UNIVERSITÄT
DES
SAARLANDES

Using Terminology and Sub-ontological Clusters to Improve Statistical Machine Translation

Liling Tan

Supervisor: Josef van Genabith
Co-supervisor: Francis Bond

Submitted in part fulfillment of the requirements for the degree of
Doctor of Philosophy
in Sprachwissenschaft und Sprachtechnologie

June 2018

Declaration of Authorship

I, Liling Tan, hereby declare that this dissertation is my own original work except where otherwise indicated. All data or concepts drawn directly or indirectly from other sources have been correctly acknowledged. This dissertation has not been submitted in its present or similar form to any other academic institution either in Germany or abroad for the award of any other degree.

Liling Tan
Saarbrücken, 01 Jan 2018

Acknowledgements

This thesis is made possible by the mentoring and wise advice of **Prof. Josef van Genabith** and **Dr. Francis Bond**. I thank them for their patience and the impartation of their knowledge of the machine translation and natural language processing. **Prof. Josef van Genabith** especially deserves thanks for creating various opportunities to further my knowledge on machine translation and deep/machine learning that forms the foundations of this thesis. I am also thankful for **Prof. Elke Teich** constant encouragements to my academic pursuit and her support in various side computational linguistics side projects during my stay in Saarland.

The PhD journey couldn't have been interesting without the collaboration and beer momements with my fellow Saarland colleagues, *Jon Dehdari, Jose Manuel Martinez Martinez, Noam Ordan, Santanu Pal, Anne Schumann, Mihaela Vela* and *Marcos Zampieri*. I am thankful also for the fun times with my fellow EXPERT partners for going through the journey together, esp. *Rohit Gupta* and *Carolina Scarton*, who had joined me in several shared task participations.

The EXPERT consortium have provided opportunities to collaborate across borders and the exchanges I have spent at WordFast, Universida de Malaga and University of Wolverhampton. My interactions with **Prof. Gloria Corpas Pastor**, *Hernani Costa* and *Anna Zaretskaya* at the University of Malaga gave me a better understanding of human-centric language technologies. It was an honor to work with **Mr. Yves Champollion** who had hosted me at WordFast. The experience had gave me valuable insights to the translation industry. Thank you **Prof. Ruslan, Dr. Constantin Orasan** and **Iain Mansell** for organizing the consortium and hosting me in UoW for the exchange; nothing sums up my stay at the Wolves better than the blogpost.[1]

The research leading to this thesis has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

[1] http://rgcl.wlv.ac.uk/2016/01/26/farewell-from-liling/

Abstract

Statistical Machine Translation, yada yada...

Überblick

Contents

1 Introduction **1**

1.1 Motivation and Objectives 1

1.2 Contributions 2

1.3 Publications 3

2 Background **5**

2.1 Statistical Machine Translation 6

2.1.1 Phrase-Based SMT 7

2.1.2 Other SMT Models 7

2.2 The Mathematics of Statistical Machine Translation 11

2.3 Using Lexical Information in Statistical Machine Translation 14

2.3.1 Dictionaries are Data too 15

2.3.2 Augmenting Manual Dictionaries to Improve Statistical Machine Translation 16

2.3.3 Grouping Multi-Word Expressions in Statistical Machine Translation 17

2.3.4 Domain Adaptation for Lexically Informed Statistical Machine Translation 18

2.3.5 Training Phrase-Based SMT Models using Domain-Specific and Generic Corpora and Dictionaries 20

2.3.6 Improving Statistical Machine Translation Using Domain-Specific Bilingual MWEs . 22

2.3.7 A Overview of Integrating Lexical Information in SMT 25

2.4 Terminology and Ontology 27

2.4.1 A Survey of Term Extraction Techniques 27

2.4.2 The Importance of Terminology in Machine Translation 34

2.4.3 A Survey of Ontology Induction Techniques 35

2.5 Ontology and Translation 38

2.6 Summary 39

3 Resources and Experimental Setup 40

3.1 Resources used in Machine Translation Experiments 40

3.1.1 Asian Scientific Paper Excerpt Corpus (ASPEC) 40

3.1.2 Workshop for Machine Translation (WMT) News Domain Corpus 40

3.1.3 JICST Dictionary 41

3.1.4 The Phrase-based SMT Configuration used throughout the Thesis 42

3.2 Resources used in Terminology and Ontology Extraction 43

3.2.1 Disease Names and Adverse Effect (DNAE) Corpus 43

3.2.2 WikiFood Corpus 43

3.2.3 SemEval TaxEval Ontologies 44

4 Information Theoretic and Language Model based Term Extraction using Pointwise Mutual Information 45

4.1 Pointwise Mutual Information 46

4.2 Term Extraction Using Language Model Pointwise Mutual Information 47

4.3 Evaluating PMI_{LM} against a Gold Standard 48

4.4 Evaluating PMI_{LM} in Information Retrieval 50

4.5 Bilingual Term Extraction using PMI_{LM} 51

4.6 Evaluating PMI_{LM} as Additional Lexical Knowledge for SMT 52

4.7 Summary 54

5 Using Terminology to Improve Machine Translation 55

5.1 Passive and Pervasive Lexical Injection 56

5.1.1 Adding Single Word vs Multi-Word Dictionary Entries to SMT 57

5.1.2 Adding a Manually Crafted Dictionary vs an Automatically Extracted Terminology to SMT 57

5.1.3 Experimental Setup 58

5.1.4 Results (Passive vs Pervasive) 59

5.1.5 Results (Mono-lexeme vs Multi-Word Expressions) 61

5.1.6 Results (Manual vs Automatic) 63

5.2 Summary 66

6 Measuring the Goodness of Machine Translation 68

6.1 Evaluating Semantic ‘Goodness’ of Machine Translated Texts 69

6.1.1 Measuing Grammatical Fluency with MT Evaluation Metrics Ensemble 70

6.1.2 Measuing Semantic Adequacy With Monolingual Word Alignments and Neural Network Embeddings 73

6.1.3 Experimental Setup and Results 77

6.1.4 Summary 80

6.2 The Awkward Disparity between BLEU / RIBES Scores and Human Judgments 81

6.2.1 BLEU 82

6.2.2 RIBES 84

6.2.3 Other Shades of BLEU / RIBES 86

6.2.4 Experimental Setup 87

6.2.5 Results 88

6.2.6 Segment Level Meta-Evaluation 89

6.2.7 Summary 90

6.3 Summary 91

7 Ontology Induction using Neural Vector Space 92

7.1 Inducing a Hypernym with *is-a* Vector 93

7.1.1 Experimental Setup 94

7.1.2 Evaluation Metrics 95

7.2 Results 95

7.3 Hyponym Endocentricity 96

7.3.1 Identifying Endocentric Parts 97

7.4 Extending Taxonomy with Wikipedia List of Lists of Lists 98

7.4.1 Limitations of LOLOL Trawler 98

7.4.2 Results 99

7.5 Summary 101

8 Using Sub-Ontological Knowledge to Improve Machine Translation 102

8.1 Exchange-based Word Clustering 103

8.2 Experimental Setup 104

8.3 Results 105

8.4 Summary 106

9 Conclusion 107

9.1 Future Work 108

9.1.1 Terminology Extraction 108

9.1.2 Ontology Induction 109

9.1.3 Using Terminology and Ontology in Machine Translation 109

Bibliography 109

List of Tables

2.1 A Comparison of Previous Work in Integrating Lexical Resources in Statistical Machine Translation 24

3.1 Disease Names and Adverse Effect (DNAE) Corpus Statistics 43

4.1 Comparison of PMI_{LM} against State-of-Art Term Extractors 48

4.2 Accuracy and Mean Reciprocal Rank for Term Extraction for WikiFood 51

4.3 An Example Output of Ranked Term Candidates and Metrics Scores 51

4.4 BLEU Scores from Phrase-based SMT Systems with Passively Added Terminology of Incremental Sizes; (*) refers to statistically significant results at p $<$ 0.05 53

5.1 BLEU Scores for Passive and Pervasive Use of the Dictionary in SMT (Japanese to English) . 59

5.2 BLEU Scores for Passive and Pervasive Use of the Dictionary in SMT (English to Japanese) . 60

5.3 BLEU Scores in adding Mono-lexeme vs MWE Dictionary to SMT (English to Japanese) . . . 61

5.4 BLEU Scores in adding Single Words vs Multi-Words Dictionary to SMT (Japanese to English) 62

5.5 BLEU Scores in passively adding a Manually Crafted Dictionary vs an Automatically Extracted Terminology to SMT (English to Japanese) 63

5.6 BLEU Scores in adding Manually Crafted Dictionary vs Automatically Extracted Terminology to SMT (Japanese to English) 65

6.1 Stasis Metric Evaluated on the STS-2016 Task; *Best* is the best performing results per domain from various systems participating in the English STS-2016 shared task and *Median* is the median scores from all participating systems in the English STS-2016 shared task. 77

6.2 Differences between Organizer's and our Phrase-based SMT system 87

6.3 BLEU and HUMAN scores for WAT 2015 88

7.1 Structural Measures and Comparison against Gold Standards for USAAR-WLV. The labels of the columns refer to no. of distinct vertices and edges in induced taxonomy (|**V**| and |**E**|), no. of connected components (**#c.c**), whether the taxonomy is a Directed Acyclic Graph (**cycles**), vertex and edge coverage, i.e. proportion of gold standard vertices and edges covered by system (**%VC** and **%EC**), no. of vertices and edges in common with gold standard (**#VC** and **#EC**) and ratio of novel edges (**:NE**). 95

7.2 Averaged F&M Measure, Precision, Recall, F-score for All Systems Outputs when Compared to Gold Standard and Manually Evaluated Average Precision of Novel Edges. 96

7.3 Results of Our Endocentric Hypo-Hypernym Identifier Against the Gold Standard Taxonomy (**#Terms** refers to the no. of terms in the domain and **#Relations** refers to the no. of hypo-hypernym pairs found in the gold-standard taxonomy. **#Correct / #Identified** refers to the proportion of hypo-hypernym pairs our system has correctly identified. **Bold** items indicates that it is highest score among the competing teams in TExEval-2. The asterisk * indicates that the trawler was used to produce submissions for this domain.) 99

7.4 Results of a Manual Evaluation on 100 Random Novel Hypo-Hypernym Pairs for Competing Teams In TExEval-2 99

7.5 Comparison of String-based Methods 100

8.1 BLEU scores mkcls→BIRA and significance across cluster sizes ($|C|$) 105

8.2 BLEU score changes and significance across cluster sizes. 105

List of Figures

4.1 Precision, Recall and F-score of Unsupervised Term Extraction of PMI_{LM} and ProMiner on the DNAE Corpus 49

4.2 Precision, Recall and F-score of Supervised Term Extraction of PMI_{LM} and ProMiner on the DNAE Corpus 49

4.3 Adding Automatically Extracted Terminology of Different Sizes to SMT 53

5.1 Overview of the Passive and Pervasive Use of Dictionary in the WAT Experiments 60

5.2 Overview of the Mono-Lexeme and Mult-Word Expression subset of the Dictionary in the WAT Experiments 62

5.3 Passively Adding a Manually Crafted Dictionary vs an Automatically Extracted Terminology (English-Japanese) 64

5.4 Passively Adding Manually Crafted Dictionary vs Automatically Extracted Terminology (Japanese-English) 66

6.1 L1 Error Analysis on the answer-answer domain 79

6.2 L1 Error Analysis on the post-editing domain 79

6.3 Correlation between BLEU, RIBES differences and *Positive* HUMAN Judgements (HUMAN Scores of 0, +1, +2, +3, +4 and +5 represented by the colored bubbles: *grey, orange, blue, green, red and purple*; larger area means more segments with the respective HUMAN Scores) 89

6.4 Correlation between BLEU, RIBES differences and *Negative*HUMAN Judgements (HUMAN Scores of 0, -1, -2, -3, -4 and -5 represented by the colored bubbles: *grey, orange, blue, green, red and purple*; larger area means more segments with the respective HUMAN Scores) 89

8.1 End-to-end translation model training times for English-Russian and Russian-English for various cluster sizes using `mkcls` and `BIRA`. 106

Chapter 1

Introduction

1.1 Motivation and Objectives

The main motivation behind this thesis is to explore the issues of (i) creating novel state-of-art terminology and ontology extraction systems and (ii) integrating semantic (terminological and ontological) knowledge into statistical machine translation within the same domain as the domain of the training data.

Unlike normal domain adaptation strategies that integrate domain specific semantic knowledge into models trained on generic data, **the objective of this thesis to incorporate domain specific terminological and ontological knowledge into machine translation models already trained from the domain specific data** of the same domain.

The thesis seeks to answer the following questions:

- How can terminological knowledge be extracted from monolingual and parallel corpora?
- How can ontological information be induced using state-of-art distributional approach such as word embeddings and using simple string-based patterns?
- Can terminological / lexical knowledge be used in Statistical Machine Translation to improve translation quality?
- Can sub-ontological knowledge (i.e. word clusters) be used to improve Statistical Machine Translation?

In brief, we report the following in this thesis to address these research questions:

- We introduce a **novel information theoretic approach for term extraction** based on Pointwise Mutual Information (PMI) and language models (Chapter 3)
- We investigate various setups to using additional lexical information to phrase-based statistical machine translation to produce **statistically significant but marginal BLEU gains**[1] (Chapter 4)
- To understand more about machine translation metrics, esp. BLEU, we introduce a **semantically motivated adequacy metric** to measure the 'goodness' of translation (Chapter 5.1)
- We **meta-evaluated the correlation between BLEU/RIBES MT evaluation metrics against human judgements** (Chapter 5.2)
- We introduce an **unsupervised novel hypernym induction** techniques using neural network word embeddings (Chapter 6)
- We integrate sub-ontological knowledge into SMT by **scaling word clustering** that is being used in SMT word alignments (Chapter 7)

1.2 Contributions

The work presented in this thesis contributes to the field of Natural Language Processing (NLP) and Statistical Machine Translation (SMT) in the following ways.

Innovations:

- We create a novel approach to extract terminology monolingually using a pre-trained language model (PMI_{LM}) and extend it for bilingual terminology extraction with the use of word/phrase alignments
- Different from the normal use of a dictionary for the purpose of domain adaptation where normally, a domain-specific lexicon is appended to a translation model trained on generic texts, we are investigating the use of an in-domain dictionary in statistical machine translation.
- We propose a semantically adequacy motivated evaluation metric that combines state-of-art word embeddings and an MT evaluation metrics ensemble
- We introduce a new method to induce hypernyms in an unsupervised manner using the 'is-a' non-content word embedding

[1] Our best systems with added lexical information achieved 24.14 BLEU for Japanese to English translation (baseline: 23.91 BLEU) and 17.38 BLEU for English to Japanese translation (baseline: 16.75 BLEU).

Empirical Findings:

- Using additional lexical information (automatically extracted terminology or a manually crafted dictionary) in SMT can provide marginal BLEU score increment (often statistically insignificant) but more often degrades system performance
- Passively adding lexical information in SMT more than once can improve SMT but it also provides marginal BLEU improvements (sometimes statistically significant) and the number of times to add the lexicon becomes an additional hyperparameter that is not in general justified by the marginal gains
- BLEU / RIBES score correlation only occurs when the translation is inherently good; when it is not, BLEU / RIBES tends to be overly optimistic due to their reliance on crude surface string (ngram) based nature.
- Using sub-ontological information in machine translation can provide substantial speed gains in training a phrase-based SMT system

1.3 Publications

Various parts of this thesis were published in peer-reviewed natural language processing and computational linguistics conference and workshop proceedings.

Chapter 4 describe the Pointwise Mutual Information (PMI) terminology extraction related experiments are published in Tan (2016a) and Tan (2016b). The papers describes the preliminary findings of the terminology extractor where the language model based PMI extractor proposed in this thesis outperformed the classic C-value based term extraction on a subset of food-related articles on Wikipedia. The extracted terms was evaluated against a food domain terminology provided in SemEval-2015 taxonomy evaluation task. A list of candidate terms were extracted, each Wikipedia sentence and if the sentence contains the words in the SemEval-2015 taxonomy task, we compute the mean reciprocal rank (MRR) and accuracy of the candidate terms extracted from these sentences.

Chapter 6 summarizes the pursuit of finding a metric of ‘goodness’ of translation quality and this work was previously published in (Tan et al., 2015c; Vela & Tan, 2015a; Bechara et al., 2016a) in collaboration with various research partners within the EXPERT projects. We have pit ours metric in SemEval and WMT competitions.

Our best system was ranked in top 10 out of 130 submissions from 40 over teams competing in the SemEval semantic similarity task. We've combined state-of-art deep learning word embeddings with a deluge of machine translation evaluation metrics to create a robust ensemble metric that attempts to measures the similarity of two sentences. We showed that by using the distributed semantic representation captured by the embeddings and the syntactic grammaticality encapsulated in the surface word similarity from the MT metrics, we effectively assess the adequacy and fluency of translations that yields high similarities between the MT hypotheses and their reference translations

The second part of Chapter 6 takes a deeper dive into understanding what exactly are MT evaluation metric evaluating. It presents the joint publication where we meta-evaluated the disparity between the popular BLEU metric and human judgments. This work was previously published in the Workshop for Asian Translation (Nakazawa et al., 2015) as the error analysis of our submission to the shared task (Tan, 2016c).

Part of Chapter 7 comes from the two year participation in the SemEval taxonomy evaluation shared tasks where we created a function phrase vector that identifies the hypernym of a term within an embedding vector space (Tan et al., 2015b; Tan, 2016e).

