 2007. Mandarin Part-Of-Speech Tagging and Discriminative Reranking. *In EMNLP 2007*. Prague, Czech Republic.

[10] Boris Katz and Jimmy Lin. 2003. Selectively Using Relations to Improve Precision in Question Answering. *In EACL 2003 Workshop on Natural Language Processing for Question Answering*

[11] Adam Meyers, Michiko Kosaka, Satoshi Sekine, Ralph Grishman and Shubin.Zhao. 2001. Parsing and GLARFing. *In Proceedings of RANLP-2001*.

[12] Teruko Mitamura, Mengqiu Wang, Hideki Shima and Frank Lin. 2006. Keyword Translation Accuracy and Cross-lingual Question Answering in Chinese and Japanese. *In EACL 2006 Workshop on Multilingual Question Answering.*

[13] F. J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translaton. *In ACL 2003.*

[14] Barry Schiffman, Kathleen R. McKeown, Ralph Grishman and James Allan. 2007. Question Answering using Integrated Information Retrieval and Information Extraction. *In HLT-NAACL 2007*. Rochester, US.

[15] Dan Shen, Geert-Jan M. Kruijff, Dietrich Klakow. 2005. Exploring Syntactic Relation Patterns for Question Answering. *In IJCNLP 2005*.

[16] A. Smeaton, R. O'Donnell, and F. Kelledy. 1994. Indexing Structures Derived from Syntax in TREC-3: System description. *In TREC-3*.

[17] T. Strohman, D.Metzler, H. Turtle and W.B. Croft. 2005. Indri: A Language-model based Search Engine for Complex Queries (extended version). Technical Report IR-407, CIIR, Umass Amherst.

[18] T. Strzalkowski, L. Guthrie, J. Karlgren, J. Leistensnider, F. Lin, J. Perez-Carballo, T.Straszheim, J. Wang and J. Wilding. 1996. Natural Language information retrieval: TREC-5 report. *In TREC-5*.

[19] Dilek Hakkani-Tür and Gokhan Tur. 2007. Statistical Sentence Extraction for Information Distillation. In *ICASSP-2007*.

[20] Dilek Hakkani-Tür, Gokhan Tür and Michael Levit. 2007. Exploiting Information Extraction Annotations for Document Retrieval in Distillation Tasks. *In Interspeech 2007.*

[21] Roman Yangarber, Ralph Grishman, Pasi Tapanainen and Silja Huttunen. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction. *In COLING 2000*. Saarbruecken, Germany.

[22] Richard Zens and Hermann Ney. 2004. Improvements in Phrase-Based Statistical Machine Translation. *In HLT/NAACL 2004*. New York City, NY, US

[23] C. Zhai, X. Tong, N. Milic-Frayling, and D. Evans. 1996. Evaluation of Syntactic Phrase Indexing – CLARIT NLP track report. *In TREC-5*.

Multilingual multi-document continuously-updated social networks

Bruno Pouliquen, Ralf Steinberger & Jenya Belyaeva

European Commission – Joint Research Centre

Via Enrico Fermi 1, 21020 Ispra (VA), Italy

{Bruno.Pouliquen, Ralf.Steinberger}@jrc.it, Jenya.Belyaeva@ext.jrc.it

Abstract

We are presenting a fully-automatic live online system (accessible at http://langtech.jrc.it/SocNet) that produces monolingual or mixed-language social network graphs showing which groups of persons are being mentioned together in the world news of the last few hours. The basis for this system are name mentions extracted automatically from an average of 35,000 news articles per day in 32 languages. For any given person on the graph, hyperlinks lead to the list of text snippets and to the original texts where the person was mentioned, plus to a dedicated webpage containing additional information about this person gathered in the course of several years. For any link between persons, hyperlinks lead to the list of text snippets and to the full texts where both persons are mentioned. Building multilingual social networks that even cross writing systems (Arabic, Greek, Chinese, etc.) is made possible by exploiting the name database built up by the multilingual online NewsExplorer system (Steinberger et al. 2005), which automatically associates name variants to the same person identifier. We also discuss differences between live social networks generated from the news in different languages for the same time period.

Keywords

Social Networks, multilinguality, multi-document summarisation, Named Entity Recognition, name variant merging, visualisation.

1. Introduction

To a large extent, the factual part of news is about themes or events (taking place at certain locations at a certain time) and about persons or organisations. The news analysis system NewsExplorer (Steinberger et al. 2005; accessible at http://press.jrc.it/NewsExplorer) tries to give views of the news from the axes *events* (news clusters), *locations*, *named entities* (mainly persons and organisations) and *time* (via time lines, i.e. historical linking of news). In addition to linking news via these entities and axes, news items in NewsExplorer are also linked across languages. In this paper, we present an additional way of allowing users access to news: we present live social networks, i.e. graphs displaying groups of persons that are frequently mentioned together in the news of the last few hours and up to 1 day. Probably the most interesting aspect of the presented approach is the high multilinguality of the system (32 languages) and the fact that names are linked across languages (and writing systems) even if spelt differently and when the names have been inflected. Users can view the multi-document, multilingual and cross-language live system at the site http://langtech.jrc.it/SocNet. Additionally to the most recent multilingual social networks displayed at that site, it is also possible to produce social networks separately by language or by the country of origin of the news, as well as for documents covering a specific theme. These customised social networks are not accessible to the public, but in this paper we compare the multilingual networks with monolingual networks in four languages (section 5).

This social network generation tool takes as input the *Europe Media Monitor* (EMM; Best et al. 2005) news data and makes furthermore use of the following technology: (a) multilingual name recognition software, (b) approximate name matching software that identifies name variants for the same person, (c) multilingual language-dependent morphological name inflection generation software, and (d) network generation and visualisation software. Tools (a), (b) and (c) are part of the NewsExplorer system, which analyses news every day, links news over time (topic detection and tracking) and across languages (cross-lingual topic tracking), extracts new and known names, collects information about people and visualises the results in various ways.

The 12 co-occurrence graphs visible at the above-mentioned site are updated every two hours. Graph production starts completely anew every 24 hours at midnight so that users will always see the social network graphs of world-wide news of today. Information found in the news of all 32 languages are fully aggregated and all the results are visualised together.

Section 2 points to work with a similar focus. Section 3 summarises the text analysis technology underlying the social network generation. Section 4 focuses on the network generation, size reduction and visualisation. In Section 5, we discuss the network generation results, comparing the mixed language network with various monolingual networks for a sample 8-hour snapshot for Friday 13 July. Section 6 concludes the paper and points to future work.

2. Related Work

Due to the large volume of various types of information on the internet, there are now various applications that try to produce person profiles and to exploit similarities for various purposes (e.g. to provide focused advertising, to provide meeting forums, etc.). Some social network services like *LinkedIn* (LinkedIn 2007) or *MySpace* (MySpace 2007) build and verify online social networks, connecting registered users by different types of interests

(company, country, research interests, etc.). The features used for the linking are typically user-provided. To our knowledge, the only tools that extract the underlying linking features fully automatically are called *Connivence Maps* by *Pertinence Mining* (Connivence 2007; based on English and French news) and S*ilobreaker*, based on *Elucidon* software (Silobreaker 2007, English only), but the producers do not say how their technology works and it is not even clear whether the networks are manually edited.

For related work on individual components of the presented system (Named Entity Recognition, name variant matching, dealing with highly inflected languages, etc.), see Steinberger & Pouliquen (2007).

3. The underlying news data and text analysis technology

The social networks under discussion are extracted from live news, using resources on person names and their spelling variants. In this section, we briefly summarise where the news data comes from (section 3.1), how person names have been extracted across many languages and over years to build a name database of currently 615,000 names (3.2), how spelling variants for the same name have been gathered and merged automatically (3.3) and how morphological inflections of known names are being recognised in Balto-Slavonic and other highly inflected languages (3.4). Section 4 will then explain how this data is used to produce live social networks.

3.1 Gathering the news data

The JRC's *Europe Media Monitor* system (Best et al. 2005) gathers an average of 35,000 news article per day in 32 languages, by continuously monitoring about 1,100 public news sites from around the world for newly published information. All new articles are downloaded, converted to the standard UTF-8-encoded XML news format RSS, full-text indexed and classified according to themes and the countries mentioned in the text. The result is published in the *EMM-NewsBrief* site (http://press.jrc.it), which is updated every ten minutes.

3.2 Multilingual Named Entity Recognition

For 19 of the 32 languages, the related *EMM-NewsExplorer* application (http://press.jrc.it/NewsExplorer, Steinberger et al. 2005) clusters all articles gathered during the previous day by similarity in order to group all articles about the same subject or event. For all clusters, references to geographical places, to persons and organisations are identified, using finite state automata to recognise known names and regular expressions to recognise new names or name variants (recognition of *new* names in 14 languages only: Da, De, En, Es, Et, Fr, It, Nl, No, Pt, Ro, Sl, Sv, Tr). Sequences of uppercase words are identified as being a name if they contain known first names or if they are surrounded by empirically collected lexical patterns consisting of titles (e.g. *Minister*), words indicating nationality (e.g. *German*), age (e.g. *32-year old*), occupation (e.g. *playboy*), a significant verbal phrase (e.g. *has declared*), and more. We refer to these patterns generically as *trigger words*. Name *stop words* are used to exclude identifying frequent uppercase words (e.g. *Monday*) as part of the name. For a detailed description of this process, see Steinberger & Pouliquen (2007). The process does not make use of part-of-speech or other linguistic information in order to keep the process simple and so that it can easily be extended to many languages.

3.3 Name variant matching and merging

For all unknown names found during the daily analysis, an approximate string matching algorithm checks whether the name is likely to be a variant of a known name or whether it is a new name. New names are added to the database with a new identifier. Names found in at least five different news clusters are added to the list of known names. Periodically, a search on Wikipedia (Wikipedia 2007) is carried out to gather name translations that can be found there, as well as photographs. Wikipedia is especially useful to find name transliterations in languages using different scripts, such as Asian languages or languages using the Cyrillic, Arabic or Hellenic scripts.

The approximate string matching algorithm to compare newly found names with the 615,000 known names and their 143,000 known variants (status July 2007) is a multi-step process, details of which are described in Steinberger & Pouliquen (2007). To avoid a performance bottleneck when comparing each of several hundred new names per day with close to a million known names and name variants, we first apply a name normalisation step. Only if the normalised new name is identical with a normalised name (or any of its variants) in the database, we apply the edit distance approximate matching algorithm (Zobel & Dart 1995) to two different name representations: once to the normalised name form and once to the normalised name form with the vowels removed. If the average similarity for the new and the known name are above an empirically set threshold, the two names will be classified as variants of each other. Otherwise, the new name will be added to the database as a new name.

The name normalisation rules eliminate diacritics, reduce two neighbouring identical consonants to single consonants, unify frequent spelling variants across languages, etc. For instance, the German name-initial 'Wl' and the name-final 'ow' for Russian names (as in *Wladimir Ustinow*) will get replaced by 'Vl' and 'ov'; the Slovene 'š' and the German 'sch' will get replaced by 'sh'; French 'ou' (as in *Oustinov*) will get replaced by 'u', etc. These normalisation rules are exclusively driven by pragmatic needs and have no claim to represent any underlying linguistic concept.

An average of 400 new person names are automatically recognised as part of the NewsExplorer text analysis every day. The NewsExplorer database keeps track of all name mentions plus the list of trigger words (the titles and phrases) they are associated with.

Lang	NewsPaper	Snippet
sl	vecer	glavnega osumljenca za umor **Aleksandra Litvinenka** v Londonu postavili pred
sl	vecer	v ponedeljek zavrnil izrocitev **Andreja Lugovoja**, da bi ga kot glavnega
tr	sabah	öldürülen eski KGB ajani **Alexander Litvinenko**'nun davasi, Ingiltere-Rusya
tr	sabah	cinayetin zanlisi olarak istedigi **Andrei Lugovoy**'u Rusya'nin iade etmemesi
en	dailytimesPK	suspected of killing Kremlin critic **Alexander Litvinenko** in London last year,
en	dailytimesPK	when British prosecutors alleged that **Andrei Lugovoi** used a rare radioactive
pt	DiariodeNoticias	assassínio do ex-oficial do KGB **Alexander Litvinenko**. A revelação foi feita
pt	DiariodeNoticias	acederia ao pedido de extradição de **Andrei Lugovoi** (outro ex-agente do KGB)
en	taipeitimes	Kremlin following its refusal to extradite **Andrei Lugovoi**, the former KGB....
en	taipeitimes	KGB agent suspected of murdering **Alexander Litvinenko** last November.
en	eirepost	Lugovoi over the murder of **Alexander Litvinenko**, describing the decision
en	eirepost	Russia's refusal to extradite **Andrei Lugovoi** over the murder of Alexander
sl	delo	in nekdanjega tajnega agenta KGB **Andreja Lugovoja**. London - Britanija in
sl	delo	in ostrega Putinovega kritika **Aleksandra Litvinenka** , ki je bil nekoc prav tako
en	rian	- Russia considers the **Alexander Litvinenko** case a purely criminal matter,
en	rian	Moscow has refused to extradite **Andrei Lugovoi**, a former Kremlin bodyguard,

Table 1. Text snippets in newspapers of various languages showing both the names *Alexandre Litvinenko* and *Andrei Lugovoi*.