Finally, a portion of Chapter was previously published in collaboration (Dehdari et al., 2016b,c). It describes the experiments of using word clusters (i.e. sub-ontological knowledge) to improve machine translations. The publications were more focused on the novel clustering mechanism while the thesis chapter zoomed in on the word clusters effects in improving machine translation quality.

A full list of publications including the above described and covering other related but not directly related topics can be found in the following 2 pages.

Chapter 2

Background

This chapter provides an literature review of the related work from the field of Machine Translation (Section 2.1 and 2.2), the use of lexical information in Statistical Machine Translation (Section 2.3), a survey of the state-of-art terminology extraction (Section 2.4.1 and 2.4.2) and ontology induction techniques (Section 2.4.3).

The *most relevant background* to the work described in the following chapters are:

- **Section 2.2** on Phrase-based Machine Translation[1]
- **Section 2.3.7** on an overview of integrating additional lexical information in SMT
- **Section 2.4.1** on a brief survey of term extraction techniques and
- **Section 2.4.2** that motivates the importance of terminology in machine translation
- **Section 2.4.3** on a brief survey of ontology induction techniques
- **Section 2.4.4** describing the phrased-based SMT configuration used throughout the thesis[2]

[1] This is the main machine translation paradigm we used in this thesis.

[2] Unless explicitly stated in the prose, all SMT experiments described in this thesis used this phrase-based configuration.

2.1 Statistical Machine Translation

Machine Translation (MT) is the computational task to automatically translate between human languages.

The history of automatic translation traces back to 17th century ideas of a universal language at the demise of Latin as the global *lingua franca* (Hutchins, 2000).

Before the invention of modern day computing machines, attempts were made to create *mechanical dictionaries* that tried to map words/ideas/concepts into numerical code to mediate between languages. The most influential approach is perhaps Leibniz's (1714) monadic theory that encapsulates symbolic thoughts and formal logic (Busche, 2009) and assumes that cross-lingual understanding between languages is a matter of mapping natural language into monads. In one of his memos, *The Art of Discovery* (1685), Leibniz wrote:

"*This [monadic] language will be the greatest instrument of reason [for] when there are disputes among persons, we can simply say: Let us calculate, without further ado, and see who is right.*"

In the early days of modern day machine translation in the 1960s, Becher's (1962) work titled *'Zur mechanischen Sprachübersetzung: ein Programmierungversuch aus dem Jahre 1661'* expresses that the historical "mechanical dictionary" approaches foreshadowed certain principles of machine translation. The aim of *mechanical dictionaries* are not unlike the modern day natural language processing (NLP) task of creating a multilingual ontology and word/concept mapping like Open Multilingual WordNet (OMW) (Bond & Paik, 2012) and multilingual Ontonotes (Weischedel et al., 2010).

In the 1930s, patents for a *"general purpose translation using a mechanical multilingual dictionary"* and *"mechanical translations via universal grammatical functions"* were granted by France and Russia to Georges Artsrouni and Petr Trojanski respectively. The Statistical Machine Translation (SMT) paradigm is largely attributed to Warren Weavers memorandum to the Rockfella foundation in 1949 (Weaver, 1955):

"*It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code". If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?*"

Data-driven SMT has developed rapidly with the introduction of IBM word alignment models (Brown et al., 1990, 1993c) and a *word-based* SMT model which translates word-for-word. The word-based model was later superseded by phrase-based models (Och & Ney, 2002; Marcu & Wong, 2002; Zens et al., 2002; Koehn et al., 2003; Koehn, 2004a) which rely on the word-alignment methods developed by its word-based predecessor

(Al-Onaizan et al., 1999).

2.1.1 Phrase-Based SMT

Phrase-Based Statistical Machine Translation (PB-SMT) models translate contiguous sequences of words from the source sentence to contiguous words in the target language. In this case, the term *phrase* does not refer to the linguistic notion of syntactic constituent but the notion of *n*-grams . Knight (1999) showed that decoding word/phrase-based models involve search problems that grow exponentially with sentence length.

Phrase-based models significantly improve on word-based models, and work especially well for closely-related languages. This is mainly due to the modelling of local reordering and the assumption that most orderings of contiguous *n*-grams are monotonic. However, this is not the case for translation between language pairs with divergent syntactic constructions; e.g. when translating between SVO-SOV languages.

Tillmann (2004) and Al-Onaizan & Papineni (2006) proposed several sophisticated lexicalized reordering and distortion models to address most long-distance reordering issues. Alternatively, to overcome reordering issues with a simple distortion penalty, Zollmann et al. (2008) memorized a larger phrase *n*-gram sequence from very large training data and allow larger distortion limits; it achieves similar results to more sophisticated reordering techniques with fewer training data. In practice, reordering is set to a small window and Birch et al. (2010) showed that phrase-based models tend to perform poorly even for short and medium range reordering.

2.1.2 Other SMT Models

Hierarchical Phrase-Based SMT

Hierarchical phrase-based machine translation (aka *hiero* or HPB-SMT) extends the phrase-based models concept of phrase from naive contiguous words to a sequence of words and sub-phrases (Chiang, 2005). Within the hiero model, translation rules are created using make use of the phrases extracted from the PB-SMT and the reordering of the subphrases. For example, *'She likes him"* is translated to German as *"Er gefällt ihr"*[3]; this reordering can be expressed as a lexicalized *gappy* synchronous hierarchical rule using X_1 and X_2 as placeholders for subphrases.

$$< X_1 \; likes \; X_2, \; X_2 \; gefaellt \; X_1 > \tag{2.1}$$

[3] *Er* is the third-person singular masculine pronoun, while *ihr* is the third-person singular feminine pronoun

Hierarchical phrase-based models can also model discontiguous phrases such as the long distance dependence between the verb and its negation in German, e.g.

$$< X_1\ do\ not\ like\ X_2\ ,\ X_2\ gefaellt\ X_1\ nicht > \tag{2.2}$$

In the hiero model, these translation rules are the production rules of a Synchronous Context-Free Grammar (SCFG). Target language translations are generated from both the SCFG parses and the surface string inputs. The translation rules are induced from a parallel corpus without linguistic annotation from any grammatical formalisms. The induction is based on distributional properties of the PB-SMT style phrases. Arguably, SCFG provides minimal subphrasal lexical selection syntax that is agnostic to any linguistic commitments or assumptions.

Although the hiero model provides more robust translation possibilities, the size of the grammar is exponential because of the arbitrary re-orderings between the source and target language. Zhang et al. (2006) introduced a linear-time algorithm for factoring syntactic re-orderings by restricting synchronous rules on the source-side grammar to binary branching nodes when possible. The binarization of the SCFG rules significantly improves the speed and accuracy of the hiero models.

Most open-sourced implementations of machine translation decoders support both phrased-based and binarized SCFG hiero models, they differ primarily by the programming language used in the implementation and the varying computing structures used to parse the trees and the formalisms used to present the same phrased-based and hiero model (Koehn et al., 2007; Hoang & Koehn, 2008; Li et al., 2009; Weese et al., 2011; Dyer et al., 2010).

Still, neither the basic phrase-based model nor the basic hierarchical phrase-based model incorporate any linguistic structure such as syntax, morphology, or semantics beyond surface strings.

Factored SMT

In the early days of SMT, the importance of linguistic information to translation was recognized (Brown et al., 1993c):

"*But it is not our intention to ignore linguistics, neither to replace it. Rather, we hope to enfold it in the embrace of a secure probabilistic framework so that the two together may draw strength from one another and guide us to* *better natural language processing systems in general and to better machine translation systems in particular.*"

Factored SMT embarked on the task of effectively incorporating linguistic information from taggers, parses and morphological analyzers into the machine translation pipeline. It is motivated by fact that (i) linguistic information provides a layer of disambiguation to address the ambiguity of natural language, (ii) generalized translation of out-of-vocabulary (OOV) words can overcome sparsity of training data and (iii) arbitrary limits are replaced with linguistic constraints put in place in the decoding process too keep the search space tractable (Hoang & Lopez, 2009; Koehn et al., 2010; Hoang, 2011).

The factored model extends the phrase-based model by reformulating each word as a vector of factors and each input factor produces an equivalent output factor in the target language. For example, a vector of a source word can be made up of its surface form lemma, POS tags and morphological annotations decoded into a vector in the target language. Then the surface form of the target language word is generated given the decoded vector. The decoding process to find the most appropriate translation follows the log-linear model as in the phrase-based model where the translation step and the generation step are regarded as additional components to the log-linear model (in addition to the existing language model and reordering model in phrase-based MT) (Koehn & Hoang, 2007a).

The integration of morphological information remains an impetus for deploying factored based models when (i) translating from morphologically rich languages (that cause model sparsity) due to under-sampled morphological variants of the same word (Bojar, 2007) or (ii) translating from morphological poor languages (that causes spurious ambiguities) to morphological rich languages that requires word declensions and conjugations that are lacking in the source language (Ramanathan et al., 2009).

Syntax-Based SMT

Early work on syntactically informed[4] SMT developed in tandem with the various components of phrase-based SMT (i.e. reordering, distortion, hiero, etc.). Based on the finite state automata, Wu (1997) and Wu & Hkust (1998) introduced the notion of Inversion Transduction Grammars (ITG) and Stochastic Bracketing Transduction Grammars (SBTG) where every terminal symbol (word/phrase) is marked for two output streams, (i) the non-terminal node to parse upwards towards the top of the parse tree and (ii) the equivalent terminal node on other language; the non-terminal nodes are represented as classes of derivable substring pairs.

[4] "Syntactically informed" refers to the linguistic theory of syntax.

Yamada & Knight (2001) presented a syntax-based SMT model that transforms a source language parse tree to its target language counterpart by applying stochastic operations at each node. By doing so, they can exploit the rich parsing resources for English. In their approach, they flattened trees to allow more reordering possibilities. The decoding process for their syntax-based model emulates a bottom-up parsing problem where nodes are translated individually and hypothesis pruning and hypothesis combinations are applied when the parser goes towards the top of the tree.

The stochastic operations proposed in Yamada & Knight were formalized as a theory to automatically derive from word-aligned corpora a minimal set of syntactically motivated transformation rules (Galley et al., 2004). These transformation rules (aka *GHKM rules*) map the input string (words/phrases) to the output tree fragments and Galley et al. (2006) showed that learning the probabilities of the rules with the EM algorithm produce “contextually-richer tree" (i.e. SCFG rules with more nodes). Likewise, syntactic parses and the surface string input can be feed into the decoder to generate phrases (Huang et al., 2006) and the annotations used to generate the transformation rules do not need to be restricted to syntactically structured trees, Liu & Gildea (2008) extended the model with semantic role labels. These models are usually dubbed the *String2Tree* and *Tree2String* SMT models.

Rather than using a single best syntactic parse, Mi et al. (2008) showed that using a forest of n-best syntactic parses of the source sentence improves upon the 1-best tree-based translation models. Neubig (2013) maintains a working implementation of the Forest2String MT systems using tree transducers[5].

The idea of generating GHKM rules from annotations is not restricted to a string input or output; *Tree2Tree* models can be learnt from word-aligned corpora. Zhang et al. (2008a) mapped source language tree fragments to target language fragments using Synchronous Tree Substitution Grammar (S-TSG). The Tree2Tree model is able to capture non-syntactic constituent phrases and discontinous phrases using linguistically structured features, additionally, it supports stratified structure reordering of larger trees. Similarly, Shieber (2007) proposed a Tree2Tree model using probabilistic Synchronous Tree Adjoining Grammar[6] (S-TAG).

Dependency Models and Deep Tecto MT

Lin (2004) introduced path-based models that extract a set of transfer rules from the corpus and each rule transforms a path in the source language dependency tree into a target language dependency tree fragment;

[5] http://www.phontron.com/travatar/

[6] It is worth noting that the non-probabilistic S-TAG grammar formalism precedes the ITG, SBTG and SCFG (Shieber & Schabes, 1990)

effectively searching for the best translation problem becomes a graphical problem to find the minimum path that covers the source language dependency tree. Menezes & Quirk (2005) proposed a shallower dependency treelet approach uses the source side dependency as a tree-based ordering model and suggested that the model can be improved by adding information such as semantic roles or morphological features.

Like the hiero model, Xiong et al. (2007) extracted transfer rules using the source side dependency treelet and gappy target language string fragments. Shen et al. (2008) presented a String2Tree model that extracted gappy string fragments in the source language mapping to the target language dependency treelet; this exploits a target dependency language model that was previously unavailable for the dependency Tree2String model.

Čmejrek et al. (2003) proposed a tectogrammatical model (TectoMT) that is also based on dependency trees, in addition, TectoMT includes morphological analysis and generations (Eisner, 2003; Žabokrtskỳ et al., 2008; Popel & Žabokrtskỳ, 2010). Similar to how phrase-based models map smaller fragments to larger ones, Bojar & Hajič (2008) extended the TectoMT model by mapping arbitrary connected fragments of the dependency tree.

2.2 The Mathematics of Statistical Machine Translation

The sections above described a zoo of SMT models with varying levels of linguistic information incorporated into the different models. In this section, we describe in detail the phrase-based SMT models that are used in this thesis.

Let's consider the scenario of translating French text into English, to formalize the convention, we will use ***s*** to denote a source language (i.e. French) sentence and ***t*** to denote a target language (i.e. English) sentence.

The objective of the MT system is to find the best translation $\hat{\boldsymbol{t}}$ that maximizes the translation probability p($\boldsymbol{t}|\boldsymbol{s}$) given a source sentence ***s***:

$$\hat{t} = \underset{t}{argmax}\ p(t|s) \tag{2.3}$$

Applying Bayes' rule, we can factorize p($\boldsymbol{t}|\boldsymbol{s}$) into three parts:

$$p(t|s) = \frac{p(t)}{p(s)} p(s|t) \tag{2.4}$$

Substituting our p($\boldsymbol{t}|\boldsymbol{s}$) back into our search for the best translation $\hat{\boldsymbol{t}}$ using *argmax*:

$$\begin{aligned} \hat{t} &= \underset{t}{argmax}\ p(t|s) \\ &= \underset{t}{argmax}\ \frac{p(t)}{p(s)} p(s|t) \\ &= \underset{t}{argmax}\ p(t)p(s|t) \end{aligned} \tag{2.5}$$

We note that the denominator p($\boldsymbol{s}$) can be dropped because for all translations the probability of the source sentence remains the same and the *argmax* objective optimizes the probability relative to the set of possible translations given a single source sentence.

At first glance, the formulation seems counter-intuitive in that in order to achieve the best translation in the target language $\hat{\boldsymbol{t}}$, we compute the component p($\boldsymbol{s}|\boldsymbol{t}$). This is derived from Bayes' rule and corresponds to casting the translation problem as an instance of the noisy channel model in information theory (Shannon, 2001).

We can explain this anecdotally, consider our machine translation system as a human translator who is a native speaker of the target language (let's say English). When he/she hears the source language sentence (i.e. French), he/she will conceive of an English sentence and we can consider the p($\boldsymbol{t}$) component as the grammaticality of that English translation.

The translator then tries to check that he/she has achieved the best translation of the source sentence $\boldsymbol{s}$ given the hypothesized English sentence, $\boldsymbol{t}$. And we can consider this process as the p($\boldsymbol{s}|\boldsymbol{t}$) component of our machine translation system.

Brown et al. (1993c) provides another anecdotal account to describe the noisy-channel model[7]:

> *As a representation of the process by which a human being translates a passage from French to English, this equation is fanciful at best. One can hardly imagine someone rifling mentally through the list of all English passages computing the product of the a priori probability of the passage, p(*t*), and the conditional probability of the French passage given the English passage, p(*s|t*). Instead, there is an overwhelming intuitive appeal to the idea that a translator proceeds by first understanding the French, and then expressing in English the meaning that he has thus grasped. Many people have been guided by this intuitive picture when building machine translation systems.*

[7] The notation in the quotation has been modified to suit the notation used in this thesis.

In another words, the source sentence probabilistically passes through the noisy channel p(***t***|***s***) to result in the target sentence:

$$\begin{array}{ccccc} p(s) & -> & p(t|s) & -> & f \\ source & & channel & & target \end{array} \tag{2.6}$$

We need to reason backwards to the best $\hat{\boldsymbol{s}}$ in the source that corresponds to $\hat{\boldsymbol{t}}$. We know the source probability p(s) and the channel probability p(***t***|***s***); i.e. how likely is ***s*** corrupted into ***t***

Log-linear Models

Extending the noisy channel model, Och & Ney (2002) simplified the integration of additional model components using the *log-linear model.* The model defines feature functions *h(x)* with weights λ in the following form:

$$P(x) = \frac{exp(\sum_{i=1}^{n} \lambda_i h_i(x))}{Z} \tag{2.7}$$

where the normalization constant *Z* turns the numerator into a probability distribution. In the case of a simple model that contains the two primary features from the noisy channel model, we define the components as such:

$$\begin{aligned} h_1(x) &= p(t) \\ h_2(x) &= p(s|t) \end{aligned} \tag{2.8}$$

and the *h(*x_1*)* and *h(*x_2*)* are associated with the λ_1 and λ_2 respectively.

The flexibility of the log-linear model allows for additional translation feature components to be added to the model easily, e.g. adding p($POS_s|POS_t$) to account for the translation of the part-of-speech (POS) transfers across the source and target language.

Additionally, the weights λ associated with the ***n*** components can be tuned to optimize the translation quality over the parallel sentences, ***D*** (often known as the development set):

$$\lambda_1^n = \underset{\lambda_1^n}{argmax} \sum_{d=1}^{D} \log P_{\lambda_1^n}(s_d|t_d) \tag{2.9}$$

2.3 Using Lexical Information in Statistical Machine Translation

At the transition from the word-based to phrase-based SMT paradigm, Tanaka & Baldwin (2003) highlighted that compound nouns are a major issue in machine translation because of their low frequencies and the high productivity of noun-noun compounds.

The noun-noun compound issue in machine translation has been particularly difficult for languages that do not separate them with whitespaces, e.g. Chinese (Wang et al., 2007; Chang et al., 2008, 2009; Pu et al., 2015), Japanese (Kitamura & Matsumoto, 1996; Cherny, 2000; Baldwin & Tanaka, 2004; Pinkham & Smets, 2008; Tsuji & Kageura, 2006), Vietnamese (Khanh & Ishizaki) and German (Rackow et al., 1992; Popović et al., 2006; Stymne, 2008; Stymne et al., 2013; Weller et al., 2014; Cap et al., 2014). It remains unsolved even with the recent advancement in character-level neural machine translation (Tran et al., 2014; Daiber & Sima'an, 2015; Sennrich et al., 2015a).

Additionally, researches have also investigated improving translations of other MWEs such as phrasal verbs (Simova & Kordoni, 2013; Kordoni & Simova, 2004; Cholakov & Kordoni, 2014) and named entities (Hermjakob et al., 2008; Nothman et al., 2013; Tan & Pal, 2014). Often these bilingual MWEs are extracted automatically using a combination of statistical and heuristics-based approaches and added as additional parallel training data prior to training the machine translation system (Tsvetkov & Wintner, 2012).

Another wealth of approaches on injecting lexical information in the SMT springs from the task of adapting machine translation systems to a specific-domain. While MWE motivated SMT research to focus on translating the MWEs correctly, domain adaptation for machine translation researches focused on clever ways to incorporate various resources (generic / in-domain parallel corpora, monolingual corpora and automatically extracted or manually crafted dictionary / terminologies) (Nanba et al., 2009; Sánchez Cartagena et al., 2011; Arcan et al., 2014; Miñarro-Giménez et al., 2015) or integrating them into Computer Assisted Translation (CAT) tools to aid human translation (Tezcan & Vandeghinste, 2011; Skadiņš et al., 2013)

The following sections of the thesis briefly survey seminal research that incorporated MWEs in SMT or used various lexical resources to adapt the machine translation system to a specific domain. The train of thought remains the same in finding novel and effective ways to incorporate additional lexical information to the standard SMT pipeline.

2.3.1 Dictionaries are Data too

"Machine translation depends vitally on data in form of large bilginual corpora, [but] bilingual dictionaries are also a source of information". Brown et al. (1993b) showed that a bilingual dictionary can be used as an additional parameter to the machine translation system and including a dictionary can improve the maximum likelihood estimates of the bilingual text alignments.

Given a dictionary entry with the source language word/phrase ***s*** and its corresponding ***m*** number of translations $\boldsymbol{t}_1$**, ...** $\boldsymbol{t}_m$ and if we consider that the lexicographer has chosen the translations from a random sample of ***c*** instances, the probability of translation for the dictionary entry is:

$$p(f_1, ..., f_m|e) = \sum_{c} \sum_{c_1>0} ... \sum_{c_m>0} \begin{pmatrix} c \\ c_1 ... c_m \end{pmatrix} p(c|s) \prod_{i=1}^{m} p(t_i|s)_{c_i} \tag{2.10}$$

Decrypting the equation from the right to left, the product of *p(*t*|s)* is the usual probability of the translation given the source word/phrase and the subscript in *p(*t*|s)*$_{c_i}$ refers to the subset of the global word/phrase translation probabilities given the sample of example sentences that the lexicographer has chosen.

The $[\sum_{c1>0} ... \sum_{c_m>0} \begin{pmatrix} c \\ c_{1...m} \end{pmatrix} p(c|s)]$ part of the equation can be simply thought of the “effective multiplier" that changes the word/phrase translation probabilities given the lexicographer’s choice of the sampled instances. For all possible ***c*** number of examples that a lexicographer can generate for a given source word, ***s***, the lexicographer chose ***m*** number of examples to explain the ***s*** and ***m*** translations for ***s***.

Simplifying the computation, Brown et al. (1993b) mathematically estimated the binomial distribution as a Poisson distribution, as such:

$$p(f_1, ..., f_m|e) = exp^{-\lambda(s)} \prod_{i=1}^{m} (exp^{-\lambda(s)p(t_i|s)} - 1) \tag{2.11}$$

where they introduced the $\lambda(s)$ variable to represents the Poisson distribution mean and that simplifies the translations of the word ***s*** as the product of each translation. To put it in Brown et al.’s (1993b) words:

> Imagine that a lexicographer, when constructing a [source language] entry for the English word or phrase ***s***, first chooses a random size ***c***, and then selects at random a sample of ***c*** instances of

> the use of *s*, each with its French translation [in the target language]. We imagine, further that he includes in his entry for *s* a list consisting of all the translations that occur at least once in his random sample. The probability that he will, in this way, obtain the list t_1, ... t_m is p(t_1, ... t_m | s)
>
> ...

Although, Brown et al. (1993b) provided a mathematical account to incorporate dictionary information into an SMT system, they did not empirically test the effects of the “effective multiplier" in a standard machine translation task setup evaluated using Word Error Rate (WER) or other machine translation evaluation metrics. However, they did show that the effective multiplier generally has a greater effect on low frequency words (<5).

2.3.2 Augmenting Manual Dictionaries to Improve Statistical Machine Translation

Vogel & Monson (2004) showed that a statistical machine translation system can be improved when the training data has been extended with dictionary entries and their various forms transformed using simple morphological rules (e.g. plural forms and verb inflections). They augmented Chinese-English dictionary entries with new English translations by (i) first automatically generating plural forms of the noun entries and adding the definite and indefinite determiners and generating verb forms by inflecting them with -s, -ed, -ing and with the infinitive form ‘to’ and (ii) then filtering out word forms generated in step (i) if they do not appear in a large monolingual English corpus.

$$p(s|t) = \prod_j \sum_i p(s_j|t_i) \tag{2.12}$$

The probabilities of each lexical entry is assigned by calculating the product over the probability of the source words *s* and the probability of each source word s_j is the sum of the translation probabilities of the source word s_j given its t_i, where *p(s_j | t_i)* is the probabilities generated from the word/phrase alignment process. However, there is no indication of how the probabilities are assigned if the lexical entry does not appear in the bilingual data which was used to generate the word/phrase alignments. Additionally, they suggested that the probability of the lexical entries can be renormalized and it might help to improve the NIST scores when only the lexicon (without bilingual corpus) is used to train the model.

Simply by adding the lexicon to the baseline system using the probabilities assigned as mentioned above, Vogel & Monson (2004) achieved improved NIST scores and the system is further improved using the augmented dictionary.

Vogel & Monson (2004) leveraged on the isolating (non-inflecting) nature of Chinese (source language) which simplifies the probability assignments of the augmented dictionary without considering multi-to-multi translations since only the English translations can be augmented. However, their best attempt at adding the augmented dictionary showed statistically significant[8] but minimal improvement (+0.38) to the baseline system, from 5.40 to 5.78 BLEU. This is in the scenario where they have trained the system using only the augmented dictionary and a large monolingual corpus for the language model; they did not use any parallel data other than the augmented dictionary.

Since the dictionary augmentation only increases the word/phrase alignment fertility when translating from the isolating language (Chinese) to the inflectional one (English), the evaluations were uni-directional from Chinese-English.

2.3.3 Grouping Multi-Word Expressions in Statistical Machine Translation

As Statistical Machine Translation evolved from word to phrase based approaches, the alignment models are still inherently based on word to word inferences (Vogel et al., 1996; Och & Ney, 2003). Lambert & Banchs (2006) proposed the idea of grouping Multi-Word Expressions (MWE) prior to training the SMT system in the hope that the alignments created between the source and target language text depend on linguistic constituents instead of contiguous ngrams that acknowledge ngrams with partial constituents as valid phrases to be used in the decoding process. They motivate their proposal with the following example:

The phrase “fire engine" is a fixed expression that is translated to “camion de bomberos" in Spanish, and the viterbi word-to-word alignments will probably provide the following word alignments *“camion<->truck", “bomberos<->firefighters", “fugeo<->fire" and “maquina<->engine"* which will not yield the “camion de-bomberos" translation since there are no combinations of 1-to-1 alignments which will lead to “camion de bomberos".

Intuitively, the example does behoove the grouping of MWEs prior to SMT training. But they overlook the basic assumption that word alignments are performed at sentence level and sub-phrasal word-to-word alignment occurs because the phrase appears on both the source and target language; given that there are occurrences of the phrase pair in the parallel corpus, there will be probability assigned to ‘*‘bomberos<->fire"* but the probability will be lower than *“bomberos<->firefighters"* that is a natural dictionary entry independent of the context. Moreover, given a well-built language model, the SMT should provide a low perplexity to the contiguous

[8] at 95% confidence interval (Zhang et al., 2004a)

phrases if they are observed in the monolingual corpus.

Furthermore, Lambert & Banchs (2006) overlooked the translation model training after the word alignment process in phrase-based machine translation which capitalizes on the alignment '*consistency*' which essentially requires asymmetric alignment points to extract 'consistent phrases' that are saved in the phrase table used in decoding. By pre-identifying the asymmetric alignments and single-tokenizing them, the phrase extraction algorithm has lesser information when building the phrase table.

Lambert & Castell (2004) introduced a novel approach to extract bilingual MWEs extraction using asymmetric word alignments between the source and target language phrases (i.e. non-exclusive one-to-one alignments between contiguous phrases). These asymmetric phrases are then scored by their minimal probability computed using their relative frequency between the source and target phrase (i.e. the $argmin$ between relative frequency of the source phrase and the target phrase). The top scoring candidates are heuristically filtered using language specific rules.

Using the automatically extracted bilingual MWEs as a dictionary, Lambert & Banchs (2006) single-tokenized all instances of those MWEs in the SMT training corpus, e.g. whitespaces within the phrase *"camion de bomberos"* would be replaced by underscore *"camion_de_ bomberos"* and the components of SMT would treat the phrase as a single token instead of a multi-token entity.

When comparing the results of the Spanish-English phrase-based model built on a corpus with single-tokenized MWEs against a baseline system trained on normal text, Lambert & Banchs (2006) reported a slight decrease in BLEU scores. They reported a drop of -3.0 BLEU (ES-EN: 54.7 to 51.7) and -0.2 BLEU (EN-ES: 47.2 to 47.0). The significance of the difference in BLEU scores is unreported thus the effect of grouping MWEs in SMT is inconclusive.

2.3.4 Domain Adaptation for Lexically Informed Statistical Machine Translation

Koehn & Schroeder (2007a) typified the domain adaptation scenario in machine translation systems where we have a vast amount of monolingual and parallel data in a generic domain but limited amount of in-domain data. The shared tasks held by the popular Workshop for Machine Translation (WMT) across the years follow the same setup to train machine translation systems using generic domain data and tune the systems using domain-specific data (Callison-Burch et al., 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015).

Using the English-French portion of the Europarl corpus[9] and NewsCommentary corpus[10], Koehn & Schroeder (2007a) trained several French to English factored translation models and language models using various combinations of the in- and out-domain data and tested them on a test set made of news commentaries:

- Training a single translation model and language model using only the Europarl corpus (25.11 BLEU)
- Training a single translation model and language model using only the NewsCommentary corpus (25.88 BLEU)
- Training a single translation model and language model using both the Europarl and NewsCommentary corpora (26.69 BLEU)
- Training a single translation model using both the Europarl and NewsCommentary corpora and using only the NewsCommentary in the language model (27.46 BLEU)
- Training a single translation model using both the Europarl and NewsCommentary corpora and using an interpolated language model trained with both corpora (27.12 BLEU)
- Alternative decoding using a single translation model trained on both in- and out-domain corpora and two separate language models trained using only the Europarl corpus and only the NewsCommentary corpus (27.30 BLEU)
- Alternative decoding using a single language model trained on the in-domain data and two translation models trained separately using only the Europarl corpus and only the NewsCommentary corpus (27.64 BLEU)

The best results was obtained using alternative path decoding (Birch et al., 2007) from two different translation models and a single language model trained using the in-domain data.

One possible reason for the difference in the scores achieved by alternative path decoding using two translations models (27.64 BLEU) and simply combining the data (26.69 BLEU) is the access to varied factor probabilities from different models. When the factor probabilities are computed separately the normalizing denominators are domain dependent, which will result in varying ranks in the translation probabilities of in-domain phrases. However, the +1.05 BLEU improvement was neither tested for statistical significance nor manually evaluated; it may not reflect true improvement in translation quality.

[9] 37 and 34 million French and English tokens respectively

[10] 1.2 and 1.1 million French and English tokens respectively

Relatively, the simple data concatenation of the in-domain data with the generic corpus achieved a +1.58 BLEU improvement (25.11 to 25.88 to 26.69). Also, the same improvement is reflected when only the in-domain data was used to train the language model.

Hypothetically, the concatenation of the in-domain data with general domain data could have effectively introduced domain specific words that might not have been in the generic corpus or increase the counts of the domain specific words. These findings empirically reflect the notion of “effective multiplier" that Brown et al. (1993b) theoretically proved; SMT domain adaptation can be achieved by influencing the frequencies in the in-domain lexicon.

2.3.5 Training Phrase-Based SMT Models using Domain-Specific and Generic Corpora and Dictionaries

Wu et al. (2008) enumerated several scenarios where researchers have access to one or more of the following resources (i) a generic parallel corpus, (ii) an in-domain dictionary, (iii) an in-domain target language corpus and (iv) an in-domain source language corpus. Wu et al. (2008) developed a heuristics-based domain adaptation algorithm approach that checks which of the resources is/are available before training the model differently depending on the available resource(s).

Their default assumption is that there is always a generic parallel corpus and an in-domain dictionary. Wu et al. (2008) developed a domain adaptation algorithm for SMT that follows the following steps:

1a. ***if an in-domain target language corpus is available***, train two separate language model using the in- and out-domain target language corpus and use the additional language model as an extra feature function in the decoding process. Using the extra language model feature, train a phrase-based system.

1b. ***otherwise*** train a standard phrase-based system on the out-domain parallel texts and in-domain dictionary.

2. ***if an in-domain source language corpus is available***, translate the source language corpus with the system built in step 1, then append the source and translation to the generic parallel text and dictionary to retrain the model and evaluate the new model on a development set. Then repeat the translation of the source language corpus and retrain the model until there’s no improvement made to the development set.

The recursive retraining in step 2 is similar to the approach introduced by Ueffing et al. (2007). To combine the lexical information from both the generic corpus and the in-domain dictionary, Wu et al. (2008) created two phrase tables separately from the generic corpus and the in-domain dictionary. They experimented with four different ways to assign the probabilities of the dictionary phrase table:

i. ***simply adding the dictionary as additional data to the generic corpus***

ii. ***using uniform probabilities*** ($1/n$, where n is the no. of possible translations for a source phrase)

iii. ***using constant probabilities*** (as reported in the paper, they set the constant to 1 for their experiments)

iv. ***using corpus probabilities*** when an in-domain source corpus is available[11]

When translating from Chinese to English[12] using a model trained on the CLDC Chinese-English bilingual corpus[13] (**generic corpus**) and the Basic Travel Expression Corpus (BTEC) (Paul, 2006) (**in-domain corpus**) and the LDC Chinese English translation Lexicon (LDC2002L27) (**generic dictionary**) and a manually created **in-domain dictionary**[14] was used as the dictionary, their baseline system achieved 13.59 BLEU and simply adding the dictionary as additional training data improved the system by +1.93 BLEU. Using uniform, constant and corpus probabilities yielded +2.41, +2.79 and +2.13 BLEU improvement respectively over the baseline. They showed that they were able to exploit the monolingual source corpus to improve translations.

Furthermore, when they added the target language corpus, their combined system trained using the algorithm described above scored 21.75 BLEU points, a significant improvement over the baseline system trained on the generic corpus. Without the dictionary input and the transductive learning cycles, the generic corpus with the target language in-domain corpus, the model scored 17.16.

Additionally, Wu et al. (2008) analyzed the number of lexical entries and the out-of-vocabulary percentage of the various resources and concludes that:

- Using a general-domain dictionary alone (without the in-domain dictionary) can improve the system (17.16 (baseline) $->$ 19.11 BLEU)
- Using a manually crafted in-domain dictionary alone improves the system more than using a general dictionary (21.16 BLEU)

[11] they repeatedly translated the source corpus synthetically as in step 2 of the algorithm, then estimate the probabilities using the translation probabilities of the dictionary entries using the translation probabilities from the phrase table

[12] Tuned and tested on IWSLT 2006 development and evaluation data

[13] http://www.chineseldc.org/doc/CLDC-LAC-2003-004/intro.htm

[14] They collected dictionary entries from phrase books and the dictionary was verified with a native Chinese speaker

- Using an automatically extracted in-domain dictionary[15] alone improves the system too (19.88 BLEU)
- Combining the generic and manually-crafted in-domain dictionary improves the system further than just the in-domain dictionary alone (21.34 BLEU)
- Combining the generic and automatically extracted in-domain dictionary improves the system but marginally (20.49)

Wu et al. (2008) empirically tested various improvements that different resources can make to an SMT system and showed significant BLEU improvements over their baseline system. They have also tried various ways to incorporate lexical information from manually and automatically extracted dictionaries into an SMT system.