3.4 Dealing with morphological inflection

The current list of *known names*, i.e. the names that were found in at least five independent news clusters, consists of approximately 50,000 names plus 135,000 variants. These names can be identified in text of any language through a simple lookup procedure, i.e. no lexical patterns are required. This works well for languages with little morphological proper noun variation (e.g. most Western European languages, Arabic, Bulgarian, etc.). However, for Balto-Slavonic, Finno-Ugric and other languages, looking up the base form of a name will yield poor results as the names will not be found when they are inflected. For instance, Estonian *Bushiga* and Slovene *Bushom* are both inflections of *Bush*. Table 1 shows some morphological variants of the names *Alexander Litvinenko* (e.g. *Litvinenka*, *Litvinenko'nun*) and *Andrei Lugovoi* (*Lugovoja*, *Lugovoy'u*). As acquiring or developing morphological resources for all 32 EMM languages is out of our reach, we use relatively simple, hand-crafted paradigm expansion rules that generate – for each of the known names and their variants – a number of morphological variants. These rules, described in more detail in Pouliquen et al. (2005), either add various name endings to the same name or they substitute endings to generate a set of new endings. For the name of the Secretary-General of the Council of the European Union *Javier Solana*, for instance, we generate various inflection forms so that the strings *Javierja Solane* (sl), *Javierom Solanom* (sk), *Javierem Solaną* (pl), *Javierjem Solano* (sl), *Javiera Solany* (pl) will all be found and identified as variants of *Javier Solana*.

These morphological paradigm extension heuristics do not solve all problems, but the most frequent morphological variants can normally be captured and over-generated (wrong) variants are not harmful as they will simply not be found. For the lookup procedure, we use FLEX (Paxson 1995) to produce a finite state automaton. This tool is useful for the efficient lookup of large name lists including character-level regular expressions for suffixes, etc. It also allows looking up person names in languages that do not use white space to separate words, such as Chinese.

4. Social network generation and visualisation

The input for the work on social network generation consists of a stream of all incoming EMM news articles (32 languages), in which references to known persons have been marked up using the finite state automaton described in Section 3.4. Names not previously known are not recognised in the live system, but the list of known names is updated every day.

For efficiency purposes, we build a constantly updated index that records, for each recognised name and for each pair of names, all 300-character text snippets around the names. Table 1 shows multilingual text snippets for the names *Alexander Litvinenko* and *Andrei Lugovoi*. The index is reset at midnight every day so that it always contains the latest news and name mentions. We plan to turn this index into a 24-hour rolling window from which articles older than 24 hours will get deleted. This will give more consistency to the networks shown and will be more useful for users living in different time zones.

The index is read every two hours to update the graphs.

4.1 Turning links into a network

When two names are mentioned in the same article, a *link* is created between these two names. The more frequently the two names are mentioned, the stronger the link. In the input example shown in Table 1, the tool will thus build a

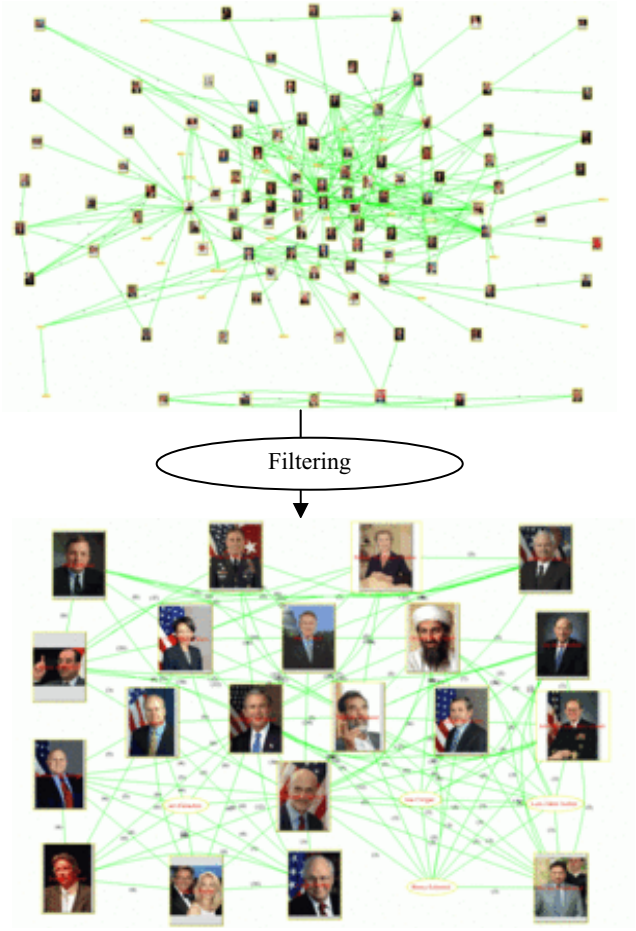

Figure 1. Example of the result of graph filtering, we retain only persons having lots of links to others.

link of weight (strength) 8 between the two persons Alexander Litvinenko and Andrei Lugovoi because they are mentioned together in eight different documents. We also have the possibility to filter graphs by language, country or subject area. This can be useful for users wanting to analyse news articles for a very specific domain. When considering only English documents, the link between Litvinenko and Lugovoi in Table 1 would be 4.

Any set or subset of links can be used to build a graph with edges (links) and vertices (persons, nodes). When considering the co-occurrence relationship under discussion, links are obviously non-directional, but in other types of relationships, it may be necessary to show the direction of relations. In the *criticism* and *support* relationship extracted by Tanev (2007), for instance, links need to be directed because such relationships are not necessarily bidirectional, and in the quotation relationship (person A refers to person B in direct speech; Pouliquen et al. 2007), users may also want to see who makes reference to whom. These more specific types of relationships and the resulting social networks are not yet incorporated in the public version of NewsExplorer.

Each sub-graph contains connected persons, but these persons may be connected indirectly, i.e. if person A is linked to person B and person B to person C, A and C will occur in the same sub-graph. The presence or non-presence of drawn edges indicates whether the link is direct or indirect.

4.2 Reducing the size of a network

The network of links can grow rather big in any given 24-hour period. For all languages together, we find approximately 50,000 links per day. In order to reduce the amount of information to the stronger links, we reduce the size by setting a threshold on the number of links required for each name pair. Depending on the number of articles, we set this threshold to between 1 (for instance for graphs fed by single languages with less articles) and 4 (for the multilingual graph fed by all 32 EMM languages).

We use a simple algorithm to divide the graphs into sets of sub-graphs, which may be connected or not. Using the previously mentioned threshold, the network of all links can be cut down into sub-graphs, i.e. graphs are automatically separated if they either have no links or if the link strength is lower than the threshold. However, if there are links above the threshold, both graphs will be joined into one.

Towards the end of a 24-hour period, the graphs are often illegibly large (see, for instance, the first graph in Figure 1). This might be an indication that 12-hour windows may be more appropriate.

For practical reasons, we display only the first 12 biggest sub-graphs. For visualization purposes, we further reduce the size of graphs if the total number of persons (nodes, or vertices) is larger than 120. We do this by deleting those persons having only one link and loop until no more vertices can be deleted. If the remaining number of edges is above 140, we remove the persons having more than two edges. If the remaining number of edges is more than 160, we delete the ones having three edges, and so on, according to the following algorithm:

```
threshold=1; min=100;
  While (numberOfVertices > min+20*threshold)
  {
    Do {
      DeleteVertexHavingLessThan(threshold);
    } until (noVertexDeleted);
    threshold += 1;
  };
```

4.3 Visualising the network

The graphs are displayed using *GraphViz* (GraphViz 2007), and more specifically the *Neato* tool which computes automatically the best place for each vertex to be displayed. We additionally make use of our database of images downloaded from Wikipedia. GraphViz allows for various output formats. We have chosen *gif + html ImageMap* as the output format, as it allows us to visualize the network on any web browser.

This simple visualisation does not allow to zoom or to change the angle of view, but it allows us to provide two types of hyperlinks: (1) When clicking on the link between two persons, the user is shown the latest text snippets in which the related persons have been mentioned

Table 2. Evaluation of person name recognition in various languages. The second column shows the number of manually identified person names.

Language	*No. of Names*	*Precision*	*Recall*	*F-1*
English	405	92	84	88
French	329	96	95	95
German	327	90	96	93
Spanish	274	85	84	84
Italian	298	92	90	91
Russian	157	81	69	74

together. Such a list of text snippets is shown in Figure 3. (2) When clicking on a node (a photograph or a person name), the user sees the most recent text snippets surrounding the name. Figure 2 shows the text snippets surrounding the name *Alexander Litvinenko* (the Russian spy who was murdered in London). This allows users to see in what context the person was mentioned. From this page, two additional hyperlinks lead to the original article where the text snippet was found and to the dedicated *NewsExplorer* person page where the user can see additional background information about this person collected over the last few years: a list of name variants, name attributes (lists of trigger words), lists of associated persons, the latest articles about this person, quotations by and about this person, and more.

5. Evaluation of the results, discussion

There is no obvious method to evaluate our automatically generated social networks quantitatively. What can be evaluated quantitatively are Precision and Recall of the Named Entity Recognition (NER) and – as we merge name variants for the same person – the results of the merging algorithm. Such results have been described in Steinberger & Pouliquen (2007) for the languages English, French, German, Spanish, Italian and Russian, repeated in Table 2 for the reader's convenience.

All persons that have been found to be mentioned together in a news article are linked somehow. The usefulness of such a link cannot be evaluated quantitatively, but is by definition of qualitative nature and depends heavily on the user interests. An example will make this clearer: In the case where one of the two linked names is that of a politician and the other that of the journalist writing the article, people who are interested in finding out about the politician may consider that the occurrence of the journalist's name is unwanted noise. People looking for the journalist, however, may very well want to know who and what the journalist writes about. For this reason, we consider correct all social network links found in the news of the day, i.e. if two persons have been mentioned together in the news, they are linked. In that sense, all social networks summarised in Table 3 are correct so that we could claim near to 100% Precision. Incorrect links would then only be those where two unrelated articles are accidentally merged into one. In order to give at least some idea of the performance and usefulness of our system, we will first discuss some errors made by the system and highlight some observations made during the analysis (Section 5.2). We will then compare the social networks across languages (5.3).

5.1 The test set

Our social networks are generated live and change continuously. For evaluation purposes, we froze a time snapshot of 7 hours and 45 minutes, starting at midnight on Friday 13 July 2007. The following numbers give an idea of the relative impact of the various languages in this time snapshot.: EMM gathered 4358 news articles in 28 languages during this time, with the most prominent languages being En: 1688, Es: 800, Pt: 274, Nl: 221, Fr: 214, De: 189, Tr: 98, Hu: 96, Da: 95, It: 91. In these articles, 12,415 name mentions of 3,417 different persons and 16,954 links between persons were found. The following number of person name mentions were found for the various languages: En: 5971, Es: 1953, Fr: 849, Pt: 659, De: 491, Nl: 443, Ar: 247, Tr: 229, Da: 209, Ro: 197, It: 190, Hu: 175, Sv: 116. The networks and links for this test set have been frozen and can be found at http://langtech.jrc.it/entities/socNet/test/last.html.

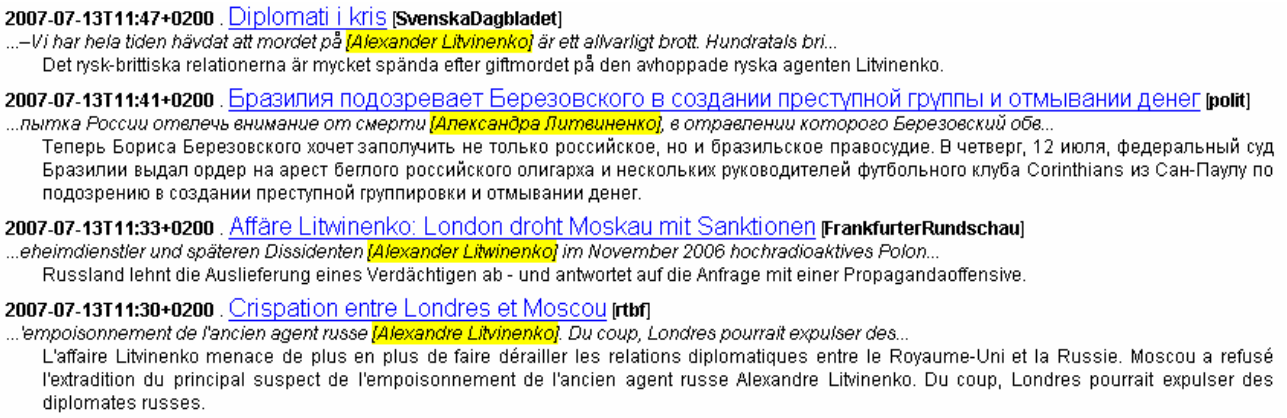

Figure 2. Clicking on a person (here *Alexander Litvinenko*), shows the context (article, description and snippet). Clicking on the document title leads to the original document. At the top of the page, another hyperlink leads to the dedicated NewsExplorer page about this person.