However, the evaluation data and the host of resources they have used in their experiments are not openly available, preventing other researchers to verify and improve upon their work.

2.3.6 Improving Statistical Machine Translation Using Domain-Specific Bilingual MWEs

Motivated by the research reported in Lambert & Banchs (2006) and Wu et al. (2008) work on using MWEs in SMT and in Koehn & Schroeder (2007a) on improvements made to domain-adapted MT, Ren et al. (2009) experimented with three different ways of integrating lexical knowledge from MWEs into SMT; viz

i. ***Adding a bilingual MWE dictionary as additional data before training the SMT models*** (`BiMWE`)

ii. ***Using a binary feature function that uses 1 to represent the existence of at least one MWE translation*** matching entries in the MWE dictionary and

iii. ***Adding the MWE dictionary as an additional phrase table and assigning all feature probabilities to 1*** for alternative path decoding.

They evaluated the methods using two Chinese-English domain-specific patent corpora, one in the Traditional Chinese Medicine (TCM) domain and the other in the chemical industry domain. The training corpora were of similar sizes with 120,000 sentences that contains around 4.5 tokens for each sentence. The models were tuned and tested on a development and test set that comprised 1,000 sentences (30-40,000 words in total).

All three methods of integrating MWE attained marginal improvement ($<$ +1.0 BLEU) over their baseline phrase-based system that achieved 26.58 BLEU in the TCM domain.

[15] They extracted the in-domain dictionary from BTEC using the methods described in Wu & Wang (2007)

In their error analysis, they manually evaluated the automatically extracted MWEs and found that only 76.69% are correct. They introduce a Log-Likelihood Ratio (LLR) scoring mechanism with Gaussian priors to filter the noisily extracted MWE list and retrained their MT model with `BiMWE` and found that the system improved from 26.61 to 27.15 BLEU. However, it is still marginally better than their baseline system without MWEs. Similarly, all three MWE integration methods with the filtering mechanism had marginal gains ($<$ +0.5 BLEU) over their baseline (18.82 BLEU) in the chemical industry domain.

Ren et al. (2009) concluded their experiments with "to our disappointment, however, none of these improvements are statistically significant".

Previous Work		**+Dict**	**+MWE**	**Single-tok**	**In-Domain**	**Domain Adapt.**	**Passive**	**Intrusive**	**Pervasive**	**+BLEU (sig.)**
Vogel & Monson	(2004)	✓				✓	✓			✓
Lambert & Branch	(2006)		✓	✓						
Koehn & Schroeder	(2007)	✓				✓	✓	✓		✓
Wu et al.	(2008)	✓				✓	✓			✓
Ren et al.	(2009)	✓	✓			✓	✓	✓		
Pal et al.	(2010)		✓				✓			✓
Tsvetkov & Wintner	(2012)		✓		✓		✓			✓
Skadins et al.	(2013)	✓				✓	✓	✓		✓
Simova & Kordoni	(2013)		✓	✓	✓		✓			✓
Meng et al.	(2014)	✓				✓	✓	✓	✓	✓
Tan and Pal	(2014)	✓	✓				✓			
Hellrich & Hahn	(2015)	✓				✓	✓			
Tan et al.	(2015)	✓					✓		✓	✓
This Thesis	(2016)	✓	✓	✓	✓		✓		✓	✓

Table 2.1: A Comparison of Previous Work in Integrating Lexical Resources in Statistical Machine Translation

2.3.7 A Overview of Integrating Lexical Information in SMT

Table 2.1 presents an overview of the state of art in integrating lexical information in statistical machine translation. Based on BLEU score evaluation, the results shows that it is possible to achieve statistically significant BLEU gains albeit generally only a minor increment in absolute BLEU scores.

The **Passive** column refers to the integration lexical information by adding additional training data prior to the model training. The joint **+Dict** and **Passive** columns show that the most common approach to integrating lexical information is the passive addition of manually crafted or automatically extracted parallel dictionary or terminology to the start of statistical machine translation training process (Vogel & Monson, 2004; Koehn & Schroeder, 2007b; Wu et al., 2008; Ren et al., 2009; Skadiņš et al., 2013; Meng et al., 2014; Tan & Pal, 2014; Miñarro-Giménez et al., 2015; Tan et al., 2015a). In some studies, they also added automatically extracted Multi-Word Expressions (MWEs) to training data prior to the training process (Ren et al., 2009; Tan & Pal, 2014), indicated by the joint **+Dict**, **+MWE**, **Passive** columns in Table 2.1.

There were also studies that explores the effects of solely adding parallel MWEs (Lambert & Banchs, 2006; Pal et al., 2010; Tsvetkov & Wintner, 2012; Simova & Kordoni, 2013; Kordoni & Simova, 2004), represented by the joint **+MWE** and **Passive** columns. This is in line with the (Brown et al., 1993b) hypothesis of increasing the relevant “effective multiplier" of MWEs by increasing their frequencies in the training data. In exploring effectively multiplying the frequencies of MWEs, Lambert & Banchs (2006) experimented with single tokenizing the MWEs to trick the SMT system to treat the MWEs as a single token when extracting and decoding the ngrams; they achieved negative results from their experiments (details in Section 2.3.3).

However, as shown by the joint **+MWE**, **Single-tok** and **Passive** columns, Tsvetkov & Wintner (2012) and Kordoni & Simova (2004) repeated similar experiments on a different datasets and found that single tokenizing MWEs provides BLEU gains to SMT systems. Like in some other lexical information adding studies, they achieved statistical significant BLEU increments with little absolute BLEU gains. Focusing on only specific types of MWEs, e.g. named entities, verb compounds or phrasal verbs, Pal et al. (2010) and Simova & Kordoni (2013) showed positive results in single tokenizing MWEs to improve statistical machine translation.

Previously research on exploiting additional lexical information primarily targets the task of domain adaptation where external resources from a different domain are included to the training data to scale the machine translation system from one domain to another. In doing so, addition lexical resources from different domains are needed. Alternatively, Tsvetkov & Wintner (2012) and Kordoni & Simova (2004) attempted to inject domain-specificity **without using external resources from another daomain**. This is usually done by capitalizing on the lexical distribution extracted from the training data. The machine translation improvements made in this thesis follow the same train of thought where that lexical resource comes from the same domain as the training data.

The **Domain Adapt.** column in Table 2.1 indicates the work focused on adapting machine translation from one domain to another while the **In-Domain** column indicates the researches focusing on improving domain specificity in machine translation using the resources extracted from the training data.

Other than passively adding lexical information to the training data, previous studies have attempted various ways of injecting lexical information in the various steps in the statistical machine translation training processing. Koehn & Schroeder (2007b) introduced the idea of using more than a single pre-trained translation model and language model that can be used as domain adaptation. In that sense, the lexical information is added not only by *passively* adding them prior to the training process affecting the translation model and the language model in a monolithic one-off manner. Rather, introducing additional alternate decoding paths with the domain specific parallel corpus and/or dictionary, the lexical information addition becomes an intrusive injection to the SMT training process. We denote such “injection” as **Intrusive** on Table 2.1[16].

Similarly, framing the additional lexical information for domain adaptation, Wu et al. (2008) experimented with various alternate decoding with a permutation of in-domain and out-domain language models and translation models. Furthermore, to isolate the dictionary from the in-domain parallel corpus, they create phrase tables solely from the in-domain and generic dictionaries separately and jointly before injecting the additional translation model into the alternate decoding step. They reported positive results from several experiment setups (See Section 2.3.5). Injecting lexical information from a different approach, Skadiņš et al. (2013) added a dense binary feature to indicate the existence of an in-domain term in the SMT training process, the simple yet effective feature showed statistically significant +6.0 BLEU improvements but there was no documentation of the absolute BLEU gains and the dataset used in their experiments. Likewise, Meng et al. (2014) proposed a term disambiguation, term consistency and term bracketing features that attempted to improve the SMT decoding process. But they achieved marginal and statistically insignificant BLEU increments over the baseline models.

The last genre of lexical information addition to statistical machine translation comes from the last step of the SMT search process (aka decoding). Since the decisions made using the additional lexical information at the decoding stage would be finalized in the machine translated output, we denote such lexical information

[16] Although the intrusive addition of lexical information remains an active field in SMT, it would not be covered beyond this overview under the limited scope of the thesis.

addition techniques at the decoding step as **Pervasive**. From the literature, there was only one previous work that enforces specific knowledge in the decoding process; when decoding, Meng et al. (2014) whenever a hypothesis just translates a source term in a possible target term, they check whether the translation exists in a bilingual term bank. Different from the encoding the existence of a recognized in-domain term as a binary feature and allowing the log-linear decoding to decide the best possible decoding path like in Skadiņš et al. (2013), the pervasive technique used in Meng et al. (2014) ensures translation consistency of a specific term.[17][18]

To make a comparison between *passive* and *pervasive* addition of lexical information, we empirically investigate the effects of both approaches on the same dataset and provide further insights on how lexical information can be reinforced in statistical machine translation. The extension of this work is described in Chapter 5, where we compare the coverage of this work in the experiments on using additional lexical information in statistical machine translation.

2.4 Terminology and Ontology

2.4.1 A Survey of Term Extraction Techniques

A **term** is the *designation of a defined concept in a special language by a linguistic expression*; a term may consist of one or more words. A **terminology** refers to the set of terms representing the system of concepts of a particular subject field (ISO 1087). The International Organization of Standardization (ISO) history of terminology traces back to Wüster's (1969) seminal article on *Die vier Dimensionen der Terminologiearbeit*[19] which the ISO Technical Committee 37 (ISO/TC 37) builds upon in providing the common standards related to terminology work.

A later formulation states that a term is *any conventional symbol representing a concept defined in a subject field*; a terminology is the aggregate of terms, which represent the system of concepts of an individual subject field (Felber, 1984). The core characteristic of a term is defined as **termhood**, i.e. *the degree to which a*

[17] This is often referred to as "forced decoding" and can be easily enabled using the XML decoding feature using the Moses Machine Translation Toolkit

[18] We note that during our interaction with the industry partners of the EXPERT project, they have informed us that these pervasive techniques are essential when delivering high quality humanly post-edited translations using the outputs from machine translation systems. We note that in the case of commercial applications, there is no little or no reference translation to determine the BLEU score of the machine translation outputs and the only measure of "goodness" of translation comes from the satisfaction given by the clients of the commercial companies. The Volvo incident (Section 2.4.2) reiterates the necessity to look into these pervasive lexical information addition techniques and more importantly gain insights how the research community should change our notion of machine translation evaluation to suit actual industry needs.

[19] The Four Dimensions of Terminological Work

linguistic unit is related to a domain-specific context (Kageura & Umino, 1996). In the case of multi-token terms, additional substantiation is necessary to check its **unithood**, i.e. *the degree of strength or stability of syntagmatic combinations and collocations* (Kageura & Umino, 1996).

Single token terms can be perceived as a specialized vocabulary[20] that is used specifically in a domain. The surface word representing the single token term is often polysemous and the usage of the term within a specialized domain may narrow down the set of possible senses or single out a disambiguated sense of the word. For example, the term "*classifier*" can refer to:

(i) a morpheme used to indicate the semantic class to which the counted item belongs, or

(ii) a pre-trained model to identify/distinguish different classes within a dataset.

The first definition is mainly used within linguistic research, the second within the machine learning domain. However, when *"classifier"* is used in computational linguistics, its usage is ambiguous. The latter definition of classifier tends to be used more often than the former.

In English, terms are more often multi-word expressions (MWE), primarily nominal phrases, made up of a head noun and its complement adjective(s), prepositional clause(s), or compounding noun(s). Commonly, a complex term can be analysed in terms of a head with one or more modifiers (Hippisley et al., 2005).

Rule-based Term Extraction

The linguistic properties of a term can be characterized by its syntactic context. Previous approaches to term extraction use these linguistics properties in form of Part-Of-Speech (POS) patterns. For example, Justeson & Katz (1995) and Daille (1996) used the following POS patterns to extract nominal phrasal terms:

EN: ((Adj|(Noun)+|((Adj|Noun)*(NounPrep)?)(Adj|Noun)*)Noun$_{head}$

FR: Noun$_{head}$ (Adj|(Prep(Det)?)?Noun |V$_{inf}$)

In the case of English, the compulsory head noun is in the final position preceded by its modifiers whereas in French, it is in the first position followed by its modifiers. The multi-word nature of Romance languages produces more terminological phrases, whereas for Germanic languages, the compounding nature of nouns derives more single token lexicalized terms. For example, an equivalent POS pattern for German would have to consist of a combination of POS and morphemic pattern:

[20] aka. domain-specific vocabulary

DE: ((AC|NC)+|((AC|NC)*(Noun|Prep)?)(AC|NC)*)Noun$_{head}$

Similar to the (Adj|Noun) pattern in English, the German (AC|NC) pattern is a combination of adjective/noun with occasional connective morpheme where a connective morpheme might be necessary to join the adjacent adjectives/nouns. For example, in the German compound noun *Mausefalle* (*Mousetrap*), the *Falle* (*trap*) is head noun in the final position and the word *Maus* (*mouse*) attaches to the head noun with the *-e-* connective morpheme between the nouns.

Linguistic patterns such as these are usually used as filters to generate a list of multi-word terms of high unithood followed by further statistical measures to re-rank or reduce the list (e.g. Bourigault et al., 1996). Frantzi et al. (2000) differentiated two types of filters, viz. a close filter is strict about which strings it permits and an open filter allows more strings in the POS patterns. For example, the English pattern is an open filter that allows a wider range of multi-word term candidates than simply using a /Noun+/ that only allows delexicalized compounding nouns.

Although state-of-art term extraction systems do not solely rely on linguistic patterns, the pattern templates are used as filters to remove candidate terms from the system output (e.g. Zhang et al., 2004b; Gómez Guinovart & Simoes, 2009).

Statistical Term Extraction

The basis of all statistical properties in multi-word term extraction relies on the frequency of a token or an n-gram in a corpus. Frequency counts are combined to compute co-occurrence measures (aka. word/lexical association measures) that quantify the probabilistic occurrence of a word with its neighbouring words. Cooccurrence measures are used to estimate the propensity for words occurring together.

Psycholinguistic evidence shows that word association norms can be measured as a subject's responses to words when preceded by associated words (Palermo & Jenkins, 1964) and humans respond quicker in the case of highly associated words within the same domain (Church & Hanks, 1990).

Common co-occurrence measures, e.g. Dice coefficient, Mutual Information (MI), Pointwise Mutual Information (PMI), Log-Likelihood Ratio (LLR) and Phi-square (ϕ^2) rely on three types of frequency information; (i) the frequency of a word occurring in the corpus, (ii) the joint frequency of a word occurring with another word, (iii) the total number of words in the corpus. Formally we describe them as follows:

Let f_i be the frequency of the occurrence of a word, i

Let f_j be the frequency of the occurrence of another word, j

Let f_{ij} be the frequency of the word i and j occurring simultaneously

Let $f_{ij'}$ be the frequency of the word i occurring in the absence of j

Let $f_{i'j}$ be the frequency of the word j occurring in the absence of i

Let $f_{i'j'}$ be the frequency of both words i and j not occurring

Let N be the size of the corpus

We further simplify the notion by having $a = f_{ij}$, $b = f_{ij'}$, $c = f_{i'j}$ and $d = f_{i'j'}$.

These basic statistical properties of word co-occurrence are combined in various ways to form more complex co-occurrence measures. The common co-occurrence measures are defined as follows:

$$Dice(i,j) = \frac{2 * a}{(f_i + f_j)} \tag{2.13}$$

$$PMI(i,j) = \log a - (\log f_i + \log f_j) \tag{2.14}$$

$$MI(i,j) = \log a - \log(a+b) - \log(a+c) \tag{2.15}$$

$$\begin{aligned} LLR(i,j) = {} & a * \log a + b * \log b + c * \log c + d * \log d \\ & -(a+b) * \log(a+b) - (a+c) * \log(a+c) \\ & -(b+d) * \log(b+d) - (b+d) * \log(c+d) \\ & +(a+b+c+d) * \log(a+b+c+d) \end{aligned} \tag{2.16}$$

$$\phi^2 = \frac{(a * d - b * c)^2}{(a+b)(a+c)(b+c)(b+d)} \tag{2.17}$$

Distributional properties can be viewed as localized statistical properties. The statistical properties in the previous section make use of global count occurrences of words to extract co-occurrence statistics between words. The distributional properties relate to (i) the number of documents that a word occurs within a corpus and/or (ii) the differing counts of a word occurring across two or more corpora.

A common measure is the term frequency – inverse document frequency (*tf-idf*). The term frequency reflects the global counts of a word and the inverse documen frequency measures the spread of the word throughout the document collection. Formally,

Let f_i be the no. of times a term occurs in all documents
Let n_i be the no. of documents where the term i occurs
Let N_{doc} be the total no. of documents in a corpus
The term frequency: $tf = f_i$
The document frequency: $df = n_i / N_{doc}$
In logarithmic space, the inverse document frequency: $idf = log(N_{doc} / n_i)$
And, $tf\text{-}idf = tf * idf$

A high word frequency might favor the global statistical co-occurrence measure however if the mass of the counts comes from a low number of documents, it will reflect a low tf-idf score deeming the term to be document-specific.