5.2 Observations on the extraction of links

Links between persons are either not shown because their names were not recognised in the text (NER Recall) or because they were filtered out when reducing the network size (see Section 4.2). We found two examples where the system missed important persons: the British Queen Elizabeth II was not recognised because she was only referred to by the term "The Queen", and the US president George W. Bush was missed because his known name variants were never mentioned. For disambiguation purposes, in NewsExplorer, we only recognise person names if at least two name parts (such as first and last name) are mentioned at least once in the article. In the case of the US president George Bush, his mention was not detected because he was only referred to by the strings "the Bush administration", "the President", and similar references.

In one case, *John F. Kennedy* was wrongly identified although the text referred to the airport with the same name. This disambiguation would have been difficult to solve as the word 'airport' was not mentioned: "… John F. Kennedy terror suspects …".

We also found one example of erroneous name merging: The news text referred to the Pakistani doctor *Mohammed Anif*. When following the hyperlink to the NewsExplorer page, we found that this person was merged with two other persons with the same name: a prisoner and a cricket player. In this case, the social network link was thus correct, but the additional NewsExplorer background information was partly wrong.

Another problem we came across was due to partially duplicate news articles as it happened that several identical text snippets led the system to create a strong link between two persons. The tenth English cluster and the second German cluster in Table 3 are such examples: The links are based on identical text snippets coming from two different newspapers, who even chose different titles for their stories.

Entirely or partially identical news articles are a frequent phenomenon as a small number of news agencies provide information to many newspapers. Newspapers often either copy the whole article or large parts of it. Furthermore, they sometimes publish news updates with articles that change only slightly from one to the other. In NewsExplorer, these duplicate or near-duplicate articles are automatically eliminated as part of the clustering procedure (see Section 3.2). The social network analysis, however, operates on single articles so that every duplicate will be counted as one. One possibility to reduce the impact of duplicate articles at least to some extent would be to count only those name pairs that were found in different news sources.

5.3 Social networks across languages

The page http://langtech.jrc.it/SocNet shows the live social networks as identified in world-wide news in 32 different languages. As the number of articles per language and per country differs, the relations found in news articles of some languages and countries will clearly have a bigger impact than the relations found in other languages. For this reason, we list the number of articles per language separately for each of the 12 largest live social networks as part of the graph. This gives the user an idea how much each language contributed to each graph.

English is by far the most prominent EMM-NewsExplorer language, with approximately 7,000 English language news articles from around the world per day, followed by Spanish (3000), German (2500), Dutch (2000), Portuguese (1800), etc. EMM sources are continuously updated and changed so that the relative importance of the languages can change.

News articles in some of the languages clearly come from one country (e.g. Bulgarian, Farsi and Polish news are exclusively from Bulgaria, Iran and Poland). News in some other languages, however, represents various countries: English news may be dominated by British and US-American news sources, but comes from all around the world. German news comes from Germany, Austria and Switzerland, Dutch news from Holland and Belgium, etc.

Figure 3. Context of a link: Here *Ali Larijani* and *Mohammed ElBaradei* are highlighted in articles in which they appear together.

Table 3 summarises the main contents of the twelve biggest social network of the first eight hours of Friday 13 July 2007. The first two columns show the main mixed language networks with two different link thresholds (see Section 4.2). In brackets, we show which languages mainly contributed to each of the networks. The four remaining columns of Table 3 show the main monolingual networks for the four languages English, French, German and Arabic. Where appropriate and useful, we also show the most central names of each of the networks.

Table 3 shows that some social networks are clearly related across different languages. These are mainly linked to international politics and to various types of sports. Other networks are more country-specific and are not shared across languages. Whether or not these national networks make their way into the multilingual graph largely depends on the relative impact of each language.

The networks for international politics across languages mainly show the same persons, but different people take the more central roles. The position of the persons on the GraphViz network is automatically determined depending on the number of links to other persons. For instance, in the English network for international politics, the most prominent persons are G. Bush, H. Clinton, M. McConnell, M. Chertoff, Bin Laden and S. Hussein. The same persons can broadly be found in the related Arabic network, but the most central roles are taken by M. Jamil, Bin Laden, T. Blair, G. Bush, A. al-Zawahiri and M. Abbas. In the French related network, the most central persons are G. Bush, S. Hussein, N. Al-Maliki, C. Rice and H. Clinton. Interestingly, during the evaluation period, international politics scored only second in the French network, giving place to a French network of names, whereas for all other languages international politics took the first position.

For all languages but English, the number of articles is not large enough to fill all 12 social networks (indicated by the dashes ---), at least for the 8 hour test period. The fact that the first network is usually extremely tightly-knit whereas other networks are often scarce (consisting sometimes of only two or three names) indicates that the minimum link threshold should probably differ depending on the number of names per network.

6. Conclusion - Outlook

Altogether, the live social networks provide a lot of food for thought and they do show who is in the news right now across languages and internationally. They may also show how different countries present the same themes from different angles, depending on the people they mainly mention. It goes without saying that these live social networks are not a ready-made socio-political analysis of current events across languages and countries, but that they should be seen as a tool and one of the types of input that may help analysts do their work. However, they clearly also give observers a good first impression of current events world-wide and across countries. We reckon that the user group that could reap the biggest benefit of our technology are analysts or researchers when investigating social networks produced for a selection of documents of their interest.

Work necessary to make the social networks more useful includes the adaptation of the link threshold according to the number of nodes on the graph. In addition to displaying the multilingual social networks, it would be useful to also give access to the monolingual social networks. We would like to work on improving the weight of links by boosting the link if it is fed from different sources or even from several languages and if the names are mentioned close to another in the text. We will eventually use different visualisation software and we plan to display thematic information for each social network, by either providing a list of keywords or by displaying the medoid article(s) for the documents that fed the network. Finally, we would like to experiment with graph theory algorithms to infer additional information from our graphs including cliques, walks and sub-graphs, as well as to highlight those paths that link different social networks.

7. Acknowledgement

We thank the entire EMM team and especially group leader Clive Best and chief developer Erik van der Goot for providing the valuable EMM news data and a very reliable and robust large-scale system. Martin Atkinson has started to develop a customised visualisation tool that will be used in the future.

8. References

[1] Best, Clive, Erik van der Goot, Ken Blackler, Teofilo Garcia, David Horby (2005). *Europe Media Monitor – System Description*. Report No. EUR 22173 EN.

[2] Connivence. 2007. See http://www.connivences.info/ (last visited 28.03.2007).

[3] GraphViz (2007). See http://www.graphviz.org/ (last visited 13.07.2007)

[4] LinkedIn. 2007. http://www.linkedin.com/ (last visited 28.03.2007).

[5] MySpace. 2007. http://www.myspace.com/ (last visited 28.03.2007).

[6] Paxson, Vern. 1995. *Flex – Fast Lexical Analyzer Generator*. Lawrence Berkeley Laboratory, Berkeley, CA. Available at ftp://ftp.ee.lbl.gov/flex-2.5.4.tar.gz (last visited 28.03.2007).

[7] Pouliquen Bruno, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani, Jan Žižka (2005). *Multilingual person name recognition and transliteration*. Journal CORELA. Numéros spéciaux, Le traitement lexicographique des noms propres.

[8] Pouliquen Bruno, Ralf Steinberger, Camelia Ignat & Tamara Oellinger (2006). *Building and displaying name relations using automatic unsupervised analysis of newspaper articles.* Proceedings of JADT'2006. Besançon, France.

[9] Pouliquen Bruno, Ralf Steinberger & Clive Best (2007). *Automatic detection of quotations in multilingual news.* Proceedings of RANLP'2007.

[10] Silobreaker. 2007. http://www.silobreaker.com/Corporate/ (last visited 28.03.2007).

[11] Steinberger Ralf, Bruno Pouliquen, Camelia Ignat (2005). *Navigating multilingual news collections using automatically extracted information*. Journal of Computing and Information Technology - CIT:13.4, pp. 257-264.

[12] Steinberger Ralf & Bruno Pouliquen (2007). *Cross-lingual Named Entity Recognition*. In: Satoshi Sekine & Elisabete Ranchhod (eds.). Linguisticae Investigationes LI 30:1, pp. 135-162. Special Issue *Named Entities: Recognition, Classification and Use*.

[13] Tanev Hristo (2007). *Unsupervised Learning of Social Networks from a Multiple-Source News Corpus*. Proceedings of RANLP'2007.

[14] Wikipedia. 2007. http://www.wikipedia.org/ (last visited 13.07.2007).

[15] Zobel Justin & Philip Dart, 1995. *Finding approximate matches in large lexicons*. Software – Practice and Experience, Vol. 25:3, pp. 331-345

	All (3)	**All (4)**	**En (2)**	**Fr (2)**	**De (2)**	**Ar (1)**
1	International Politics (En-Fr-Es-De-Pt): Bush, Rice, Hussein, Musharraf, Chertoff, Petraeus, Boucher, …	International Politics (En-Fr-Es-De-Pt)	International Politics (Bush, H. Clinton, McConnell, Chertoff, Bin Laden, Hussein)	French politics: Sarkozy + Strauss-Kahn + Jospin + Barroso, …	US politics + Irak (Bush + various senators + al Maliki)	International Politics: M. Jamil, Bin Laden, Blair, Bush, al-Zawahiri, Abbas, …
2	Cycling (En-Nl-De-Da-Fr)	Cycling (En-De-Fr-Nl-Da)	Cycling	International Politics: Bush + Hussein + Al-Maliki + Rice + H. Clinton	German calendar (birthdays etc. for this date)	Saad, Hariri, Amine + Pierre Gemayel, …
3	German calendar (De)	German calendar (birthdays etc. for this date) (De)	UK politics + public life: Churchill, cook J. Oliver, actor A. Hepburn)	Canadian politics	Libyan HIV-scandal: Nicolas + Cécilia Sarkozy + Ferrero-Waldner + Gaddafi	Iran politics: Rafsandjani, Khamenei, Khatami, …
4	JFK terror suspects (En): Abdul Qadir + …	JFK terror suspects (En): Abdul Qadir + …	Football + public life; D. + V. Beckham, Paris Hilton	Cycling	German politics: Stoiber + Huber + Seehofer + Beckstein	Sadr + Abdul Aziz-al-Hakim
5	Argentinean politics (Es-Pt)	Football (En-Fr-Tr-Sk-Pt-Es)	New Zealand rugby + public life	Hariri assassination: Brammertz + Eido	German-Turkish politics: Merkel + Köhler + Necdet Sezer + Kolat	Chomsky + Fisk + Miller
6	Spanish politics (Es-Ca-Ro)	Dutch historical (Hitler, Stalin, Gandhi, …) (Nl-En)	Nigerian politics	European Union + European Parliament: Solana + Pöttering	Cycling	Iran nuclear conflict: El Baradei + Larijani + Mohammed Said
7	Dutch historical (Hitler, Stalin, Gandhi, …) (Nl-En)	Argentinean politics (Es-Pt)	Cricket	French politics: Carrez + de Courzon	UK politics: Blair + Brown	Annan + Miró + …
8	Japanese politics (En)	Baseball (En)	Japanese politics	---	---	Public Life: David Beckham, J. Lennon + …
9	Iran nuclear (En-Pt-Ar-Es)	Public Life, Stars (En-Da-Pt-Sv-Pl-De)	JFK terror suspects	---	---	Pakistan politics: Musharraf + Youssef Mohamad
10	Baseball (En)	Spanish politics (Es-Ca-Ro)	Tennis	---	---	---
11	Cricket (En)	Turkish politics (Tr-Sv)	Iran nuclear conflict	---	---	---
12	Canadian politics (Fr-En)	Slovene politics (Sl)	Kosovo news + UN resolution + Russia: Lavrov + Miliband + Litvinenko + …	---	---	---

Table 3. Main subjects or actors in the 12 top social networks generated for the first eight hours of Friday 13 July 2007. The first two columns show mixed-language networks, with the languages in brackets indicating the dominant languages contributing to each of the social networks. Columns three to six show monolingual English, French, German and Arabic networks. The number in brackets in the header row indicates the threshold (minimum number of links required for links to be displayed). Dashes (---) indicate that there were no more social networks with a link strength above the threshold.

Unsupervised Learning of Social Networks from a Multiple-Source News Corpus

Hristo Tanev
European Commission
Joint Research Centre
I21020 Ispra, Italy
hristo.tanev@jrc.it

Abstract

Social Networks provide an intuitive picture of inferred relationships between entities, such as people and organizations, which allows different analyst tasks to be performed. In this paper we describe an unsupervised syntax-based algorithm for learning of Social Networks from different news sources. The algorithm performs automatic paraphrase learning and multiple-source relation extraction. We put forward a novel syntactic graph matching algorithm which facilitates the method scalability. Finally, we demonstrate that automatically acquired Social Networks may be exploited successfully for certain analyst tasks.

Keywords social network, syntactic patterns, syntactic network, relation extraction, paraphrase learning

1.Introduction

Social Networks provide an intuitive picture of inferred relationships between entities, such as people and organizations; they allow for better understanding of the social systems and make possible application of a broad range of Social Network Analysis (SNA) approaches. SNA [11] focuses on a network-based view of the interrelations between entities to identify underlying groups, communication patterns, evolution of the relations [1], and other information.

However, manual building of Social Networks is feasible only on a very limited scale. For this reason, different authors describe automatic approaches for Social Network extraction (see among the others [4] and [5])

In this paper we present an unsupervised methodology for automatic learning of Social Networks from a multiple-source syntactically parsed news corpus and evaluate this approach against a real world data set. We compiled a English-language multi-source corpus of news articles using the Europe Media Monitor (EMM) family of applications [2]. In order to overcome the efficiency problems which emerge from using syntactic information on real-world data, we put forward an efficient graph matching algorithm.

This paper is structured as follows:

In section 2 we present related work. In section 3 we describe our method. In section 4 we describe the Syntactic Network model and a pattern matching algorithm based on it. In section 5 we present evaluation. Finally, we present our conclusions in section 6.

2.Related Work

Some approaches extract Social Networks from e-mail communications, FOAF links [5] and statistical co-occurrences [4]. A common disadvantage of these approaches is that they cannot detect the type of the extracted relations: For example, statistically derived relationship can reflect friendship, kinship, co-authorship, rivalry, etc. It is not possible to classify precisely the relations without analyzing deeper the context.

On the other hand, text-based Relation Extraction [12] provides more accurate means for Social Network extraction which allow for capturing a relation even from a single mention.

A predominant supervised paradigm for Relation Extraction is Support Vector Machines (SVM) [12]. In her PhD thesis Natasha Singh [8] showed that using syntactic features from a full parser brings significant improvement in the performance of SVM-based Relation Extraction. However, the use of syntactic features with SVM brings into light different efficiency problems: The SVM methods require each pair of names appearing in one sentence to be classified separately. This, combined with the time complexity of the syntax-based kernel functions [12] restricts the scalability of SVM methods based on syntactic features.

Another method for relation extraction based on full parsing is the syntactic pattern matching: [7] describes an unsupervised approach based on syntactic paraphrases. They apply each paraphrasing pattern against each tree from the corpus. This approach becomes quite inefficient when many patterns are matched against a big corpus. Moreover, their paraphrase acquisition based on [9] uses a huge quantity of Web search engine queries, which significantly slows down the learning process.

3.Unsupervised Learning of Social Networks

In contrast to SVM which requires manually labeled data to be provided, we use an unsupervised approach for learning of syntactic patterns which requires only a minimal human input. Our method is similar to the one described in [7]. However, our algorithm differs in several points: (i) We use news clusters instead of search engine queries for paraphrase learning. (ii) We use co-reference resolution (iii) We avoid matching each template to each sentence separately. We rather, perform efficient pattern matching which maps many templates to

many structures in a pretty much "all at once" manner. (iv) We learn simultaneously different relation types and use the information from one relation type when learning the others (see point 7 of the algorithm in section 3.1).

Our unsupervised social network acquisition method has three main stages:

1. Learning of syntactic patterns which paraphrase certain predefined relations
2. Extract relations from the news corpus
3. Aggregate the relations into a Social Network.

As a case study we used the "meeting" and "support" relations between two people.

Relation "meeting" holds between two people, if they met in a certain time interval. Relation "support" holds between two people, if one supports the other or have common opinions on important matters. More detailed description of our definition about these two relations is provided in the evaluation section.

EMM news clusters. For the paraphrase learning we used the daily EMM news clusters. Each news cluster consists of the news articles obtained from many news sources in one day whose main topic is the same. For example, the news about the meeting of the heads of state of Germany, France, and Russia are clustered into one news cluster.

3.1 Learning syntactic templates

The algorithm we put forward here is inspired by the TEASE approach described in [9]. Our learning schema includes the following main steps:

1. Provide manually a very small number of *seed* syntactic templates which express the main relation. For example, for the relation "X support Y" we use the syntactic patterns $X \leftarrow subj - support - obj \rightarrow Y$ and $X \leftarrow subj - praise - obj \rightarrow Y$.
2. Match these templates against the news clusters in the corpus. Each pair of fillers of the slots X and Y is called an *anchor pair*. For example, if in the news the text "*Bush praised the Prime Minister Hamid Karzai"* appears, the algorithm will extract the anchor pair *(X:Bush; Y:Hamid Karzai)*
3. Normalize the anchor pairs using co-reference resolution and name variant detection; details of this step are discussed in section 3.3. After this step, the example anchor pair will become *(X:George W. Bush; Y:Hamid Karzai)*.
4. For each extracted anchor pair, search in the same cluster all the sentences where both names of the anchor pair occur. The assumption is that the same relation will hold between the same pairs of names in the whole news cluster, since all articles in it have the same topic.
5. From all the sentences in which at least one anchor pair appears, learn syntactic templates using our pattern-learning SyntNet GSL algorithm (see section 4.3).
6. Scoring: A template score is equal to the number of anchor pairs for which it appears.
7. Filtering: (i) Filter out all templates which appear for less than 2 anchor pairs. (ii) Accept all templates which are rooted in words which are also roots of highly scored patterns. (iii) Take out generic patterns like "X say Y", "X have Y", "X is Y", etc. using a a predefined template list (iv) Take out all templates which are included in templates with higher score in the other relation

The output of the learning algorithm is a set of syntactic templates which paraphrase the relation expressed by the seed templates or an entailment relation holds between them. For example "X meet with Y" paraphrases "X meet Y". See *T1* and *T2* on Figure 1 as examples.

3.2 Extract relations from the news corpus

We match the patterns learned in the previous step against a syntactically parsed news corpus to extract pairs of people for which certain relation holds. We use an efficient pattern matching algorithm, whose details are described in section 4.

3.3 Information aggregation and co-reference resolution

We build two Social Networks – one for "meeting" and one for "support" relations. Each pair of people is connected via an arc in one of the Social Networks, if at least one instance of the corresponding relation is detected in the text. Co-reference resolution plays an important role during information aggregation. We used the EMM database in which each article identifier is associated with a set of names detected from a Named Entity recognizer. If we detect a partial name mention (e.g. *Bush*), we try to map it to some of the full names associated to the article (e.g. "*George W. Bush*"). We used also the EMM database of name variants to merge variants of the same name in one vertex in the Social Network. We did not perform pronominal anaphora resolution.

4. Syntactic Network and Efficient Algorithms for Template Learning and Matching

Our approach uses a news corpus parsed by a dependency parser, MiniPar [3].

The parser produces from each sentence a directed graph in which the nodes represent words and the arcs - syntactic relations between them (see *G1* and *G2* in Figure 1).

In order to make feasible efficient structure mining at syntactic level, we used the *Syntactic Network* model (*SyntNet* for short) which was presented earlier in [10]. It merges all the syntactic graphs in one big graph which

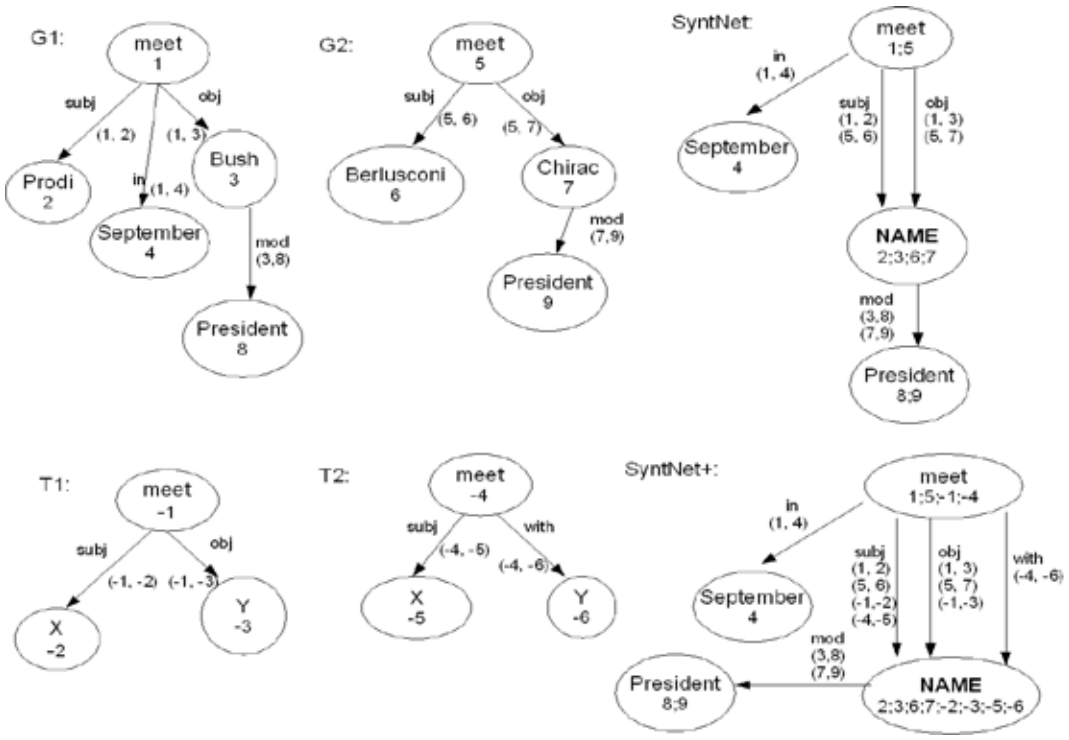

Figure 1: Merging sentences and templates in a Syntactic Network

allows for efficient searching of structures and calculation of their frequencies.

4.1 Structure of the SyntNet

Two dependency syntactic graphs *G1* and *G2* corresponding to the sentences *"Prodi met President Bush in September"* and *"Berlusconi met President Chirac"* are shown on Figure 1 together with their corresponding SyntNet.

Building the SyntNet passes through the following steps which will be illustrated using the example on Figure1:

1. All the vertices from the syntactic graphs obtain unique indices (see the numbers on the nodes of *G1* and *G2*). Arcs obtain pair of indices which show which vertices they connect. For example, the arc ***meet***-*subj*-***Prodi*** in *G1* is indexed with the index pair *(1, 2)*, because it connects vertices with indices *1* and *2*.
2. All the vertices from the corpus (from *G1* and *G2* in this example) which have the same label and do not represent person names are merged in one vertex in SyntNet having the same label. For example, vertices number *1* and *5* have the label *"meet"* and they are merged in one aggregation vertex *"meet"*in SyntNet. To show which vertices are merged in the aggregation vertex, it is indexed with the indices of these vertices. For example, the vertex *"meet"* in SyntNet has two indices: *1* and *5*, since it is obtained from the merging of vertices *1* and *5*. The indices of one vertex in SyntNet form its *index set.*
3. All the person names from the whole corpus are merged in one vertex in SyntNet - see the vertex labeled "*NAME"*. This vertex is also indexed with the corresponding index set.
4. In a similar way, we merge all arcs which are labeled equally and which connect vertices which are merged. The index sets of the aggregation arcs are sets of index pairs which represent separate arcs. In SyntNet on Figure 1 the arc ***meet***-*subj*-***NAME*** has an index set {*(1, 2)*; *(5, 6)*}, which shows that arcs ***meet***-*subj*-***Prodi*** - *(1, 2)* and ***meet***-*subj*-***Berlusconi*** - *(5, 6)* are merged in it.

A valuable property of the SyntNet is that isomorphic syntactic structures overlap, which allows for efficient tracing of common sub-structures. The worst case time complexity of building SyntNet is $O(|w| \log |w|)$, where $|w|$ is the number of the words in the parsed corpus.

4.2 Efficient matching of syntactic templates against SyntNet

Matching the templates against a corpus of parsed sentences is performed when we search for anchor pairs which fill the slots of the seed templates (see section 3.1) and during the relation extraction, when we extract pairs of related people from the corpus (see section 3.2). Taking into account that during our relation extraction experiments we had to match about 100 syntactic templates against more than a million of syntactic graphs, a straightforward approach of matching each template against each structure seemed to be quite time consuming.

We used the SyntNet graph model to substitute many to many matchings with one tracing of SyntNet: (i) First, we represent the corpus of parse graphs via a SyntNet, as it was described in the previous section. (ii) In a similar way, we add the syntactic templates to SyntNet and obtain SyntNet+ (iii) We trace SyntNet+ and using its indices, we find all paths in it which appear both in the corpus and in the templates. (iv) Finally, using the output from the previous step, we find the matching templates and the corresponding sub-structures in the corpus which they match.

Before continuing, we have to introduce some concepts. Each path in SyntNet, e.g. ***meet***-*obj*-***NAME***-*mod*-***President*** may represent a structure which has one or more occurrences in the corpus or which does not occur at all. If the path occurs in the corpus, its occurrences may be represented via their indices.