Other than using the distributional properties of words within a corpus, it is also helpful to compare the distribution of words across corpora. By comparing a domain specific corpus distribution to a general corpus, we can determine the weirdness of ratio of term frequencies across the corpora (Ahmad et al., 1999). The weirder a term, the more domain-specific a term is and the more likely it is to be a term candidate to form the terminology of a specific domain. We can simply refer to the relative frequency ratio across the corpora as such:

$$weirdness(i) == \frac{f_i^D * N^G}{f_i^G * N^D} \tag{2.18}$$

where f_i^D is the frequency of a term i in a domain-specific corpus and f_i^G is the frequency of the same term in a generic corpus; N^G and N^G is the total number of tokens in the generic corpus and a domain specific corpus.

Frantzi et al. (1998, 2000) introduced a method to use both linguistic and statistical information using C-value and NC-value. They start with a set of POS patterns and a stop word list to pre-filter possible n-grams before they calculate the n-gram's termhood using the C-value metric and the concept of nested terms. Nested terms refer to those terms that appear within other longer terms and may or may not appear by themselves in the corpus (Frantzi et al., 1998), e.g. 'floating point' is a nested term because it is also found in 'floating point arithmetic'.

For non-nested terms, the C-value accounts for the length of the term candidate and its frequency. For nested terms, the C-value subtracts the average number of times the term is nested in other term n-grams. Thus if 'floating point' occurs as a nested term candidate as often or more than it does as an independent term, then it will have low C-value. Formally:

Let NG be the set of all n-grams possible from a corpus.

Let T be the set of all n-grams possible after using a POS pattern filter such that $T \subset NG$

Let t be a candidate term that is filtered from the full list of n-grams, and

Let T_N be the set of terms that contains nested terms with t such that $t \subset T_N \subset T$

Given the definition of T_N, we calculate the *C-Value* of a term t as follows:

$$C - value(t) = \begin{cases} \log |t| * f_t & if\ t\ is\ nested \\ \log |t| * (f_t - 1/f_{t_N} * \sum_{i \in T_N} f_i & otherwise \end{cases} \tag{2.19}$$

From the C-Value equation, the C-value will be high for long non-nested strings with high frequency. The limitation of the C-value is that it can only be applied to multi-word terms.

And extension of the C-value is the NC-value which accounts for the context in which the term occurs. The NC-value re-ranks the term candidates extracted from the C-values by looking into the previous words occurring before the term. This is motivated by the notion of extended terms, where the terms constrain the modifiers they accept (Sager et al., 1980). This contextual constraint manifests itself as a weight to account for the number of nested terms within in the candidate term; it is then normalized by the cumulative context weight (CCW). Formally it is defined as follows:

$$NC - value(t) = 0.8 * C - value(t) + 0.2 \sum_{c}^{C_t} f(c|t) \tag{2.20}$$

where

$C - value(t)$ is the C-value of t

C is the set of distinct context words of t

$f(t|c)$ is the conditional probability of t given that c occurs within the context

Frantzi et al.'s 2000 notion the $f(t|c)$ probability differs from the common Bayes rule derivation of $f(t, c)/f(c)$, instead, the conditional probability here is calculated by taking $f(t,c) * c_t/n_t$, where c_t is the no. of terms that have the context word c and n_t is the no. of total terms extracted and filtered from the corpus.

C-values and NC-values have proved to perform well (Zhang et al., 2008b; Lossio-Ventura et al., 2013) (Zhang et al. 2008,). However it does not measure termhood as defined by Kageura & Umino (1996). The formulation of the NC-value measures how consistently a phrase can be a term but it does not exactly contribute to select any n-grams to be a term. Inherently, the term candidate selection is handled by the POS pattern filter and the NC-value reranks the terms to further threshold the list of candidates. Although it was a solution created close to a decade ago, it is still a common algorithm used for commercial and academic term extraction[21]

[21] https://code.google.com/p/jatetoolkit/wiki/JATEIntro

2.4.2 The Importance of Terminology in Machine Translation

In today‘s globalized world, the ability to localize information into a foreign market is crucial to business expansion and machine translation (MT) is the an important means to help translate the sheer amount of information that global businesses need to process daily.

Businesses benefit from translating documents automatically by accelerating corporate communication and an MT system sensitive to domain-specific terminology is crucial in ensuring uniform and clear corporate language (Porsiel, 2011). Yet “*terminology is the biggest factor in poor translation quality*" (Warburton, 2005) and “*businesses often fail to see terminology management as a way to cut costs*" (ClientSideNews, 2006). Lionbridge (2010) reported, “*approximately 15 percent of all globalization project costs arise from rework, and the primary cause of rework is inconsistent terminology".*

The use of MT is worthwhile if the following prerequisites are given: there must be a specific corporate terminology in the largest possible scope and of the best possible quality in both the source and target languages (Porsiel, 2008). As highlighted, the two main points in ensuring that terminology is useful for improving MT requires (i) the largest possible scope and of best possible quality, i.e. *recall and precision* and (ii) in both source and target languages, i.e. *bilingual.*

In 2008, the Swedish car manufacturer Volvo was found partly to blame for a car accident which killed two school children. The expert engineer appointed by the court criticized the poor translation of a Third-Party Intermediary in the car manual on power-assisted brakes, the expert engineer stated, ‘... *there is some cause for highlighting the fact ... that the technical product information EWP S 2631 (D 710) issued by the manufacturer and translated by the import was imprecise and poorly written*‘. Volvo was fined 200,000 Euros for involuntary manslaughter and bodily injury (c.f. Hoffmeister, 2014). This called for a global concerns when handling terminology translations in technical manuals, especially in current translation workflows that incorporate machine translation with human post-editing.

2.4.3 A Survey of Ontology Induction Techniques

Aristole's metaphysical categorization of worldly concepts[22] attempted to classify concepts into a universal hierarchical structure; later known as an ontology. Ontologies are often perceived as a hierarchical organization of concepts in a tree-like structure where:

- ***Concepts*** are the atomic units that relate to each other within the taxonomy
- ***Relations*** are the links that bind the concepts
- ***Root*** is a top most concept on the hierarchy
- ***Instances*** are particular referents/instantiations of a concept.

For instance, *dog*, *mammals* and *animals* are concepts within a taxnomy. They would be organized as follows within a taxonomy:

$$\top \rightarrow \textit{animals} \rightarrow \textit{mammals} \rightarrow \textit{dog}$$

The $\top$ symbol preceding the first concept indicates the *root* of the ontology. And the $\rightarrow$ symbols specify the relations[23] between the connecting concepts. As an instantiation of the lowest level concepts, one could refer to it as a referent, e.g *Odie*, the dog from the *Garfield* comics would be a referent, i.e. an instantiation of the *dog* concept.

Traditionally, broad-coverage semantic taxonomies such as CYC (Lenat, 1995), SUMO (Pease et al., 2002a) have been manually created with much effort and yet they suffer from coverage limitations. This motivated the move towards unsupervised approaches for ontology induction and knowledge extraction (Lin & Pantel, 2001; Snow et al., 2006; Velardi et al., 2013).

Ontological induction approaches can be broadly categorized as (i) pattern/rule based, (ii) clustering based, (iii) graph based and (iv) vector space approaches.

Pattern/Rule Based Approaches

Hearst (1992) first introduced ontology learning by exploiting lexico-syntactic patterns that explicitly link a hypernym to its hyponym, e.g. "*X and other Ys*" and "*Ys such as X*". These patterns could be manually

[22] Aristotle, *Metaphysics*, I, 4, 985

[23] in this case, a *hypernym* → *hyponym* relation

constructed (Berland & Charniak, 1999; Kozareva et al., 2008) or automatically bootstrapped (Girju, 2003). These methods rely on surface-level patterns and incorrect items are frequently extracted because of parsing errors, polysemy, idiomatic expressions, etc.

In the recent taxonomy induction shared tasks at SemEval (Bordea et al., 2016, 2015b), rule based systems using substring heuristics have gained popularity and shown to attain high precision across multiple domains (e.g. Lefever, 2015; Panchenko & Biemann, 2016; Tan, 2016e). However, these systems still suffer from low recall but the precision-recall trade off favors rule-based systems due to the high false positive rates produced by non-rule based systems (e.g. Pocostales, 2016) that have achieved comparative recall rates as the rule-based ones.

Clustering Approaches

Clustering based approaches are mostly used to discover hypernym (is-a) and synonym (is-like) relations. For instance, to induce synonyms, Lin (1998) clustered words based on the amount of information needed to state the commonality between two words.[24]

Caraballo (1999) was first to combine clustering and pattern-based methods by hierarchically clustering words and assigning the hypernyms by identifying the “*A is a (kind of) B*" and “*X, Y, and other Zs*" patterns where B is considered as a hypernym of A and Z is the hypernym of X and Y.

Contrary to most bottom-up clustering approaches for taxonomy induction (Caraballo, 2001; Lin, 1998), Pantel & Ravichandran (2004) introduced a top-down approach, assigning the hypernyms to clusters using co-occurrence statistics and then pruning the cluster by recalculating the pairwise similarity between every hyponym pair within the cluster.

Besides inducing generic hypernym-hyponym taxonomies, similar clustering approaches were also applied in inducing domain-specific knowledge bases that focused on inferring relations between named entities (e.g. Hasegawa et al., 2004).

Graph-based Approaches

In graph theory (Biggs et al., 1976), similar ideas are conceived with a different jargon. In graph notation, ***nodes/vertices*** form the atom units of the graph and nodes are connected by directed ***edges***. A *graph*, unlike

[24] Commonly known as Lin information content measure.

an ontology, regards the hierarchical structure of a taxonomy as a by-product of the individual pairs of *nodes* connected by directed *edges*. In this regard, a single *root* node is not guaranteed nor a tree-like structure.

Disregarding the overall hierarchical structure, the crux of graph induction focuses on the different techniques of edge weighting between individual node pairs and graph pruning or edge collapsing (Kozareva & Hovy, 2010; Navigli et al., 2011; Fountain & Lapata, 2012; Tuan et al., 2014).

Vector Space Approaches

Semantic knowledge can be thought of as a two-dimensional vector space where each word is represented as a point and semantic association is indicated by word proximity. The vector space representation for each word is constructed from the distribution of words across context, such that words with similar meaning are found close to each other in the space (Mitchell & Lapata, 2010; Tan, 2013).

Although vector space models have been used widely in other NLP tasks, ontology/taxonomy induction using vector space models has not been popular. It is only since the recent advancement in neural nets and word embeddings that vector space models are gaining ground for ontology induction and relation extraction (Saxe et al., 2013; Khashabi, 2013).

Projecting a Hyponym to its Hypernym with a Transition Matrix

Fu et al. (2014) discovered that hypernym-hyponyms pairs have similar semantic properties as the linguistic regularities discussed in Mikolov et al. (2013c). For instance:

$$v(\texttt{shrimp})\text{-}v(\texttt{prawn}) \approx v(\texttt{fish})\text{-}v(\texttt{goldfish})$$

Intuitively, the assumption is that all words can be projected to their hypernyms based on a transition matrix. That is, given a word x and its hypernym y, a transition matrix Φ exists such that

$$y = \Phi x, \text{ e.g. } v(\texttt{goldfish}) = \Phi \times v(\texttt{fish})$$

Fu et al. proposed two projection approaches to identify hypernym-hyponym pairs, (i) uniform linear projection where Φ is the same for all words and Φ is learnt by minimizing the mean squared error of $||\Phi x\text{-}y||$

across all word-pairs (i.e. a domain independent Φ) and (ii) piecewise linear projection that learns a separate projection for different word clusters (i.e. a domain dependent Φ, where a taxonomy's domain is bounded by its terms' cluster(s)). In both projections, hypernym-hyponym pairs are required to train the transition matrix Φ.

2.5 Ontology and Translation

Another aspect of semantic knowledge within translation is the usage of ontology in human and machine translation. Beyond the flat structure of a list of domain-specific words and phrases in the terminology, an ontology provides a hierarchical structure between the words. The resulting 'tree of domain-specific knowledge' is helpful for human translation when the translators are not familiar with the domain.

The goal of ontology development is the sharing of common knowledge of the information structure across different departments working in the same organization (Musen, 1992; Gruber, 1993). Although humanly crafted broad-based ontologies (i.e. upper ontologies) that attempt to catch the *world knowledge* exist (Lenat, 1995; Pease et al., 2002a; Navigli & Ponzetto, 2012), it can be more viable to automatically create a domain-specific ontology given a domain-specific corpus in the target language. In this thesis we introduce (i) ***a novel state-of-art technique to generate hypernyms between terms to create an ontology using neural net embeddings*** and (ii) ***explore the endocentricity of hyper-hyponyms relations based on their surface string representation***. This will be discussed in Chapter 6.

We hope that the techniques and findings we have on ontology induction can be further developed beyond the scope of the thesis to support automatic ontology creation to help translators in their knowledge understanding part of their translation workflow.

Beyond aiding human translators, ontology can be provide additional knowledge to the statistical machine translation training process. A simple way to incorporate ontological knowledge is to develop word clusters instead of a full hierarchical knowledge graph. Within the word alignment phase of the SMT training steps, there is a module that performs word clustering in order to reduce sparsity of the relative distortion computation[25] (Och & Ney, 2003). In this thesis we present joint work on the ***evaluation of a new predictive exchange clustering algorithm that can be used to replace existing clustering software***, `mkcls` within the (M)GIZA++ word alignment tool that is used in the phrase-based machine translation. This will be discussed in Chapter 6.

[25] The discussion on word alignment algorithm is out of scope in this thesis but the Moses developers have an informational deck of slides describing it at `http://www.statmt.org/book/slides/04-word-based-models.pdf` (Koehn, 2009)

2.6 Summary

Integrating additional lexical information into statistical machine translation is nothing new but the marginal gains in BLEU scores across the literature is somewhat disturbing given the industrial importance of translating terminologies correctly. This thesis seeks to take a closer look at how to use additional lexical resources in the SMT training process.

Ontologies are useful knowledge resources that translators use to familiarize themselves with domain specific knowledge. The manual creation of ontology is time and labor consuming, this thesis proposes a novel approach to ontology induction using neural nets and explores the endocentricity of hypernyms within the ontology to better understand how we can effectively automatically induce high precision ontologies.

The lack of ‘semantic knowledge’ in the statistical machine translation paradigm suggests that we should look into ways to inject meaning into the probabilistic system. In this thesis, we will present joint work on scaling sub-ontological knowledge (i.e. word clusters) and investigate the improvements made by this to statistical machine translation.

Chapter 3

Resources and Experimental Setup

This chapter describes the experimental setups and resources used in this thesis.

3.1 Resources used in Machine Translation Experiments

3.1.1 Asian Scientific Paper Excerpt Corpus (ASPEC)

The Asian Scientific Paper Excerpt Corpus[1] (ASPEC) was created by Japan Science and Technology Agency (JST) in collaboration with the National Institute of Information and Communications Technology (NICT). The Japanese-English subset of the corpus comprises of a 3 million sentences from Japanese-English paper abstracts. The corpus was attends to the demand for machine translation of scientific papers given the increasing number of scientific publications globally. For tuning and evaluation, there are ~1,800 sentences held out for the development and test sets.

3.1.2 Workshop for Machine Translation (WMT) News Domain Corpus

The Workshop for Machine Translation news domain translation task provides an array of parallel and monolingual dataset for the participants. We use the English-Russian parallel data from the Common Crawl corpus, News Commentary, and Wikipedia Headlines and Yandex corpus[2] with the monolingual data from the

[1] `http://lotus.kuee.kyoto-u.ac.jp/ASPEC/`

[2] `https://translate.yandex.ru/corpus?lang=en`

News Commentary and News Crawl 2008-2014 as the training data. For development and evaluation, we the `news-test2014` dataset for tuning and the `news-test2015` dataset for testing.

3.1.3 JICST Dictionary

The JICST is a Japanese-English (JA-EN) translation dictionaries (JICST, 2004) from the Japan Science and Technology Corporation. It contains 800,000 entries for technical terms extracted from scientific and technological documents. The entries in the dictionary are manually verified to be correct and specific to the science and technology domain. On average 3-4 dictionary entries are found for each sentence in the WAT corpus development set.

This dictionary will be the main source of hand-crafted lexical information that we are attempting to inject into the data of the same domain to highlight the use of in-domain terminology influence on machine translation.