Definition 1: A path, represented by the indices of its vertices is called an *index path.*

Definition 2: If we have a corpus *C* and a SyntNet *SN* which represents it, each path *p* in a *SN* which occurs in *C* has a set of *corresponding index paths*, each representing one occurrence of *p* in *C*. The set of corresponding index paths of *p* is denoted by *ip(p).*

As an example, consider the path p_1=***meet***-*obj*-***NAME***-*mod*-***President*** in SyntNet on Figure 1. Its $ip(p_1)$ contains two corresponding paths ***1***-*obj*-***3***-*mod*-***8*** and ***5***-*obj*-***7***-*mod*-***9***. Each of them represents one occurrence of the path p_1. On the other hand, the SyntNet path p_2=***meet***-*subj*-***NAME***-*mod*-***President*** has an empty set $ip(p_2)$, since it does not occur in the corpus.

We will illustrate the efficient template matching algorithm with an example in which we match a pair of templates *(T1: X meet Y* and T2: *X meet with Y)* against the SyntNet on Figure 1.

1. We index the vertices of both templates with unique negative indices to differentiate from the positive indices of the corpus
2. We merge *T1* and *T2* with SyntNet. Slots *X* and *Y* are merged with the "*NAME"* vertex. The result SyntNet we call SyntNet+ (see Figure 1)
3. In SyntNet+ we identify the vertices which represent template roots (only the vertex *"meet"* in this example) and from these roots we traverse all this paths in SyntNet+ which have corresponding index paths with positive and negative indices. In this way we find these paths which appear both in the templates and in the corpus. We use a sub-algorithm called *Index Tracing* to accomplish this task, whose details will be explained later.

   Let's consider the example on Figure 1: The Index Tracing will find two paths in SyntNet+ - ***meet***-*subj*-***NAME*** with four corresponding index paths: ***1***-*subj*-***2***, ***5***-*subj*- ***6***, ***(-1)***-*subj*-***(-2)***, and ***(-4)***-*subj*-***(-5)*** and ***meet***-*obj*-***NAME*** with three corresponding index paths: ***1***–*obj*-***3***, ***5***–*obj*-*7*, and ***(-1)***-*obj*-***(-3)***. Each index path with negative indices refers to the templates *T1* and *T2* and each index path with positive indices refers to the corpus graphs *G1* and *G2*. For example, the index path ***1***-*subj*-***2*** refers to the occurrence of ***meet***-*subj*-***NAME*** in *G1*, while ***(-1)***-*subj*-***(-2)*** refers to the occurrence of the same path in the template *T1*. If a SyntNet+ path has corresponding index paths with positive and negative indices, as it is the case in this example, it occurs both in the templates and in the corpus.
4. Using all the SyntNet+ paths and corresponding index paths collected during the Index Tracing, we identify templates which completely match in the corpus and the vertices from the corpus where they match. This step of the algorithm is divided into several sub-steps:

   4.1 Each template *t* is decomposed into set of paths *P(t),* each of which has as an initial vertex the template root and its terminal vertex is a leaf. Following the example on Figure 1, *P(T1)* consists of two paths - ***meet***-*subj*-***X*** and ***meet***-*obj*-***Y*** which are converted to ***meet***-*subj*-***NAME*** and ***meet***-*obj*-***NAME***

   4.2 If all the paths from *P(t)* are found by Index Tracing, consider the template *t* for further processing, otherwise it can not be matched in the corpus.

   4.3 If template *t* is to be considered, for each path *p* from *P(t)*, we consider the set of corresponding index paths *ip(p)*, found by the Index Tracing. For example, for the path *p*=***meet***-*subj*-***NAME*** from *P(T1)*, *ip(p)* = {***1***–*subj*-***2***, ***5*** –*subj*-***6***, ***(-1)*** -*subj*- ***(-2)***, ***(-4)***-*subj*-***(-5)***}.

   4.4 For each path *p* from *P(t)* we take its *ip(p)* and form a set, denoted by *ini(p),* which consists of all the initial vertices of the paths from *ip(p)*. For example, *ini(* ***meet***-*sub*-***NAME)***={***1***, ***5***, ***-1, -4***}. It is important to note that **the positive indices in *ini(p)* refer to these vertices in the corpus from which an occurrence of the path *p* begins**.

   4.5 For each template *t* under consideration, we find *rootset(t)* as an intersection of the sets of initial indices:

$$rootset(t) = \bigcap_{p \in P(t)} ini(p)$$

   For example:

   *rootset(T1)=ini(****meet****-subj-****NAME****)*$\bigcap$*ini(****meet****-obj-****NAME****)*={**1**, **5**, **-1**, **-4**} $\bigcap$ {**1**, **5**, **-1**}={**1**, **5**, **-1**}

It is important to note that the positive indices in *rootset(t)* refer to these vertices in the corpus in which occurrences of all the paths from *P(t)* begin. **Therefore, each positive index in *rootset(t)* refers to the root of one occurrence of the template *t.***

5. For each template *t* and each root of its occurrence taken from *rootset(t)*, we may trace all the vertices of this occurrence. To do this, we use the corresponding index paths computed by the Index Tracing.
6. In each occurrence of *t,* a pair of names matches the slots of the template. We find all such name pairs from all the occurrences of *t* and return them as a final result. For each such a name pair, the semantic relation (e.g. "meeting") expressed by *t* holds. In the example on Figure 1 the template T1 will match the pairs ("Prodi", "Bush") and ("Berlusconi", "Chirac").

Index Tracing. In its general form, this sub-algorithm takes on its input a SyntNet and a vertex *v* from the SyntNet, e.g. the vertex "meet" from SyntNet on Figure 1. The Index Tracing finds all the directed paths in a SyntNet representation for which this vertex *v* is initial[1]. For each such a path p, the algorithm finds its corresponding index paths *ip(p).* We will explain the Index Tracing basic steps using as an example SyntNet on Figure 1:

1. Initialization: *CurrentPathSet={v}*. *ip(v)* is initialized with the index set of *v* (see section 4.1), for example *ip(**meet**)*={1,5}. The mapping between the single-vertex path *v* and *ip(v)* is memorized in a hash table *IP.*
2. For each path *p* from *CurrentPathSet*, find an arc *a* from SyntNet, which can be added to *p,* such that a new SyntNet path *pn* is obtained.

   As an example, let's assume *p=**meet**-obj-**NAME***. If we take the arc *a=**NAME**-mod-**President***, a new path can be formed *pn= **meet**-obj-**NAME**-mod-**President***.
3. The set of corresponding index paths *ip(pn)* is computed from *ip(p)* by expanding some of the index paths from *ip(p)* using the index set of the arc *a*. Index paths from *ip(p)* which cannot be expanded are not included in *ip(pn).*

   Considering the values of *p*, *pn*, and *a* given in the example from the previous algorithm step, we can easily see that *ip(p)*={***1**-obj-**3***, ***5**-obj-**7***}. The index path ***1**-obj-**3*** can be expanded using the index pair *(3,8)* from the index set of the arc *a*. The result will be a new index path ***1**-obj-**3**-mod-**8*** which belongs to *ip(pn)*. In a similar way we expand the other index path from *ip(p) - **5**-obj-**7*** and obtain ***5**-obj-**7**-mod-**9***. *In this way we find that ip(pn)*={***1**-obj-**3**-mod-**8***, ***5**-obj-**7**-mod-**9***}.
4. If *ip(pn)* is not empty, *pn* is added to *CurrentPathSet* and *ip(pn)* is memorized in a hash table *IP*, in which each path p_x is mapped to its $ip(p_x)$.
5. If expanding the paths from *CurrentPathSet* is still possible, go to 2.
6. Finally, return *CurrentPathSet* and the hash table *IP.*

[1] In the Index Tracing version used in SyntNet GSL, paths which finish in a certain vertex are built

4.3 SyntNet General Structure Learning (GSL)

GSL was described in [9]. From a set of anchor pairs and a parsed corpus it learns the most frequent and most general templates (see [9] for more formal description). We combined the idea of the GSL and the SyntNet and created a powerful and efficient algorithm for learning of general syntactic templates. SyntNet GSL uses the Index Tracing sub-algorithm to discover efficiently repeating sub-structures. Due to space limitations, we cannot give the details of this algorithm here.

4.4 Time complexity

If we assume that the templates are syntactic trees, but not imposing restrictions on the structure of parse graphs from the corpus, it can be shown that the worst case time complexity of the syntactic matching algorithm is bounded by $O((|s|+|t|)\ (\log MaxArcO)^{a-1})$, where *a* is the maximal number of leaves of a template, $|s|$ is the number of the sentences in the corpus, $|t|$ is the number of the templates, and the *MaxArcO* is the maximum number of occurrences of an SyntNet arc (e.g. ***meet**-obj-**NAME***) in the corpus . When we use templates with two anchors, in most of the cases we have *a=2*. Considering this, the worst case time complexity becomes closer to $O((|s|+|t|)\ \log MaxArcO)$. Note that in general this estimate is better than $\Omega(|s|.|t|)$, which is the time complexity estimate of each method which matches each template against each sentence. It can be shown also that the worst-case time complexity of SyntNet GSL is $O(|s|.\log|s|)$.

5. Experiments and Evaluation

5.1 Evaluation schema

Social Networks provide higher level view of the relations between entities, such as people. They do not represent the individual events which motivate the relations. In our evaluation schema we assume that a relation "meeting" or "support" holds between two people X and Y, if at least one corresponding event took place in a certain time interval, the event involved both X and Y, and it is reflected in the news from the test corpus.

More formally, we consider the following definitions:
Definition 3: We say that "a meeting event" involves X and Y, if X and Y met physically and had some kind of important conversation.

Definition 4: We say that "a support event" involves X and Y, if one of the three event types took place: (i) X expressed support or positive attitude for Y or vice versa; (ii) X and Y reached some agreement with mutual benefits; (iii) X and Y have very similar attitudes towards important problems.

Our test corpus contains news articles from the period 03/Oct/2006 – 31/Oct/2006. It mostly covers events which took place in this period or shortly before or after that. We noticed that many news refer to events which happened the previous or the next week. Therefore, we decided to consider these events which are mentioned in the test corpus and which took place or at least overlap with the period which begins one week before 03/Oct/2006 and which finishes one week after in 31/Oct/2006.

Now we can define more formally "meeting" and "support" relations:

1. If at least one "meeting event" is taken into consideration and it involves X and Y, then we say that a "meeting relation" holds between X and Y.
2. If at least one "support event" is taken into consideration and it involves X and Y, then we say that a "support relation" holds between X and Y. For the purpose of the evaluation, we consider the "support" relation to be symmetric.

Note that in the Social Network many mentions of the same event and even the different events of the same type between the same pair of people map into the same relation. This is relevant especially in cases when the considered time interval is limited; in our case it is 28 days. In this clue, we evaluated the capability of our Social Network extraction algorithm to detect "meeting" and "support" relations. It was out of the scope of this evaluation to measure the performance of our system with respect to the detection of the individual events which motivate the relation.

5.2 Relation Extraction experiments

For paraphrase learning we used a training corpus of 98'000 English-language news articles clustered in 22'000 EMM topic clusters published in the period 01/May/2006 – 03/Oct/2006.

For testing the method, we used 125'000 English-language news articles published in the period 03/Oct/2006 – 31/Oct/2006.

Both corpora were parsed with MiniPar [3] and Syntactic Networks were built from the parsed sentences.

For the unsupervised paraphrase learning we used two seed syntactic templates for the "support" relation - "**X** *support* **Y**" and "**X** *praise* **Y**" and one seed template for the "meeting relation" - "**X** *meet* **Y**". (We show the linear forms of the templates, in reality they are syntactic graphs as shown in Figure 1.)

Using the template learning algorithm described in section 3.1, we acquired automatically from the training corpus 6 more syntactic paraphrases for the "support" relation and 92 for the "meeting" relation.

As an example, we give the linear form of some of the "support" templates - "**X** *thank* **Y**", "**X** *welcome* **Y**", "**X** *hail* **Y**" and some of the "meeting" templates - "**X** *tell* **Y**", "**X** *and* **Y** *hold talks*", " **X** *and* **Y** *said*", "*meeting between* **X** *and* **Y**", "*discussion between* **X** *and* **Y**", "*statement by* **X** *and* **Y**", "**X** *have dinner with* **Y**", "**X** *give details about the talk with* **Y**".

Next, using our syntactic pattern matching algorithm and both the seeds and the acquired paraphrases, we extracted from the test corpus 1'670 mentions of "support" and "meeting" events. For the purpose of the evaluation we considered only the relations between the top 33 VIP - Very Important Persons. The list of VIP was obtained by taking the people which appear most frequently in the English news in the period 01/May/2006 – 31/Oct/2006. In such a way we took an evaluation sample from 152 extracted event mentions out of the total of 1'670. Considering the evaluation schema described before, these 152 event mentions map into 40 relations between VIP, namely 33 "meeting" relations and 7 "support" relations.

In order to measure the performance of our social network extraction algorithm we identified manually all the "meeting" and "support" relations in the test corpus which hold between the 33 VIP under consideration and which are explicitly mentioned in at least one sentence. We found 75 VIP relations – 36 "meeting" and 39 "support". In Table 1 we show the performance of our system for both type of relations when using both seeds and paraphrases.

Table 1 Performance of the Social Network extraction algorithm with paraphrases

	Precision	**Recall**	**F1**
meeting	0.606	0.556	0.580
support	0.571	0.103	0.174
overall	0.6	0.32	0.417

We measured the impact of the paraphrase learning by comparing our system's performance against a baseline relation extraction which uses only the seed templates. The results of the baseline are shown in Table 2.

Table 2 Social Network extraction with only seeds

	Precision	**Recall**	**F1**
meeting	0.818	0.25	0.383
support	0.5	0.026	0.049
overall	0.769	0.133	0.227

The overall improvement of the F1 measure when using the paraphrases is 0.19 with respect to the baseline. As expected, overall precision of the baseline is better (+0.169), since the seed templates are selected manually. It is interesting, however, that the precision for the "support" relation increases when using paraphrases.

When we add paraphrases, the overall recall increases by 0.187. For the "meeting" relation, it increases the recall twice (from 0.25 to 0.556) and for the "support" relation the effect of adding paraphrases is even more dramatic – about four times improvement (from 0.026 to 0.103).

All these numbers demonstrate that automatic paraphrase acquisition is important for relation extraction and significantly improves its performance.

Unfortunately, it was not possible to compare properly our experiments with others described in the literature, since the data and the evaluation settings were completely different from ours. However, it is worthwhile mentioning than the unsupervised approach described in [7] reports F1 in the range 0.28-0.34.

Efficiency. The whole process of matching 101 syntactic patterns against about 1'080'000 parsed sentences took about 9 minutes and 48 seconds on a PC with 2.8GHz processor Pentium 4 and 2GBytes of RAM, running Windows XP Professional SP2. It took *9 minutes and 5 seconds* to read in the memory the templates, the parsed sentences and to build the Syntactic Network. After SyntNet was built, it took only *43 seconds* to perform the rest of the pattern matching.

5.3 Using the network view

The information in the automatically extracted VIP Social Network can be used to analyze better the importance of the VIP. We run the PageRank algorithm [6] on the automatically extracted "meeting" network and found the top 5 ranked people. When using PageRank, we use names instead of pages and relations between people, instead of page references. We compared our ranking with a frequency-based method which ranks the people according to the number of articles from the test corpus in which they appear. The comparison of both methods is shown in Table 3.

Table 3 Top ranked VIP using page rank and frequency

	PageRank	**Frequency ranking**
1	C. Rice	G.W. Bush
2	G.W. Bush	T. Blair
3	V. Putin	C. Rice
4	E. Olmert	N. al-Maliki
5	T. Blair	S. Hussein

The importance of a person is highly subjective and it is difficult to be evaluated formally. However, intuitively it seems that PageRank works better: It returns in the top 5 a list of very important heads of states and political figures, while the frequency ranking misses some important politicians from the first list and includes "Saddam Hussein" who was much on the news titles in this period but had relatively small importance for the political world in the considered period.

6. Conclusions

We described an unsupervised syntactic approach for learning of Social Networks from news clusters. It uses two novel and efficient algorithms for syntactic template learning and matching which make our solution potentially more scalable than other syntax-based approaches.

There is still space for improvement of our method. However, the experiments show that we can already use it for some analyst tasks.

Our approach uses syntactic templates which have well defined semantics, therefore the list of templates can be manually edited by a human expert which may improve the results reported in Table 1.

A Social Network provides an alternative view to the news topics. We used this view to measure better the importance of a person. Other interesting applications of the Social Networks include detection of groups of people and evolution of the relations inside and between the groups. Relations between people encoded in a Social Network may also be used to detect relations between news topics and navigate through news collections.

Considering these and other possible applications, we think that automatic large-scale acquisition of Social Networks will become more important in the future. In this clue, exploitation of effective and efficient algorithms for Relation Extraction is becoming an important issue.

7.References

[1] A.-L. Barabási: Linked: The New Science of Networks. Perseus Publishing, Cambridge, MA, 2002.

[2] C. Best, B. Pouliquen, R. Steinberger, E. van der Goot, K. Blackler, F. Fuart, T. Oellingen, and C. Ignat: Towards Automatic Event Tracking, ISI, 2006.

[3] D. Lin: Dependency-based Evaluation of MiniPar, Workshop on the Evaluation of Parsing Systems, 1998

[4] Y Matsuo, J Mori, M Hamasaki, and K Ishida: POLYPHONET: An Advanced Social Network Extraction System from the Web, Proceedings of WWW conference, 2006

[5] P.Mika: Flink: Semantic Web Technologies for the Extraction and Analysis of Social Networks, Journal of Web Semantics 2005

[6] L.Page, S.Brin, R.Motwani, and T.Winograd: The pagerank citation ranking: Bringing order to the Web, Standford publications, 1999

[7] L. Romano, M. Kouylekov, I. Szpektor, I. Dagan, and A. Lavelli: Investigating a Generic Paraphrase-based approach for Relation Extraction, EACL, Trento, Italy, 2006

[8] N. Singh: The Use of Syntactic Structure in Relation Extraction, Ph.D. Thesis, 2004

[9] I. Szpektor, H. Tanev, I. Dagan, and B. Coppola: Scaling Web Based Acquisition of Entailment Relations, EMNLP, 2004

[10] H. Tanev and B. Magnini: Weakly Supervised Approaches for Ontology Population, EACL, Trento, 2006

[11] S. Wasserman and K. Faust: Social Network Analysis: Methods and Applications. Cambridge University Press, 1994

[12] D. Zelenko, C. Aone, A. Richardella: Kernel Methods for Relation Extraction, Journal of Machine Learning Research, vol.3, 2007

Combining Information about Epidemic Threats from Multiple Sources

Roman Yangarber*, Clive Best†, Peter von Etter*, Flavio Fuart†, David Horby†, Ralf Steinberger†

∗ Department of Computer Science
University of Helsinki, Finland
first.last@cs.helsinki.fi

† European Commission – Joint Research Centre
Ispra, Italy
first.last@jrc.it

Abstract

This paper describes an on-going effort to combine Information Retrieval (IR) and Information Extraction (IE) technologies, to leverage the benefits provided by both approaches to add value for the end-user, as compared with IR or IE in isolation. The main aim of the combined system is to pool together information from multiple sources to improve the quality of results. On one hand, multiple mentions of the same event or related events should be presented in a coherent fashion. On the other hand, grouping related events should improve the system's confidence in the discovered facts. We describe our approach and the results achieved in the project to date.

1 Introduction

The ability to obtain timely medical information from digital sources is essential for surveillance of diseases and epidemics. It directly impacts the work of health authorities and epidemiologists throughout the world. In this paper we present preliminary results from a project that aims to build a system for monitoring disease epidemics by analyzing textual information, mostly in the form of news, available on the Web. The system is built on top of two major components—MedISys, based mostly on IR technology, and PULS, the information extraction system.

We describe the system responsible for the IR component in section 2; in section 3 we describe the medical IE system, together with the heuristics it implements; section 4 describes how the two systems are integrated. In section 4.2 we present some quantitative measures of performance of the combined system. In conclusion, we discuss directions of on-going work.

2 Medical Information System: MedISys

The *Medical Information System*, MedISys, is an automatic tool that gathers reports concerning Public Health from many Internet sources world-wide in multiple languages, classifies them according to hundreds of categories, detects trends across categories and languages, and notifies users. The publicly accessible MedISys site *http://medusa.jrc.it/* presents a quantitative summary of latest reports on a variety of diseases and disease sub-types (e.g., respiratory infections), on bioterrorism-related issues, toxins, bacteria (e.g., anthrax), viral hemorrhagic fevers (e.g., Ebola), viruses, medicines, water contaminations, animal diseases, Public Health organisations, etc. At a second, password-restricted site, EU staff and national Public Health officials get access to an even larger variety of subject classes, such as news on additional diseases, on nuclear or chemical contamination, etc. Furthermore, users of the restricted site can see pay-for newswires, get access to mapping tools, and subscribe to automatically generated daily reports on various themes.

The development of MedISys was initiated by the European Commissions (EC) Directorate General Health and Consumer Affairs (DG SANCO) for the purpose of supporting national and international Public Health institutions in their work on monitoring health-related issues of public concern, such as outbreaks of communicable diseases, bioterrorism, large-scale chemical incidents, etc.

MedISys is an automatic alternative to an otherwise time-consuming and tedious manual process. Typically, employees of national Public Health organisations look through their national press to identify reports on disease outbreaks and other Public Health issues and summarise the situation or scan the docu-

ments. The usage of MedISys saves these users time, and additionally gives them access to more news reports in more languages.

MedISys currently monitors news articles from about 1100 news portals around the world in 32 languages, from commercial news providers including 25 news agencies, LexisNexis, and from about 150 specialised Public Health sites. The system categorises all documents according to about 200 classes of predefined health threats. It uses statistical procedures to detect a sudden increase of articles in any of the classes, and visualises the trends graphically. Users can access documents and the automatically derived meta-information via Web pages, RSS feeds, through daily email alerts and summary reports, and via automatically generated SMS messages.

MedISys is part of the Europe Media Monitor (EMM) product family, developed at the EC's Joint Research Centre (JRC), which also includes NewsBrief,[1] a live news aggregation system, and NewsExplorer,[2] a news summary and analysis system [5]. The following sections cover the functionality of MedISys in more detail.

2.1 Document Gathering and Format Standardisation

MedISys ingests all EMM documents, i.e. the newswires provided by major news agencies, plus the approximately 35,000 articles per day, in 32 languages found on about 1100 news portals and 150 Public Health sites. The monitored sources were selected strategically with the aim of covering all major European news portals, plus key news sites from around the world, in order to achieve good geographical coverage. Additionally, individual users can request the inclusion of further news sources, such as all local newspapers of their country, but these user-specific sources are processed separately in order to guarantee the balance of news sources and their types across languages.

Where available, EMM (and thus also MedISys) collects RSS feeds. RSS stands for "Really Simple Syndication" and is an XML format with standardised tags used widely for the dissemination of news and other documents. For all other source sites, scraper software firstly looks for links on pre-defined web pages and downloads the pages linked to. As news pages do not only contain the news article, but also menus, related news, advertising, information about other sections of the newspaper, and other non-news-related information, the main news article is extracted from each web page in a three-step process:

1. clean the HTML by removing Java script, non-standard tags and unnecessary tags,

2. convert the HTML code to XHTML, which includes repairing incorrect HTML code, and

3. convert XHTML to RSS format, using an XSLT transformation that needs to be produced manually and separately for each news site.

For details, see [1]. The result is a standardised document format in UTF-8 encoding that allows common processing of all texts. Information about the document's language, source country, download time and place are preserved as meta-data.

2.2 Document routing and classification

EMM allows the selection of articles about any subject using either Boolean combinations of search words or lists of search words with positive or negative weights, and the setting of an acceptance threshold. It is possible to require that search words occur within a certain proximity (number of words) and to use wild cards (single letter and word-final Kleene star). In EMM, each such subject definition is called an *alert*. EMM alerts are multilingual, i.e., search-word combinations may mix languages. In addition to the generic alerts pre-defined by the EMM team, users may create their own subject-specific alert definitions. Users are responsible for the accuracy and completeness of their own alerts.

A dedicated algorithm was developed at the JRC that allows the system to scan incoming articles for hundreds or thousands of alert definitions in real time. Information about the alerts found in each article is added to the RSS file. EMM NewsBrief has approximately 600 different alert definitions, including one for each country of the world (consisting mainly of the country name, and the name of the major city or cities). More fine-grained geo-coding and disambiguation are carried out downstream in the EMM NewsExplorer application, see [4].

The medical alerts in MedISys differ from the generic EMM NewsBrief alerts. In addition to the country-based alerts, MedISys employs hundreds of health-specific alert definitions. MedISys alerts are organised into a hierarchy of classes, such as Communicable Diseases, Medicines and Labs, Organisations, Bioterrorism, Tobacco, Environmental & Food, Radiological & Nuclear, Chemical, etc., each containing finer sub-groups. Figure 1 shows the entry page of MedISys with part of its menu structure exposed (on the left and bottom-left), and a trend visualisation graph (upper-middle box).