3.1.4 The Phrase-based SMT Configuration used throughout the Thesis

Unless a different experimental setting is explicitly stated, for all experiments we used the phrase-based SMT implemented in the Moses toolkit (Koehn et al., 2007) with the following experimental settings:

- `MGIZA++` implementation of IBM word alignment model 4 with grow-diagonal-final-and heuristics for performing word alignment and phrase-extraction (Koehn et al., 2003; Och & Ney, 2003; Gao & Vogel, 2008)
- Bi-directional lexicalized reordering model that considers monotone, swap and discontinuous orientations (Koehn, 2005; Galley & Manning, 2008)
- Language modeling is trained using `KenLM` with maximum phrase length of 5 with modified Kneser-Ney smoothing (Heafield, 2011; Kneser & Ney, 1995; Chen & Goodman, 1999)
- Minimum Error Rate Training (MERT) to tune the decoding parameters (Och, 2003).
- For English translations, we trained a true-casing model to keep/reduce tokens' capitalization to their statistical canonical form (Wang et al., 2006; Lita et al., 2003) and we recased the translation output after the decoding process

Additionally, we applied the following methods to optimize the phrase-based translation model for efficiency:

- To reduce the size of the language model and the speed of querying the model when decoding, we used the binarized trie-based quantized language model provided in `KenLM` (Heafield et al., 2013; Whittaker & Raj, 2001)
- To minimize the computing load on the translation model, we compressed the phrase-table and lexical reordering model (Junczys-Dowmunt, 2012)

3.2 Resources used in Terminology and Ontology Extraction

3.2.1 Disease Names and Adverse Effect (DNAE) Corpus

The DNAE corpus[3] is annotated with a subset of the medical terms of the following medical term databases:

- Medical Subject Headings (MeSH)
- Medical Dictionary for Regulatory Activities (Med-DRA)
- International Classification of Diseases (ICD-10
- Systematized Nomenclature of Medicine–ClinicalTerms (SNOMED CT)
- Unified Medical Language System (UMLS)

The corpus was used by Hanisch et al. (2005) to evaluate their ProMiner term exractor. We use the same resource to compare our approach with Hanisch et al. (2005).

Corpus Statistics		
No. of Documents	400	
No. of Sentences	4,380	
No. of Disease Mentions (no. of unique)	1,428	(628)
No. of Adverse Effect Mentions (no. of unique)	813	(440)
No. of Total D+A Mentions (no. of unique)	2,241	(1068)

Table 3.1: Disease Names and Adverse Effect (DNAE) Corpus Statistics

Table 3.1 presents the document statistics of the DNAE corpus. There are 400 documents that contain 4,380 sentences in total. There are a total of 2,241 manually annotated Disease and Adverse Effects mentions, of which 1,068 are unique.

3.2.2 WikiFood Corpus

The food domain corpus (WikiFood) was built for an ontology induction task at SemEval-2015[4] (Tan et al. 2015). The corpus contains 869 food terms and the relevant Wikipedia articles that contain these terms. Of

[3] https://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/downloads/corpus-for-disease-names-and-adverse-effects.html

[4] http://alt.qcri.org/semeval2015/task17/index.php?id=data-and-tools

those terms 752 terms are multi-words and from the 752 multi-words, we extracted 42,851 sentences (1,207,677 tokens) that contains the multi-word terms.

3.2.3 SemEval TaxEval Ontologies

The SemEval-2015 Taxonomy Extraction Evaluation (TaxEval) task addresses taxonomy learning without the term discovery step, i.e. the terms for which to create the taxonomy are given (Bordea et al., 2015a). The focus is on creating the hypernym-hyponym relations. We will be using this dataset to evaluate our hypernym induction system using the *'is-a'* vector.

In the TaxEval task, ontologies are evaluated through comparison with gold standard taxonomies. The gold standards used in evaluation are the *ChEBI ontology* for the chemical domain (Degtyarenko et al., 2008), the *Material Handling Equipment taxonomy*[5] for the equipment domain, the *Google product taxonomy*[6] for the food domain and the *Taxonomy of Fields and their Different Sub-fields*[7] for the science domain. In addition, all four domains are also evaluated against the sub-hierarchies from the WordNet ontology that subsumes the Suggested Upper Merged Ontology (Pease et al., 2002a).

[5] http://www.ise.ncsu.edu/kay/mhetax/index.htm

[6] http://www.google.com/basepages/producttype/taxonomy.en-US.txt

[7] http://sites.nationalacademies.org/PGA/Resdoc/PGA_044522

Chapter 4

Information Theoretic and Language Model based Term Extraction using Pointwise Mutual Information

This chapter[1] describes the first main contribution of the thesis, i.e. a novel term extraction technique we have developed based on information theoretic Pointwise Mutual Information (PMI) and language model probabilities. We dub the method PMI_{LM}.

The rest of this chapter will first introduce (i) the description of our new monolingual extraction technique, (ii) an intrinsic evaluation of our extraction method against a gold standard terminology, (iii) an extrinsic evaluation of the term extraction in an information retrieval query task and (iv) the bilingual extension of our term extractor based on phrase alignments, iv) an extrinsic evaluation of the automatically extracted bilingual terms on two diverse machine translation tasks.

We show that our monolingual PMI_{LM} term extractor has comparable results with state-of-art term extraction systems in both intrinsic and extrinsic evaluations. Our results of adding terms extracted using the bilingual extension of the PMI_{LM} model agree with previous studies in using additional lexical information in statistical machine translation.

[1] Part of the search presented in this thesis has been previously published in Tan (2016a,b)

4.1 Pointwise Mutual Information

The Pointwise Mutual Information (PMI) measures the association between two discrete random variables by dividing their joint probability[2] by their individual probabilities[3]. Mathematically,

$$\begin{aligned} PMI(x,y) &= \log \frac{p(x,y)}{p(x)p(y)} \\ &= \log p(x,y) - \log p(x) - \log p(y) \end{aligned} \tag{4.1}$$

Applying the chain rule on the PMI equation, we can measure the association between more than two discrete random variables[4] as such:

$$\begin{aligned} PMI(xy,z) &= \log \frac{p(xy,z)}{p(xy)p(z)} \\ &= \log p(xy,z) - \log p(xy) - \log p(z) \\ &= PMI(y,z) + PMI(z,y|x) \end{aligned} \tag{4.2}$$

In the context of terminology extraction, we can measure the collocation coherence between two words by measuring their PMI Bouma; Eichler & Neumann (2010). For example, to measure the PMI value of the phrase ‘floating point’

$$\begin{aligned} PMI(floating\,point) &= \log \frac{p(floating\,point)}{p(floating)p(point)} \\ &= \log p(floating\,point) - \log p(floating) - \log p(point) \end{aligned} \tag{4.3}$$

where $p(floating, point)$ is the probability of the phrase ‘*floating point*’ divided by the probability of the word ‘*floating*‘ and ‘*point*‘ occurring individually.

Extending the PMI application to natural text with the chain rule, we can easily find the PMI phrases with more than two words. E.g. for the phrase ‘*floating point arithmetic*’, we can recursively sum the PMI scores of the phrase and the sub-phrases and its sub-phrase.

$$\begin{aligned} PMI(floating\,point, arithmetic) = \log p(floating\,point\,arthimetic) \\ - \log p(floating\,point) - \log p(arthimetic) \end{aligned} \tag{4.4}$$

[2] i.e. probability of their coincidence

[3] assuming statistical independence between the variables

[4] See prove on https://en.wikipedia.org/wiki/Pointwise_mutual_information#Chain-rule_for_pmi

4.2 Term Extraction Using Language Model Pointwise Mutual Information

The calculation of word probabilities and PMI described so far is based on raw counts of a word or term (in Chapter 3 Section 2.5.1). N-gram language models, that are also based on counts, have developed and applied to other NLP applications such as speech processing and machine translation (e.g. Och et al., 2004; Kirchhoff & Yang, 2005; Schwenk et al., 2012; Lembersky et al., 2012; Chelba et al., 2012).

A major advantage of using a language model is the possibility of accounting for unknown words using interpolation and smoothing techniques (Chen & Goodman, 1996; Chelba et al., 2010). By using a language model, we avoid the need to optimize n-gram counting when implementing the term extraction algorithm, especially when very fast implementations of language models already exists (Heafield, 2011).

Instead of the using phrasal and word probabilities to compute the PMI score of a term,

$$PMI(term) = \log p(w_1, ..., w_n) - \log p(w_1, ...w_{n-1}) - \log p(w_n) \tag{4.5}$$

we can use a backoff language model as follows:

$$\begin{aligned} PMI_{LM}(term) &= \frac{P_{LM}(w_1, ..., w_n)}{P_{LM}(w_1, ..., w_{n-1}) P_{LM}(w_n)} \\ &= P_{LM}(w_1, ..., w_n) - P_{LM}(w_1, ..., w_{n-1}) - P_{LM}(w_n) \end{aligned} \tag{4.6}$$

where a $term$ is made up of words $w_1, ..., w_n$ and $P_{LM}(w)$ is the language model probability provided by a pre-trained language model that approximates to $\log p(w)$.

Similar to the C-value and NC-value (Frantzi et al., 1998), our approach is limited to multi-word expressions. Our approach is (i) a more flexible approach than C-value that allows querying a smoothed language model to retrieve probabilities and (ii) relies on the backoff probabilities that have the ability to score unknown words. The traditional C-value would count co-occurrence instances without smoothed counts and would not handle unknown words.

4.3 Evaluating PMI_{LM} against a Gold Standard

Intrinsically, we will evaluate the PMI_{LM} against a gold standard terminology in the Biomedical domain. Gurulingappa et al. (2010) created the *Disease Names and Adverse Effect* (DNAE) corpus with manually annotated medical terms and empirically evaluated a rule-based term extractor, ProMiner, (Hanisch et al., 2005). The ProMiner system uses a pre-processed hand-crafted synonym dictionary to identify potential name occurrences in the biomedical text and associate protein and gene database identifiers with the detected matches.

Hanisch et al. (2005) considered a subset of the medical terms of the following medical term databases that occurred in the DNAE corpus as the positive examples of the terms that a term extractor should extract. We use the same subset to compare our approach with Hanisch et al. (2005).

System	Methodology	Precision	Recall	F-score
ProMiner	Unsupervised Rule-based	**0.18**	0.76	**0.29**
PMI_{LM} (global)	Unsupervised Probabilistic	0.17	**0.89**	**0.29**
PMI_{LM} (local)	Unsupervised Probabilistic	0.15	**0.83**	0.25
BiOpeNER	Supervised Off-the-shelf	**0.55**	0.44	0.49
LM-PMI Linear	Supervised using Probabilistic Feature	0.51	0.56	0.53
LM-PMI XGBoost	Supervised using Probabilistic Feature	0.53	**0.58**	**0.55**

Table 4.1: Comparison of PMI_{LM} against State-of-Art Term Extractors

To directly compare our unsupervised approach (i.e. without the use of the mention annotations from the corpus) against the rule-based ProMiner, we use the PMI_{LM} in two different settings[5] (i) we extracted all possible unigrams to fivegrams from each sentence and filter the top 1200 unique terms weighted by the their 5-gram PMI_{LM} and we search through all documents to label the mentions of these terms, i.e. PMI_{LM} (global) and (ii) we extracted the top 2 unigrams to fivegrams from each sentence and we considered them as the biomedical term that our system identified, i.e. PMI_{LM} (local). Using these two settings, we extracted the terms from the full DNAE corpus and evaluate against the manually annotated terms.

In the unsupervised scenario, our PMI_{LM} (global) term extractor achieved a similar F-score (0.29) to ProMiner , however our PMI_{LM} (local) system underperformed (F-score=0.25) as compared to ProMiner due to poor precision. This is indirectly caused by the heuristic we impose when extracting the top 2 PMI_{LM} scoring

[5] Note that the PMI_{LM} approach is limited to multi-word expressions, i.e. ngram orders of two and above. We fall back to the language model probabilities when computing the unigram term scores.

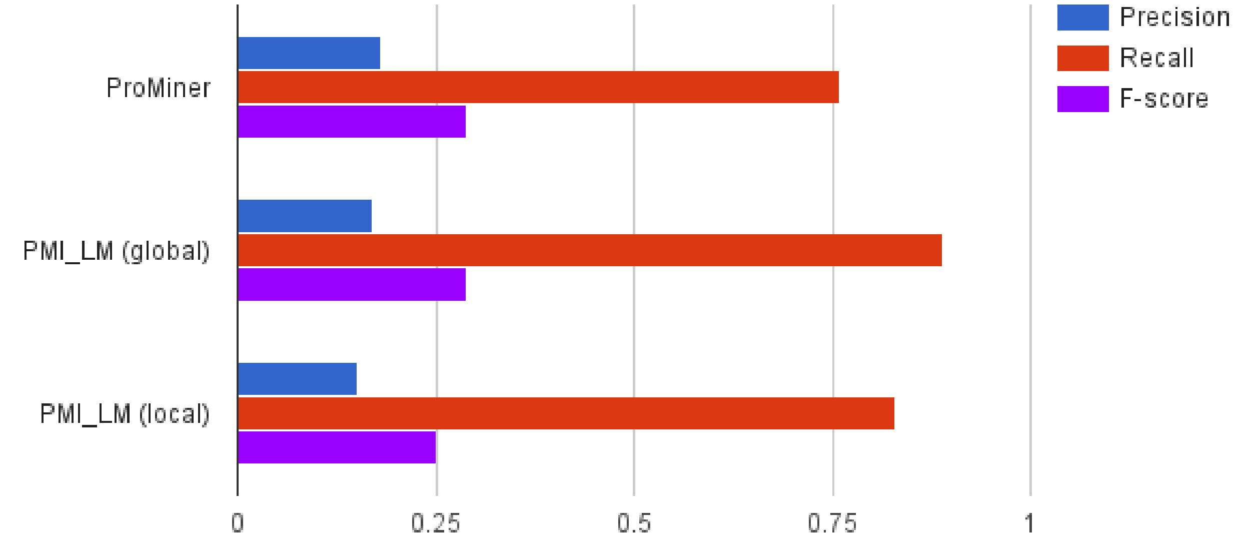

Figure 4.1: Precision, Recall and F-score of Unsupervised Term Extraction of PMI_{LM} and ProMiner on the DNAE Corpus

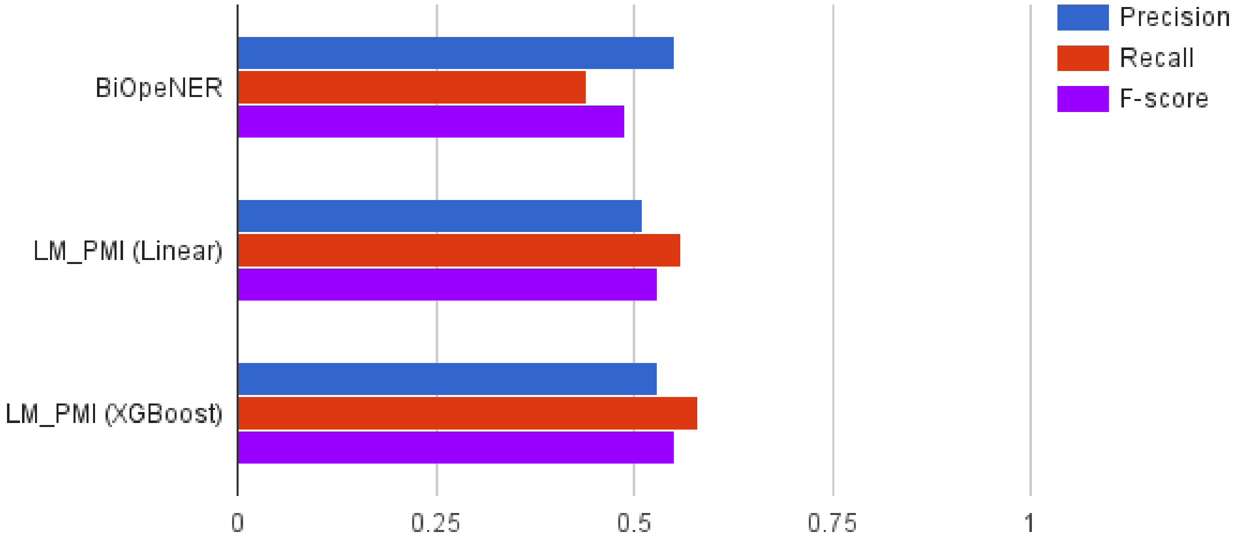

Figure 4.2: Precision, Recall and F-score of Supervised Term Extraction of PMI_{LM} and ProMiner on the DNAE Corpus

terms per sentence, when in fact there can be 0 or 1 or more than 2 terms within the sentence. On the plus side, we have achieved better recall in both local and global settings at 0.89 and 0.83 respectively as compared to ProMiner that has a recall rate of 0.76.

In a supervised scenario, we use our PMI_{LM} score of all possible ngrams as a sparse feature set to train a linear classifier[6] and an XGBoost ensemble classifier[7] (Chen & Guestrin, 2015) to identify the terms from the manually annotated DNAE corpus. We randomly split the sentences with and 90% train and 10% test set and report the harmonic cross-validation precision, recall and F-scores in Table 3.2. We compare our results against the BiOpeNER (Ellendorff et al., 2014) system that also reported harmonic cross-validation scores.

Although our system did not achieve the state-of-art precision scores that BiOpeNER had, both the linear logistic and ensemble classifier performed better on the recall scores and F-scores. The BiOpeNER system achieved 0.55 precision, 0.44 recall and 0.49 F-score whereas our best setup with the XGBoost achieved 0.53 precision, 0.58 recall and 0.55 F-score.

4.4 Evaluating PMI_{LM} in Information Retrieval

As an extrinsic evaluation, we test the PMI_{LM} term extraction approach on the food domain corpus (WikiFood) that was built for an ontology induction task at SemEval-2015. The corpus contains 869 food terms and the relevant Wikipedia articles that contain these terms. Of those terms 752 terms are multi-words and from the 752 multi-words, we extracted 42,851 sentences (1,207,677 tokens) that contains the multi-word terms.

To access the probabilities for calculating PMI_{LM}, we trained a 5-gram language model on the corpus and we evaluate the PMI_{LM} accuracy against the traditional C-value score in extracting terms from the corpus.

We extract the top five term candidates each using PMI_{LM} and C-value to match against the 1 correct term per sentence. We evaluate the metrics by calculating the accuracy of the top ranked term candidates for each sentence and matching them against the correct term for that sentence. Since the experimental task is structured more like an information retrieval task of candidate ranking, we use the mean reciprocal rank (MRR) to evaluate the ranking efficiency of the term extraction metrics. The mean reciprocal rank is calculated by averaging the ranks of the retrieved candidates against all possible candidates.

Table 4.2 presents the results of the experiment on term extraction for the WikiFood Corpus. The PMI_{LM}

[6] Using the default linear logistic classifier from Scikit-learn(Pedregosa et al., 2011)

[7] Using `hammer.py` wrapper script from (Bechara et al., 2016b). Note that we further slice the 90% training set to 80% train and 10% development to tune the system at every fold for our ensemble system

	Accuracy (*Top1*)	Accuracy (*Top5*)	MRR
PMI_{LM}	**28.29**	**45.18**	**1.632**
C-Value	23.26	32.26	2.263

Table 4.2: Accuracy and Mean Reciprocal Rank for Term Extraction for WikiFood

metric clearly scores better in terms of accuracy to rank the correct term candidate in the top position with a mean rank of 1.63 and an accuracy of 0.28 compared to the C-value's mean rank of 2.26 and an accuracy of 0.23.

C-Value		PMI_{LM}	
Bull's-Eye Barbecue Sauce	-6.491	Burger King	-3.269
Al Steak Sauce	-7.078	Bull's-Eye Barbecue Sauce	-4.909
Kraft products	-7.397	A1 Steak Sauce	-6.216
Burger King	-8.512	Barbecue Sauce	-6.304
A1 Steak	-9.148	Kraft products	-6.675

Table 4.3: An Example Output of Ranked Term Candidates and Metrics Scores

However, we do note the low accuracy scores because a term candidate with high termhood might not necessarily be the query term that we are expecting from the sentence. For example in the sentence "*In both cases, Burger King prominently used the names of the Kraft products, A1 Steak Sauce and Bull's-Eye Barbecue Sauce, in the names of the sandwiches*". The PMI_{LM} and C-value extracted the following terms in Table 4.4.

Although the absolute scores between the two metrics are on a different scale, they are comparable because they are constricted by a probability in the logarithm space. We see that there are many valid terms extracted (e.g. "Burger King" and "Bull's-Eye Barbecue Sauce") by both metrics. However, they are not used in the accuracy computation because of the aim to build a food terminology for a taxonomy and to check against a gold standard terminology.

Unsurprisingly, the C-value favours longer terms and ranks them higher. Additionally, the target "Barbecue Sauce" has been excluded in the top 5 terms for the C-value due to its preference for nested terms and that allowed "A1 Steak" into the top 5 C-value extracted terms.

4.5 Bilingual Term Extraction using PMI_{LM}

The discussion until now considers the monolingual term extraction context where the terms are extracted from a single language using the PMI_{LM} scores. Monolingual term extraction revolves around three approaches (i) rule-based methods relying on morphosyntactic patterns, (ii) statistical methods which use association/frequency measures to determine ngrams as MWE and (iii) hybrid approaches that combine rule-based and statistical methods.

However, where bilingual term extraction techniques are concerned, they operate around two main modus operandi (I) extracting monolingual terms separately and aligning them at word/phrasal level afterwards or (II) aligning parallel text at word/phrasal level and then extracting terms.

Given the probabilistic nature of the PMI_{LM}, our bilingual term extractor using PMI_{LM} would be language independent. We simultaneously (i) extract the top N terms monolingually for both the source and target languages and (ii) perform word alignment and phrase extraction to produce a phrase-table using MGIZA++ and Moses (Koehn et al., 2003; Och & Ney, 2003; Gao & Vogel, 2008). Then we filter out the terms that do not appear on the phrase table, this will result in a dictionary of aligned terms.

In this case, we can relate to our bilingual terms as globally extracted like the unsupervised experiment setting in Section 3.1 with the assumption of a top K salient terms ranked by PMI_{LM} that should be added to the training data.

4.6 Evaluating PMI_{LM} as Additional Lexical Knowledge for SMT

As a first experiment in integrating terminological information to statistical machine translation, we passively add the bilingually extracted terms, using PMI_{LM} and phrase alignments, to the training data before the phrase-based SMT training process. We use the setup as described in Section 2.4.4.

We evaluated the integration using the training, development and evaluation data from the Workshop for Asian Translation 2014 (WAT2014) Japanese-English dataset (Nakazawa et al., 2014) from the ASPEC corpus[8] (Nakazawa et al., 2016).

Additionally, we also evaluated the additional PMI_{LM} based lexical knowledge on the Workshop for Machine Translation (WMT14) Russian-English dataset. The development and test set comprises 3000 sentences each from news articles manually translated from English to Russian. The monolingual data consists of News Commentary and News Crawl articles from 2007 to 2014. The training data from WMT15 comprises Common Crawl Corpus, News Commentary, Yandex 1M Corpus and Wiki Headlines Corpus; these were web-crawled texts automatically sentenced aligned.

[8] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

#Terms added	EN-JA	JA-EN	EN-RU	RU-EN
0	16.75	**23.91**	21.0	25.9
2,000	16.68	23.74	21.6*	26.1
4,000	16.67	23.79	20.9	26.0
6,000	16.72	23.64	21.4	26.0
8,000	16.76	23.75	**22.5***	26.7*
10,000	**16.81***	23.78	22.1*	26.5*
12,000	16.59	23.45	21.4	**26.8***

Table 4.4: BLEU Scores from Phrase-based SMT Systems with Passively Added Terminology of Incremental Sizes; (*) refers to statistically significant results at p $<$ 0.05

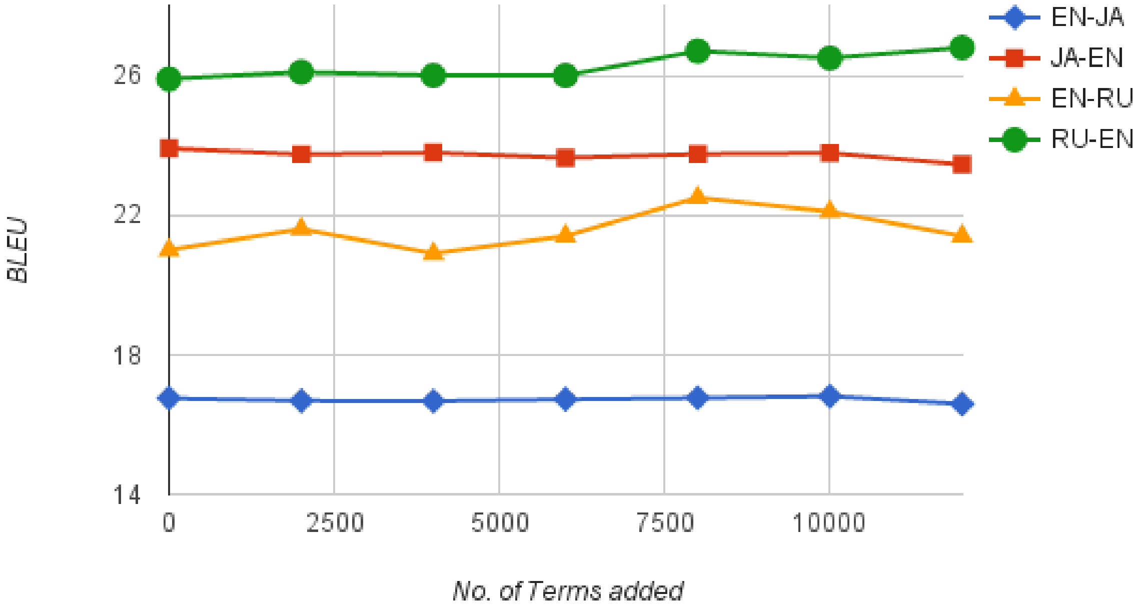

Figure 4.3: Adding Automatically Extracted Terminology of Different Sizes to SMT

Table 3.5 and Figure 3.3. presents the results of passively adding varying numbers of automatically extract bilingual terms to the training data before the statistical machine translation training process. From Table 3.5, we see that adding the top 10,000 bilingual terms to the English-Japanese data yields the highest BLEU score of 16.81 and its statistically significantly (p $<$ 0.05) better than the baseline system without the added terminology. But this is not the case of the other direction; for Japanese-English, adding the automatically extracted bilingual terminology does not improve upon the baseline but it is also not statistically significantly worse than the baseline.

As for English to Russian, the added terminology performs best (BLEU=22.5) when 8,000 terms are added. In the reverse direction, the BLEU scores can be raised from 25.9 to 26.8 by adding the 12,000 extracted terms.

The results obtain confirms the previous work on adding lexical information to the training set in the attempt of improving machine translation. Most often, previous work as described in Chapter 2 (esp. Section 2.4.3), achieved marginal BLEU gains when passively adding additional lexical information to the training data in a non domain adaptation scenario where the terminology and corpus comes form the same domain.

4.7 Summary

In this chapter, we have proposed a novel information theoretic statistical term extraction technique based on a language model (PMI_{LM}) and we have intrinsically and extrinsically evaluated the quality of the terms extracted using the PMI_{LM} based approach.

We show that the PMI_{LM} provides state-of-art performance in extracting biomedical terms in both an unsupervised and a supervised scenario. We have also shown that given the optimal configuration of the automatically extracted terminology and passive integration of the term pairs in statistical machine translation, it is possible to achieve statistically significant though marginal BLEU score improvements.

In Chapter 4, we will describe more experiments to gain insights on how adding manually crafted and automatically extracted dictionary/terminology resource can improve phrased-based statistical machine translation. In the experiments reported in this chapter, ading automatically extracted lexicon passively provides statistically significant but marginal BLEU gains.

Chapter 5

Using Terminology to Improve Machine Translation

Phrase-based Statistical Machine Translation (SMT) obtains the best translation by maximizing the conditional probability of the source sentence given the foreign sentence, $p(e|f)$. Using Bayesian deductions, we require the conditional probability of the foreign sentence given the source sentence and the a priori probability of the translation, $p_{LM}(e)$ (Brown, 1993); mathematically

$$arg \max p(e|f) = arg \max p(f|e) \cdot p_{LM}(e) \tag{5.1}$$

State-of-art SMT systems rely on (i) large bilingual corpora to train the translation model $p(f|e)$ and (ii) monolingual corpora to build the language model, $p_{LM}(e)$ (See Chapter 2, Section 2.2).

There are two primary approaches to the use of bilingual dictionary resources in statistical machine translation: (i) the ***passive*** approach of appending the parallel training data with a bilingual dictionary and prior to traiing the SMT (ii) the ***pervasive*** approach of enforcing translation as per the dictionary entries when decoding.

Previous studies have shown that both approaches of providing external lexical knowledge to statistical machine translation show the potential of improving translation quality (See Chapter 2, Section 2.4.7). In this chapter, empirically investigate the effects of both approaches on the same dataset and provide further insights on how lexical information can be reinforced in statistical machine translation.

The ***passive*** approach to improve the SMT model is to extend the parallel data with a bilingual dictionary prior to training the model. The primary motivation is to use additional lexical information for domain adaptation to overcome the this issue out-of-vocabulary words during decoding (Koehn and Schroeder, 2007; Meng et al. 2014; Wu et al. 2008). Alternatively, adding in-domain lexicon to parallel data has also shown to carry the potential to improve SMT. The intuition is that by adding extra counts of bilingual lexical entries, the word alignment accuracy improves, resulting in a better translation model (Skadins et al. 2013; Tan and Pal, 2014; Tan and Bond, 2014).

The ***pervasive*** approach to use a bilingual dictionary is to hijack the decoding process and force word/phrase translations as per the dictionary entries. Previous research used this approach to explore various improvements in industrial and academic translation experiments to enhance specifically term translation consistency.

5.1 Passive and Pervasive Lexical Injection

We view both the passive and the pervasive use of a dictionary in statistical machine translation as a type of lexically constrained statistical MT where in the passive use, the dictionary acts a a supplementary set of bi-lexical rules affecting word and phrase alignments and the resulting translation model and in the pervasive use, the dictionary constraints the decoding search space enforcing translations as per the dictionary entries.

To examine the ***passive*** use of a dictionary, we explore the effects of adding the lexicon *n* number of times to the training data until the performance of the machine translation degrades. For the ***pervasive*** use of a dictionary, we assign a uniform translation probability to possible translations of the source phrase as determined by the dictionary. For instance, in a dictionary, the English term "*abnormal hemoglobin*" could be translated to 異常ヘモグロビン or 異常血色素, we assign the translation probability of 0.5 to both Japanese translations uniformly, i.e. p(異常ヘモグロビン | abnormal hemoglobin) = p(異常血色素 | abnormal hemoglobin) = 0.5. If there is only one translation for a term in the dictionary, we force a translation from the dictionary by assigning the translation probability 1.0 to the translation.

One issue with the pervasive use of dictionary translations is the problem of compound phrases in the test sentence that are made up of component phrases in the dictionary. For instance, when decoding the sentence, "*Here was developed a phase shift magnetic sensor system composed of two sets of coils , amplifiers , and phase shifts for sensing and output .*", we fetch the following entries from the dictionary to translate the underlined multiword term:

- magnetic = 磁気

- sensor = センサ, センサー, 感知器, 感知部, 感応素子, 検出変換器, 変換素子, 受感部, 感覚器, センサー
- system = 組織体制, 制度, 子系, 系列, システム, 体系, 方式, 系統, 秩序, 体制, 組織, 一方式
- magnetic sensor = 磁気センサ
- sensor system = センサシステム, センサ系, センサーシステム

In such a situation, where the dictionary does not provide a translation for the complete multi-word string, we set the preference for the dictionary entry with the longest character length in the direction from left to right and select “magnetic sensor” + “system” entries for forced translation. Finally, we investigate the effects of using the bilingual dictionary both passively and pervasively by appending the dictionary before training and hijacking the decoding by forcing translations using the same dictionary.

5.1.1 Adding Single Word vs Multi-Word Dictionary Entries to SMT

Since the phrase-based machine translation model capitalizes on the alignment ‘consistency’ which essentially requires asymmetric alignment points to extract phrases, it remains unclear whether mono-lexical entries affect the final machine translation quality and how they affect the probability mass during the phrase extraction process. We examine the effects of adding only single word entries vs only adding multi-word entries from the dictionary to the SMT training data to compare the difference between using the full bilingual dictionary vs the mono-lexical or multi-word versions.

5.1.2 Adding a Manually Crafted Dictionary vs an Automatically Extracted Terminology to SMT

From the literature,[1] the popular approach to adding lexical information to statistical machine translation is to passively add automatically extracted a bilingual dictionary / terminology resources to SMT training data. The intrinsic quality of the automatically extracted bilingual dictionary is often disregarded since the ultimate interest is to achieve a BLEU score increment. In the Chapter 3, we experimentally show that both the intrinsic and extrinsic evaluation of the automatically extracted terminology resource using the PMI_{LM} term extraction statistics.

[1] See Chapter 2 Section 2.3

In the following result section, we would like to explore the difference between using an automatically extracted terminology and a manually crafted dictionary. We use the global PMI_{LM} extraction method described in Chapter 3 to extract the top 100,000 terms to be used as our automatically extracted terminology.

5.1.3 Experimental Setup

We experimented with the passive and pervasive uses of dictionary resources in SMT using the Japanese-English dataset provided in the Workshop for Asian Translation (Toshiaki et al. 2014). We used the Asian Scientific Paper Excerpt Corpus (ASPEC) as the training corpus used in the experiments. The ASPEC corpus consists of 3 million parallel sentences extracted from Japanese-English scientific abstracts from Japan's Largest Electronic Journal Platform for Academic Societies (J-STAGE). In our experiments we follow the setup of the WAT shared task with 18,00 development and test sentences each from the ASPEC corpus.

We use the manually crafted Japanese-English (JA-EN) translation dictionaries (JICST, 2004) from the Japan Science and Technology Corporation. It contains 800,000 entries for technical terms manually extracted from scientific and technological documents.

Similarly, we use the bilingual PMI_{LM} term extractor (from Chapter 3) to extract the top 800,000 terms from the training corpus. We seek to compare the efficacy of the manually crafted dictionary versus the automatically extracted one when using incorporating lexical information in the training data passively and pervasively.

From Chapter 4.5, in Table 4.5, we note that adding the top 10,000 terms extracted using the PM_{LM} reports statistically significant improvements from the baseline. While the experiment in Chapter 4.5 was concerned about how many unique terms to add to the training data for SMT, the experiments in this chapter using the automatically extracted terms focus on the number of times the extracted terminology is added to the training data for SMT.

For parity in comparing automatically extracted and manually crafted dictionary resources, we compute the PMI_{LM} values for all entries and extracted the 10,000 terms from the manually crafted dictionary.

The parallel data, the manually crafted bilingual dictionary and automatically extracted terminology are all tokenized with the MeCab segmenter (Kudo et al. 2004). We use the phrased-based machine translation configuration as described in Chapter 2, Section 2.4.4.

For the *passive* use of the dictionary, we simply appended the dictionary resources to the training data before the alignment and training process. For the *pervasive* use of the dictionary, we used the `xml-input` function in the Moses toolkit to force lexical knowledge in the decoding process.

Different from the normal use of a dictionary for the purpose of domain adaptation where normally, a domain-specific lexicon is appended to a translation model trained on generic texts, we are investigating the use of an in-domain dictionary in statistical machine translation.