[1] http://press.jrc.it
[2] http://press.jrc.it/NewsExplorer

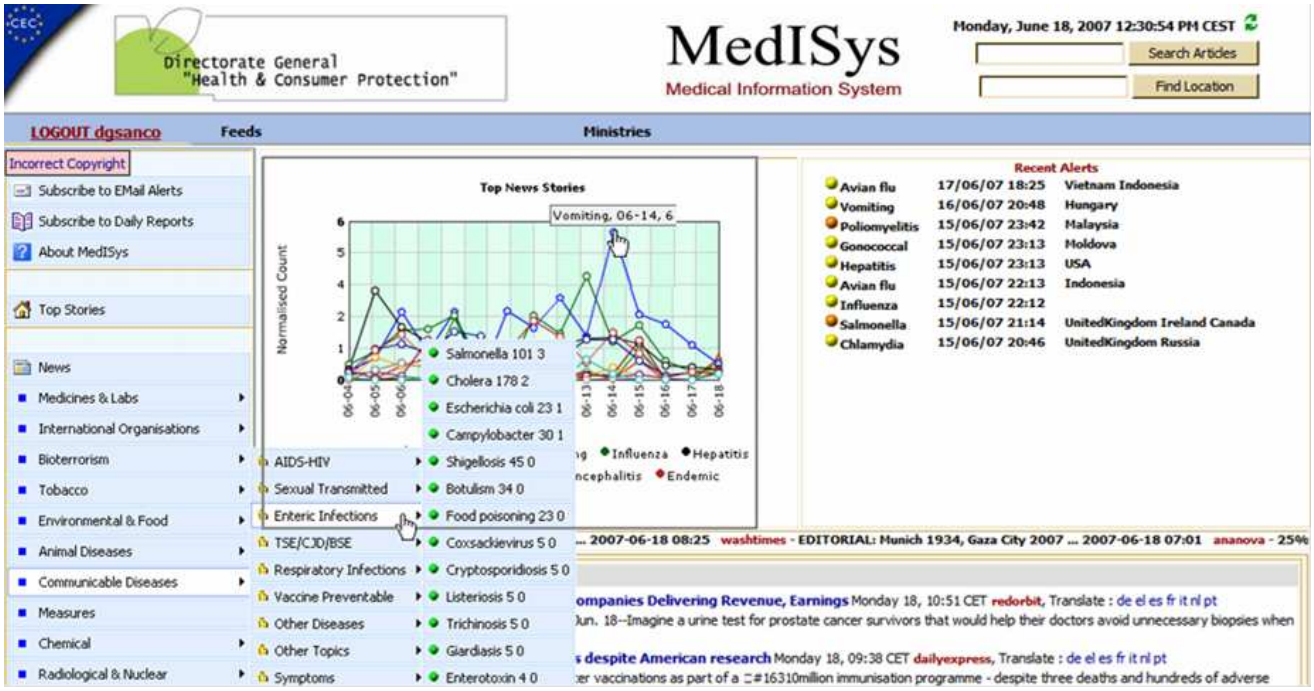

Fig. 1: *Medisys main page (restricted site).*

2.3 Multilingual multi-document trend detection

The alert definitions in MedISys are multilingual, so that the mention of a disease or symptom in the news in any of the languages can be identified. The MedISys software keeps a running count of all disease alerts for any country of the world, i.e., it maintains a count of all documents mentioning both a certain country and a given disease over a fixed time window—a period of two weeks. An alerting function detects a sudden increase in the number of reports for a given disease and country by comparing the statistics for the last 24 hours with the two-week daily rolling average. It uses the Poisson distribution, which is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate.

Figure 2 shows how the intersection of country alerts with disease alerts in combination with the trend analysis can be used to alert users to a potential health threat. This screenshot, from the public MedISys site (*http://medusa.jrc.it/*), shows increasing reporting on dengue fever related to several South-East Asian countries (highlighted on map).

2.4 Users and usage of MedISys

Customers of MedISys can use the Web interface to view the latest trends and access articles about diseases and countries. However, they can also opt to receive instant email reports, or daily summaries regarding pre-selected diseases or countries, for their own choice of languages. Specific registered users can also be granted access to the JRCs *Rapid News Service*—RNS, which additionally allows to filter news from selected sources or countries, and which provides functionality to quickly edit and publish newsletters and to distribute them via email or to mobile phones. MedISys displays the title and the first few words of each article, plus a link to the URL containing the full text.

MedISys users include the European Commission, the World Health Organisation (WHO), the Canadian Global Public Health Intelligence Network (GPHIN), the European Centre for Disease Control (ECDC) and the US CDC, the French Institut de Veille Sanitaire (INVS), the Spanish Instituto de Salud Carlos III, and other national authorities.

3 Extracting Facts about Epidemics

MedISys has proved to be a useful and an effective tool, with thousands of users accessing it daily. In considering possible extensions that would add further value, a natural choice falls on IE technology:

- IE could deliver information concerning specific incidents of the diseases tracked by MedISys, whereas IR is able to return entire matched documents (along with an indication which *alerts* fired within the document).
- IE could boost precision, since keyword-based queries may trigger on documents which are off-

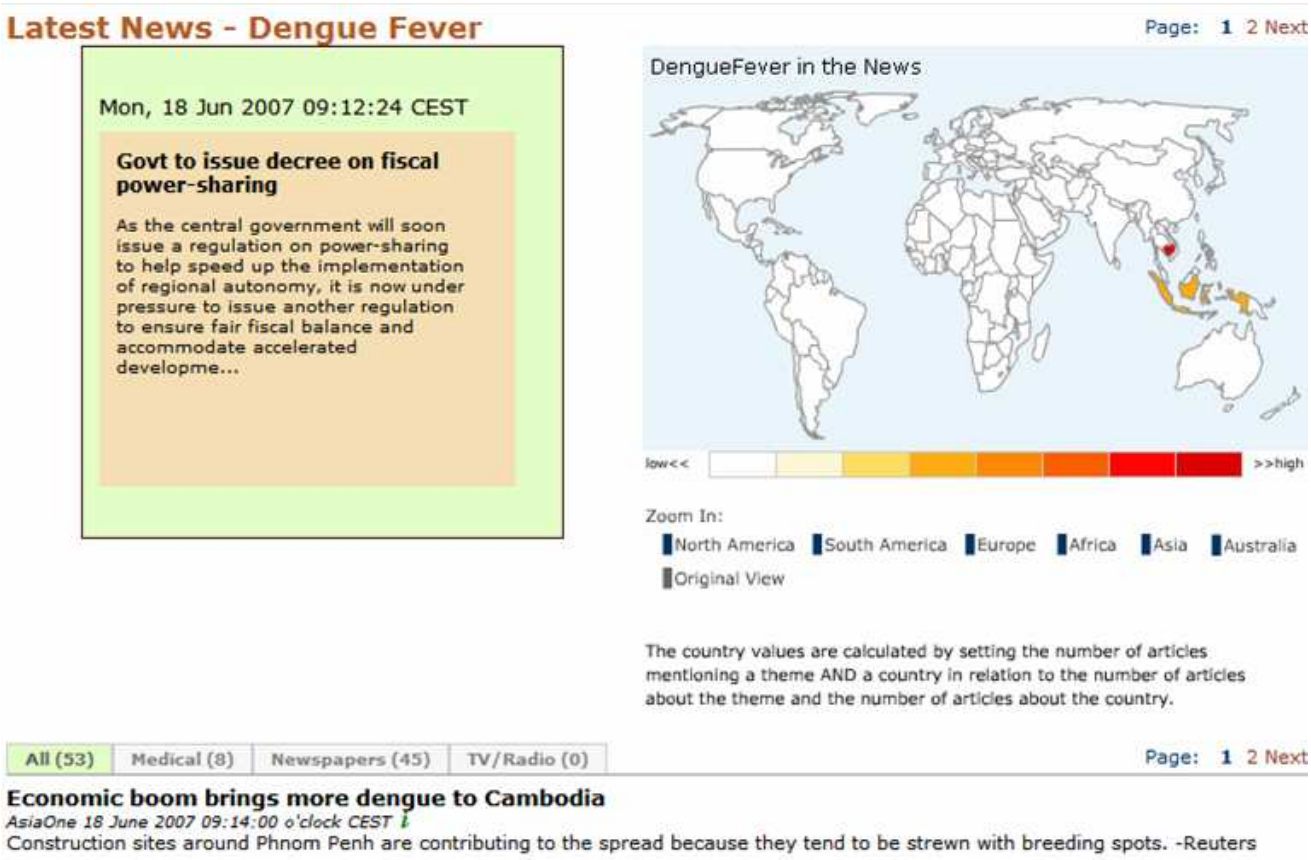

Fig. 2: *Dengue alerts with geographic distribution.*

topic but happen to mention the *alerts* in unrelated contexts. Pattern matching in IE provides the mechanism that assure that the keywords appear in relevant contexts only.

PULS, the *Pattern-based Understanding and Learning System*, is developed at the University of Helsinki to extract factual information from epidemiological reports.[3] Previously, PULS had been applied to two dedicated sources of epidemiological reports—ProMED-Mail,[4] and WHO epidemic and pandemic alert and response.[5]

We next briefly describe the key functionality provided by PULS for epidemics-related texts.

3.1 Medical Information Extraction

For each document, the IE system extracts a set of *incidents* reported in the text. An incident is an event involving some communicable disease, described in plain text. An incident is described by a set of attributes: the name of the disease, the location and country of the incident, the date of the incident, and descriptive information about the victims—their type (people, animals, etc.), their number, whether they survived, etc. The system also identifies events in which the disease is reported as *unknown*, or undiagnosed, which are especially crucial for surveillance.

For example, for the sentence "*Five people were reported to have contracted Ebola in Uganda last week,*" the system will assign the underlined values to the corresponding attributes, and create a record in a relational database. Each record extracted from the document is permanently stored, together with links to the exact offsets in the text where its attributes were found within the document.

Figure 3 presents a view of the database, as it appears on the Web site. This collection of rows was returned in response to a user "query", which is specified by constraints on some of the attribute columns. This table was constrained by publication date (April 2007), disease (avian influenza) and country (Indonesia or Cambodia). The constraints are typed into the text boxes below the column names. (Rows are ordered by publication date, by default.) Blue rows in the table correspond to confident events (defined below in section 4.1), and white rows are non-confident.

Figure 4 shows a MedISys document which generated an event (corresponding to the sixth line from the bottom of the table in Figure 3). The values of the attributes of the event are shown in the box on the right.

[3] http://doremi.cs.helsinki.fi/jrc
[4] http://www.promedmail.org
[5] http://www.who.int/csr/don/en/

[Database list] [Confident events] Events [Advanced query] [Outbreaks] [Reset page]
Viewing 248 events in 240678 documents

Published	Source	Disease	Begin	End	Country	Total	Status	Descriptor
2007.04		Avian Influenza			Indonesia|C			
2007.04.24	globalsecurity	Avian Influenza	2007.04.23	2007.04.23	Cambodia	172	†	Human Bird Flu Deaths
2007.04.24	globalsecurity	Avian Influenza	2007.01	2007.01	Indonesia	34		human cases
2007.04.24	globalsecurity	Avian Influenza	2003	2007	Indonesia	81		81 avian flu cases
2007.04.24	globalsecurity	Avian Influenza	2007.04.23	2007.04.23	Indonesia	--		two new human cases
2007.04.24	globalsecurity	Avian Influenza	2003	2007	Indonesia	63	†	63 deaths
2007.04.21	cidrap	Avian Influenza	2005.05	2005.05	Indonesia	291		291 cases
2007.04.21	cidrap	Avian Influenza	2005.05	2005.05	Indonesia	172	†	172 deaths
2007.04.19	ft	Avian Influenza	2005	2007	Indonesia	66	†	at least 66 human deaths
2007.04.19	ft	Avian Influenza	2003.09	2003.12	Indonesia	170	†	more than 170 people
2007.04.19	theglobeandmail	Avian Influenza	2003.09	2003.12	Indonesia	300		nearly 300 people
2007.04.19	ChinaPost	Avian Influenza	2003	2007	Indonesia	--		--
2007.04.17	cidrap	Avian Influenza	2007	2007	Cambodia	302	†	1,086 susceptible birds
2007.04.16	recomb	Avian Influenza	2007.04.14	2007.04.14	Indonesia	--	†	the family's chickens
2007.04.16	promed	Avian Influenza	2007.04.05	2007.04.05	Cambodia	--	†	the 13-year-old girl
2007.04.16	dailytimesPK	Avian Influenza	2007.04.12	2007.04.12	Cambodia	--	†	the Cambodian girl
2007.04.16	dailytimesPK	Avian Influenza	--	--	Cambodia	--	†	a 13-year-old girl
2007.04.15	medicinenet	Avian Influenza	2003	2003	Indonesia	33		33 people
2007.04.15	medicinenet	Avian Influenza	2003	2003	Indonesia	24	†	24
2007.04.14	JakartaPost	Avian Influenza	2007.04.13	2007.04.13	Indonesia	74	†	the country's 74 human bird flu fatalities
2007.04.11	cidrap	Avian Influenza	2007.04.11	2007.04.11	Cambodia	172	†	fatal H5N1 cases

<< 1 **2** 3 4 5 6 7 ... 11 12 13 >>

Fig. 3: *A view of extracted incidents.*

For detailed information about the design and operation principles behind the PULS system, see, e.g., [7, 2]. PULS operates by pattern matching, and relies on several kinds of domain-independent and domain-specific *knowledge bases*. An example of domain-independent knowledge is the location hierarchy, containing names of countries, states or provinces, cities, etc. An example of a domain-specific knowledge base is the medical ontology, containing names of diseases, viruses, drugs, etc., organized in a conceptual hierarchy. The system also contains a domain-specific pattern base—which contains patterns that map the surface-syntactic representation of the information in the sentence to the semantic representation in the database records. Populating the knowledge bases requires a significant investment of time and manual labor. PULS employs weakly-supervised methods to reduce the amount of manual labor as far as possible, by bootstrapping the knowledge bases from large, unannotated document collections, [6, 3].

3.2 Toward Cross-Document IE

PULS goes beyond the traditional IE paradigm in two respects. First, in a traditional IE system, documents are processed separately and independently; facts found in one document do not interact with information found in other documents. Second, for each attribute in an extracted incident, traditionally, the IE system stores only one value in the database record—the value that is the locally best guess for that attribute.

1. After PULS extracts information from each document locally, it attempts to globally unify the extracted facts into groups, which we call *outbreaks*. An outbreak is a set of related incidents. Currently, incidents are related by straightforward heuristics: they must share the same disease name and the same country, and be "reasonably close" in time. Closeness is determined by a time window, currently fixed at 15 days.[6] Any chain of incidents which are separated by no more than the time window are aggregated into the same group.

An outbreak therefore serves as a kind of a "summary" of the incidents it contains, and provides an extra level of abstraction between the user and the "low-level" facts/incidents.

2. When PULS stores a record in the database, for each attribute, in general, rather than storing a single value, PULS stores a distribution over a set of possible values. For example, the sample text (in the first paragraph of this section) might read instead *"Five more people died last week."* PULS will then try to fill in the missing attributes (i.e., the disease name, location) by searching for entities of the corresponding semantic type elsewhere in the discourse. In general, for a given attribute of an event, the document will contain several possible candidate entities, and each candidate will have a corresponding score—measuring how well it fits the event. The score depends on certain features of the candidate value. These features include whether the value is mentioned inside the *trigger*—the piece of

[6] This could be made more flexible, e.g., dependent on the disease type.

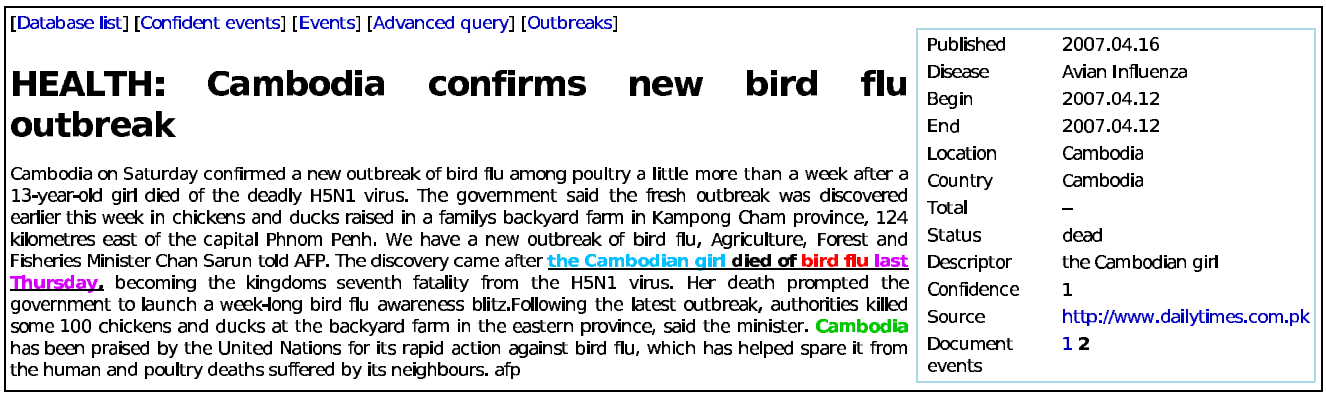

Fig. 4: *An epidemic event extracted from a document.*

text that triggered some pattern from the pattern base; whether it appears in the same sentence as the trigger; whether it appears before or after the sentence containing the trigger; whether this value is the unique value of its type, in the sentence that contains the trigger (e.g., the sentence mentions only a single country, or disease); whether the value is unique in the entire document; etc.

Using a set of candidate values rather than a single candidate is helpful in two ways. First, it allows us to compute the *confidence* of an incident, which is used in cross-document aggregation (in section 4.1). Second, it allows us to explore methods for recovery from locally-best but incorrect guesses by using global information.[7]

In the next section, we will discuss how these features of the PULS system are used in the combined, multi-source system.

4 Integration of MedISys and PULS

This section will describe the integration between MedISys and PULS, and will try to demonstrate that, even in its current, preliminary state, the integrated whole is greater than the sum of its parts.

A special RSS tunnel has been set up between MedISys and PULS. At present, PULS is able to process only English-language documents. MedISys forwards documents which it categorizes as relevant to the medical domain through the tunnel to PULS. Currently, the documents arrive as plain text, with no layout information (paragraphs, sections, etc). This is done in addition to the normal processing on the MedISys side, where running averages are monitored for all alerts, etc. A document batch is sent every 10 minutes, with documents newly discovered on the Web.

On the PULS side, the IE system analyzes all documents received from MedISys, and returns information that it extracted from the received documents back through the tunnel—in structured form (also at 10 minute intervals). This communication is asynchronous, and does not affect the functioning of both sites, which are inter-operating normally in real-time.

4.1 Multi-document Aggregation

When documents are received from MedISys, PULS performs the following processing steps:

First, the IE system analyzes the documents, extracts incidents, and stores them in the local database (*doremi.cs.helsinki.fi/jrc*). Second, PULS uses local heuristics to compute the *confidence* of the attributes in the extracted incidents.

The confidence of an attribute is computed from the set of candidate values for that attribute, based on their scores, which are in turn based on the features, as explained in Section 3.2. If the score of the best value exceeds a certain threshold, the attribute is considered *confident*.

Some of the attributes of an incident are considered to be more important than others: here, in the case of epidemic events, these *principal* attributes are the disease name, location and date. If all principal attributes of an incident are confident, the entire incident is considered confident as well.[8]

Third, the system aggregates the extracted incidents into outbreaks, across multiple documents and sources. The aggregation process requires that at least one of the incidents in each outbreak chain must be confident (that is, chains composed entirely of non-confident incidents are discarded).

Finally, PULS prepares a batch of recent incidents to return to MedISys, for displaying on its pages. The goal is to return a set of recent incidents with high confidence and low redundancy—a complete yet manageably-sized set of news for MedISys users to explore.

The batch is restricted to documents published within the last 10 days; from this period, PULS re-

[7] This line of our current research is not covered in this paper.

[8] In the PULS tables, confident attributes are set off in bold text, and confident incidents are highlighted in blue.

turns the most recent 50 incidents, filtering out duplicates: if multiple incidents of the same disease in the same location are reported, PULS returns only the most recent one.[9]

On the MedISys side, the returned events are displayed in two views. The main MedISys page displays the five most recently published events—these correspond to the most urgent news. For more detail, this box has a link to the entire batch of 50 most recent incidents. For the full view, the recent list has a link to the complete PULS database.

4.2 Performance

We now discuss some of the on-going evaluations of the currently deployed systems.

The number of documents PULS receives from MedISys is approximately 10,000 per month. From 2,700 of these, PULS extracts approximately 6,000 incidents per month, on average. (It is quite common for a relevant document to contain more than one incident.) The remaining 6,300 documents fed to PULS by MedISys produce no incidents. That is to be expected, since MedISys does not explicitly search for outbreaks, but for any *mentions* of disease names, and many documents may mention the crucial diseases in the context of new vaccines or treatments, eradication campaigns, etc.

To determine what proportion of these 6,300 documents actually do contain events—and are therefore false negatives from the perspective of PULS—we randomly selected and checked 100 MedISys documents that produced no events. Of these, 15% contained an event that the IE system missed.

From the perspective of MedISys, this roughly indicates that at least 64% of the documents fed to PULS on average contain no events. This confirms that the IE component indeed serves its purpose by helping to distinguish reports about epidemic outbreaks from other discussions concerning diseases.

About 20% of all extracted incidents are rated as confident. We tried to estimate the accuracy of the confidence heuristics. We selected 100 confident incidents at random, and checked their correctness by hand. Without employing a rigorous (e.g., MUC-style) evaluation, we consider an incident to be correct only if all of its principal attributes are correct (no partial credit). This evaluation yielded: 72% of the confident incidents are correct; in 14% of the cases, the information extraction is erroneous, i.e., PULS extracts an incident where there should be none; in 14% of the cases, the confident incident is incorrect—for at least one attribute, the top-ranked value is not correct. The latter category of error is difficult to correct, since it is usually due to an inherent complexity in the text. The former type of error is simpler to correct, as it usually entails some tuning of the knowledge bases. Thus, if we could correct the erroneous cases with some tuning labor, we might expect the confidence measure to be correct in just under 84% of the confident incidents.

Since outbreak aggregation is our primary means of reducing redundant information in the flow of news, it is important to have an estimate of the accuracy of the outbreak calculation. We analyzed a randomly chosen set of medium-sized outbreaks, 20 outbreaks, about 10 incidents each. For each incident we tried to determine whether it was appropriately included in the outbreak. We found that 68% of the incidents were correctly identified with their outbreaks. Three of the outbreaks (about 15%) were erroneous, i.e., based on incorrect confident incidents.[10]

22.5% of the examined incidents were confident (i.e., on average, the outbreaks contained only 2–3 confident incidents).

5 Conclusion and Future Work

The public and the restricted MedISys applications are currently independent of each other, and they provide different functionality. The medical event information is only available on the public site. The two systems will soon be integrated in order to allow a single entry point and visual presentation. Registered users will receive access to more functionality and more alert definitions. Depending on the users' access rights, they may get access to newswires and to commercial sources as well.

We further plan to integrate a tool that automatically extracts terms from the comprehensive medical thesaurus MeSH (Medical Subject Headings),[11] and to allow users to select articles by browsing and drilling down in the multilingual MeSH hierarchy. This will give the user an alternative entry point to the same information.

We need to resolve some technical problems to improve the quality of the input data. One problem relates to the way MedISys extracts textual content from

[9] Note that under this arrangement, a recent event that was last reported more than 10 days ago, will not appear in the result list, while an event from several months ago may appear—if it is mentioned in a very recently published report. This is a design decision that aims to balance the tension between recency of *publication* vs. recency of actual *occurrence* of an incident: both may be important to the user. Note also that in any case *all* events are always available in the PULS database for browsing.

[10] It was interesting to observe that aggregation is often useful even when the outbreak consists entirely of incorrectly analyzed incidents. E.g., in high-profile cases picked up by main news agencies, reports are re-circulated through multiple sites worldwide. Because the text is very similar to the original report, the IE system extracts similar incidents from all reports, and correctly groups them together. Although some attribute is always analyzed incorrectly, the error is *consistent*, and the grouping is still useful: it helps reduce the load on the user by aggregating related facts.

[11] www.nlm.nih.gov/mesh

source sites. Because the original focus of MedISys was on the keywords contained in the text, it ignored document layout information (such as headings, sub-headings, by- and date-lines, paragraph breaks, etc.), which provides important cues when detailed text analysis is required. The lack of this information is known to confuse the IE process, and needs to be addressed to improve IE accuracy.[12]

We are currently investigating methods for extending the measure of local confidence to global confidence, across multiple documents and sources.

We also plan to develop methods for MedISys to exploit the information returned from PULS in novel ways. One problem that needs to be studied is to what extent the outbreaks extracted by MedISys based on keyword frequencies agree with outbreaks extracted by PULS, and how they can be best integrated. Another path under consideration is to incorporate the PULS confidence as a criterion for the urgency of MedISys alerts. The current scheme, based on cumulative statistics, assumes that if something is newly prominent in many news sources, it is urgent or interesting news. However, in some cases, news that appears everywhere is "dated" news—it is already highly publicized. For timely surveillance, it is also interesting to detect outlier reports—those that have not yet achieved wide publicity, but in this case it is crucial that the system be certain that it was correctly identified. Here, a high score on the PULS confidence scale may serve as a complementary criterion for the urgency of an event.

Acknowledgements

This work was in part supported by the Academy of Finland. We wish to thank the anonymous reviewers for their thorough and helpful comments.

References

[1] C. Best, E. van der Goot, K. Blackler, T. Garcia, and D. Horby. Europe media monitor—system description. Technical Report 22173 EN, EUR, 2005.

[2] R. Grishman, S. Huttunen, and R. Yangarber. Information extraction for enhanced access to disease outbreak reports. *J. of Biomed. Informatics*, **35**(4), 2003.

[3] W. Lin, R. Yangarber, and R. Grishman. Bootstrapped learning of semantic classes from positive and negative examples. In *Proc. ICML Workshop*, Washington, DC, 2003.

[4] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A. Forslund, and C. Best. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of LREC-2006*, Genova, Italy, 2006.

[5] R. Steinberger, B. Pouliquen, and C. Ignat. Navigating multilingual news collections using automatically extracted information. *Journal CIT*, 13(4), 2005.

[6] R. Yangarber. Counter-training in discovery of semantic patterns. In *Proc. ACL-2003*, Sapporo, Japan, 2003.

[7] R. Yangarber, L. Jokipii, A. Rauramo, and S. Huttunen. Extracting information about outbreaks of infectious epidemics. In *Proc. HLT-EMNLP 2005*, Vancouver, Canada, 2005.

[12] Extracting document layout accurately is a highly non-trivial problem, since source sites are completely unstandardized, and in general the layout is hard to infer automatically.