More specifically, we seek to understand how much improvement can be made by skewing the lexical information towards the passive and pervasive use of the dictionary without additional domain knowledge i.e the dictionary comes from the same domain as the training corpus.

5.1.4 Results (Passive vs Pervasive)

	-Pervasive	**+Pervasive**
Baseline	16.75	16.87
Passive x 1	16.83	17.30**
Passive x 2	**17.31****	16.87
Passive x 3	17.26*	17.06
Passive x 4	17.14*	**17.38****
Passive x 5	16.82	17.29**

Table 5.1: BLEU Scores for Passive and Pervasive Use of the Dictionary in SMT (Japanese to English)

Table 4.1 presents the BLEU scores of the Japanese to English (JA-EN) translation outputs from the phrase-based SMT system on the WAT test set. The **-Pervasive** column indicates the number of times a dictionary is appended to the parallel training data (Baseline = 0 times, Passive xn = n time). The **+Passive** column presents the results from both the passive and pervasive use of dictionary translations, with the exception to the top-right cell which shows the baseline result of the pervasive dictionary usage without appending any dictionary to the training data.

By repeatedly appending the dictionary to the parallel data, the BLEU scores significantly[2] improves over the baseline from 16.75 to 17.31. Although the system's performance degrades when adding the dictionary passively thrice, the score remains significantly better than baseline. The pervasive use of the dictionary marginally improves the baseline without the passive use of the dictionary. The best performance is achieved when the dictionary is passively added four times with the pervasive use of the dictionary during decoding.

[2] *: p-value<0.1, **: p-value<0.001

The small fluctuations in improvement from coupling the passive and pervasive use of an in-domain dictionary give no indication of how both approaches should be used in tandem. However, using either or both the approaches improves the translation quality over the baseline system.

	-Pervasive	**+Pervasive**
Baseline	23.91	23.14**
Passive x 1	24.12*	23.13**
Passive x 2	23.79	22.86**
Passive x 3	**24.14***	**23.29****
Passive x 4	24.13*	23.16**
Passive x 5	23.67	22.71**

Table 5.2: BLEU Scores for Passive and Pervasive Use of the Dictionary in SMT (English to Japanese)

Table 4.2 presents the BLEU scores of the English to Japanese (EN-JA) translation outputs from the phrase-based SMT system on the WAT test set. The passive use of dictionary outperforms the baseline. Different from the JA-EN translation the pervasive use of dictionary consistently performs worse than the baseline. Upon random manual checking of the MT output, there are many instances where the technical/scientific term in the dictionary is translated correctly with only the passive use of the dictionary. However, it is unclear whether the overall quality of the translations have degraded from the pervasive use of the dictionary given the slight, though significant,[3] decrease in BLEU scores.

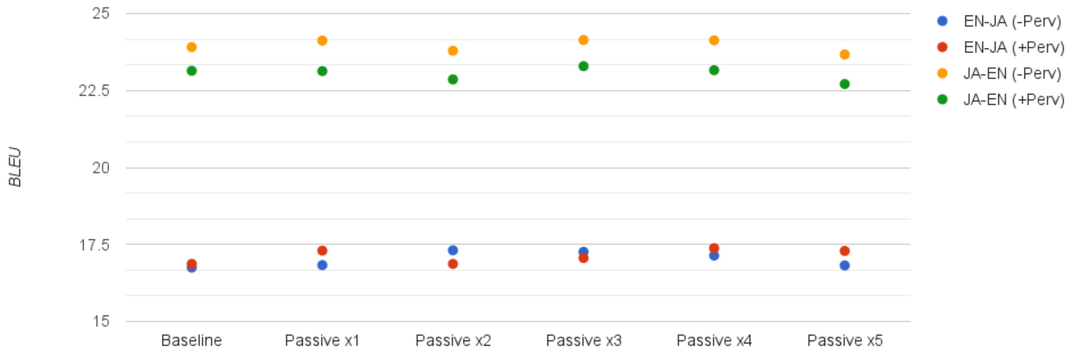

Figure 5.1: Overview of the Passive and Pervasive Use of Dictionary in the WAT Experiments

Figure 4.1 summarizes the results of our experiments on passive and pervasive use of manually created JIST dictionary with the ASPEC corpus evaluated with the WAT evaluation test set. Empirically, both passive and

[3] *: p-value<0.1, **: p-value<0.001

pervasive use of an in-domain dictionary to extend statistical machine translation models with lexical knowledge modestly improve translation quality. Interestingly, the fact that adding the in-domain dictionary information multiple times to the training data improves MT suggests that there may be a critical probability mass by increasing the frequency of in-domain terms which can impact the word and phrasal alignments in a corpus. This may provide insight on optimizing the weights of the salient in-domain phrases in the phrase table.

Although the pervasive use of dictionary information provides minimal or no improvements to the BLEU scores in our experiments, it remains relevant in industrial machine translation where terminological standardization is crucial in ensuring consistent translations of technical manuals or legal texts where incorrect use of terminology may have legal consequences (Porsiel, 2011).

5.1.5 Results (Mono-lexeme vs Multi-Word Expressions)

In a further experiment motivated in Section 4.1.1, we seek to examine the difference between adding mono-lexical entries and multi-word expressions. Table 4.3 presents the results of the experiments on passively adding subsets of the dictionary instead of the full JIST dictionary. The first column (**Mono-Lexeme**) of Table 4.3 indicates the number of times the mono-lexical entries from the dictionary is appended to the parallel training data (Baseline = 0 times, Passive x1 = 1 time, etc.). The **Multi-words** column presents the results from adding the multi-word entries from the dictionary to the data before the statistical machine translation training process. The last column of results (**Full**) in Table 4.3 is the same as the first results column in Table 4.1.

	Mono-Lexeme	**Multi-Words**	**Full**
Baseline	16.75	16.75	16.75
Passive x1	16.82*	16.99**	16.83
Passive x2	16.81*	16.59	**17.31****
Passive x3	**17.01**	**17.03**	17.26*
Passive x4	16.67	17.01	17.14*
Passive x5	16.54	16.96	16.82

Table 5.3: BLEU Scores in adding Mono-lexeme vs MWE Dictionary to SMT (English to Japanese)

Table 4.3 shows that it is possible to achieve slight statistically significant though marginal improvements from baseline (16.75) by adding the single word once or twice. Adding multi-word entries to the training data also makes statistically significant and marginal improves only when added once. Extending the training data with the full manually crafted dictionary makes the lexical addition more robust and the SMT system makes statistically significant improvements when the dictionary is added 2 to 4 times; the best score (17.31) is achieved when the full dictionary is added twice.

Here we identify a gap in the current strain of research focusing on passively adding only multi-word expressions to SMT training process. Intuitively, we may conjecture that that having mono-lexical entries (e.g. *magnetic* = 磁気 *sensor* = センサ*; system* = シス) that make up the partial multi-word expressions (*magnetic sensor* = 磁気センサ*; sensor system* = センサシステム) help in word/phrase alignment process, improving the overall translation quality. But this intuition seems to only hold when translating from English to Japanese.

	Single Word	**Multi-Words**	**Full**
Baseline	23.91	23.91	23.91
Passive x1	**23.93**	23.96	24.12*
Passive x2	23.87	24.01	23.79
Passive x3	23.54	**24.14****	**24.14***
Passive x4	22.91^{**-}	23.92	24.13*
Passive x5	23.35	23.87^{**-}	23.67

Table 5.4: BLEU Scores in adding Single Words vs Multi-Words Dictionary to SMT (Japanese to English)

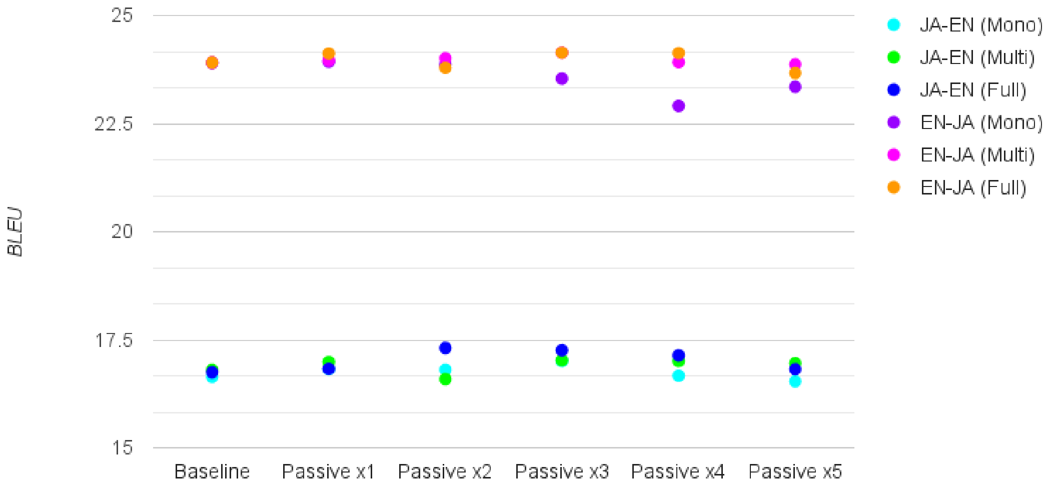

Figure 5.2: Overview of the Mono-Lexeme and Mult-Word Expression subset of the Dictionary in the WAT Experiments

Table 4.4 compares the results of adding the mono-lexical vs multi-word entries from the dictionary in the attempt to improve BLEU scores[4][5]. Interestingly, when translating from Japanese to English, the mono-lexical entries marginally reported lower BLEU scores as compared to the MWEs. And by adding only the multi-word entries, the SMT system achieves similar results to using the full dictionary.

Figure 4.2 summarizes our experiments comparing the addition of mono-lexical and multi-word entries to the statistical machine translation training process. Empirically, we have shown that the usage of the full dictionary might not be necessary depending on the directionality of translation.

5.1.6 Results (Manual vs Automatic)

	JICST (800k)	**PMI_LM (800k)**	**JICST (10k)**	**PMI_LM (10k)**
Baseline	16.75	16.75	16.75	16.75
Passive x1	16.83	15.34^{*-}	16.16	**16.81**
Passive x2	**17.31****	14.82^{**-}	16.23	16.73
Passive x3	17.26*	**15.90**$^{*-}$	**16.28**	15.68^{**-}
Passive x4	17.14*	14.32^{**-}	15.76^{*-}	15.54^{**-}
Passive x5	16.82	13.87^{**-}	15.12^{*-}	14.91^{**-}

Table 5.5: BLEU Scores in passively adding a Manually Crafted Dictionary vs an Automatically Extracted Terminology to SMT (English to Japanese)

Table 4.5 shows the result of passively adding a manually crafted dictionary (JICST) against an automatically extracted terminology using PMI_{LM} when translating from English to Japanese. For the comparison that follows, we will keep in mind that our baseline system without adding additional lexical knowledge scores 16.75 BLEU (from Table 4.1) and the significance results[6][7] presented on Table 4.5 are with respect to this baseline system.

The **JICST (800K)** and PMI_{LM} **(800K)** columns in Table 4.5 shows the results we presented on Table 4.1 when passively adding the full hand-crafted JICST dictionary to the training data and the second column shows the BLEU scores achieved by the same SMT configurations except that we passively add the top 800,000

[4] *: p-value<0.1, **: p-value<0.001

[5] *-: p-value<0.1 with negative BLEU from baseline, **-: p-value<0.001 with negative BLEU from baseline

[6] *: p-value<0.1, **: p-value<0.001

[7] *-: p-value<0.1 with negative BLEU from baseline, **-: p-value<0.001 with negative BLEU from baseline

terms automatically extracted terminology using PMI_{LM}. Table 4.5 that all results from passively adding the automatically extracted terms performed significantly worse than the baseline system that scored 16.75 BLEU.

The **JICST (10K)** and PMI_{LM} **(10K)** columns in Table 4.5 present another setting where we added the top 10,000 entries from the manually crafted dictionary ranked by their PMI_{LM} values and an automatically extracted terminology with PMI_{LM} with 10,000 terms. We achieved the best results (16.81 BLEU) when passively adding the automatically extracted terms just once. Although the absolute BLEU score is marginally higher than our baseline (16.75 BLEU), the increment is not significant. Likewise using the smaller subset of the manually crafted dictionary underperforms as compared to using the full dictionary.

Statistical significance measures the difference of the output of a system from the output of the baseline system. While the negative results of adding automatically extracted terminology are statistically significant, the BLEU score fluctuations are marginal. The overall results did not give a clear signal as to whether adding an automatically extracted dictionary is beneficial or detrimental to the SMT system.

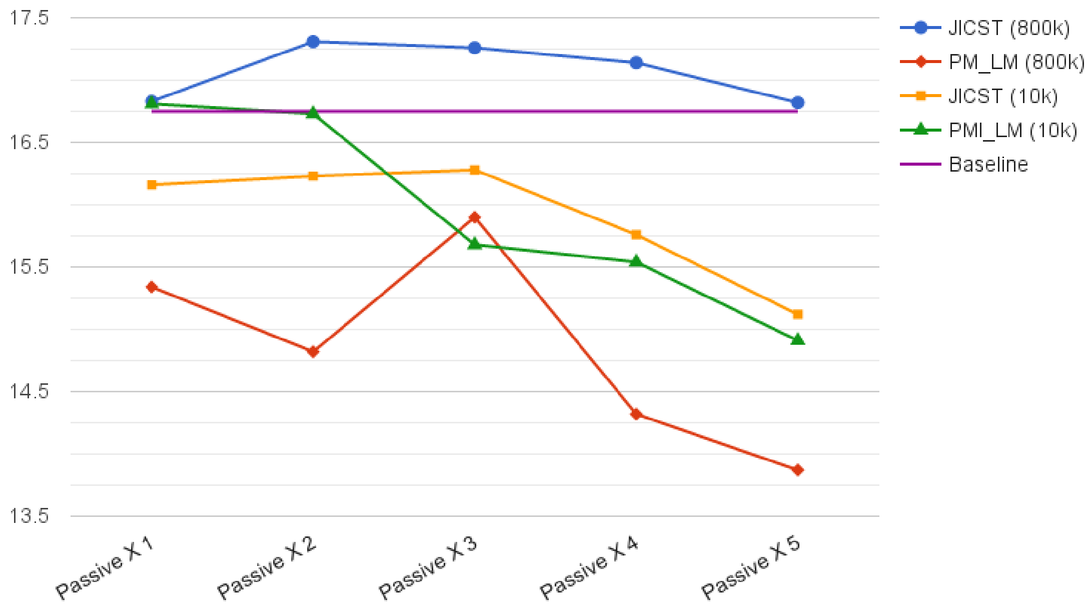

Figure 5.3: Passively Adding a Manually Crafted Dictionary vs an Automatically Extracted Terminology (English-Japanese)

Figure 4.3 provides a visual description for Table 4.5. From Figure 4.3, it is more evident that adding the additional lexical information once is optimal when using an automatically extracted terminology (referring to the first green-triangle data point). BLEU scores quickly degrades once we add the 10k terms more than two times.

Unsurprisingly, when we filter out the top 10k entries from the manually crafted dictionary, it underperformed as compared to using the full dictionary. However, we see a marginally upwards trend in BLEU scores as we add the dictionary once and twice. This suggests the *'effective multiplier'* idea that (Brown et al., 1993a) proposed[8].

The narrative changes when we increase the size of the terminology to 800k, passively adding the automatically extracted terminology thrice performs the best but other than this small anomaly, adding the terminology follows a linear downwards trend[9]. From previous experiments varying the size of the terminology (Chapter 3, Section 3.5), we see that using the top 10,000 terms performed better than 12,000 terms for both English-Japanese and Japanese-English translation[10]. Thus increasing the size of terminology to 800,000 would just be adding more noise to the system. Interestingly, we see that the phrase-based machine translation is rather robust to the noisy 70,000 bilingual terms that we've added and report only slight decrease in BLEU scores.

	JICST (800k)	**PMI_LM (800k)**	**JICST (10k)**	**PMI_LM (10k)**
Baseline	23.91	23.91	23.91	23.91
Passive x1	24.12^{*}	23.26^{*-}	**23.85**	**23.78**
Passive x2	23.79	**23.45**	23.72	23.76
Passive x3	**24.14***	22.81^{*-}	22.84^{*-}	22.93^{*-}
Passive x4	24.13^{*}	22.35^{*-}	23.04^{*-}	23.01^{*-}
Passive x5	23.67	21.93^{*-}	22.91^{*-}	23.00^{*-}

Table 5.6: BLEU Scores in adding Manually Crafted Dictionary vs Automatically Extracted Terminology to SMT (Japanese to English)

Table 4.6 shows the result of passively adding a manually crafted dictionary (JICST) against an automatically extracted terminology using PMI_{LM} when translating from Japanese to English. For the comparison that follows, we will bear in mind that our baseline system without adding additional lexical knowledge scores 23.91 BLEU (from Table 4.2) and the presented on Table 4.5 are respectively to this baseline system.
Figure 4.4 presents the results better graphically. similar to translating from English to Japanese, the best performance is achieved by passively adding manually crafted dictionary (BLEU 24.14), in this case adding it thrice. We see the same robustness of the SMT system given the noisy lexical information as the BLEU score drops marginally though the deterioration is statistically significant.

[8] See Chapter 2, Section 2.3.1

[9] Possibly, the fluke in BLEU when passively adding the terminology thrice is caused by the non-deterministic MERT tuning process that found better optimum weights.

[10] They are trained and evaluated on the same ASPEC corpus and WAT evaluation dataset

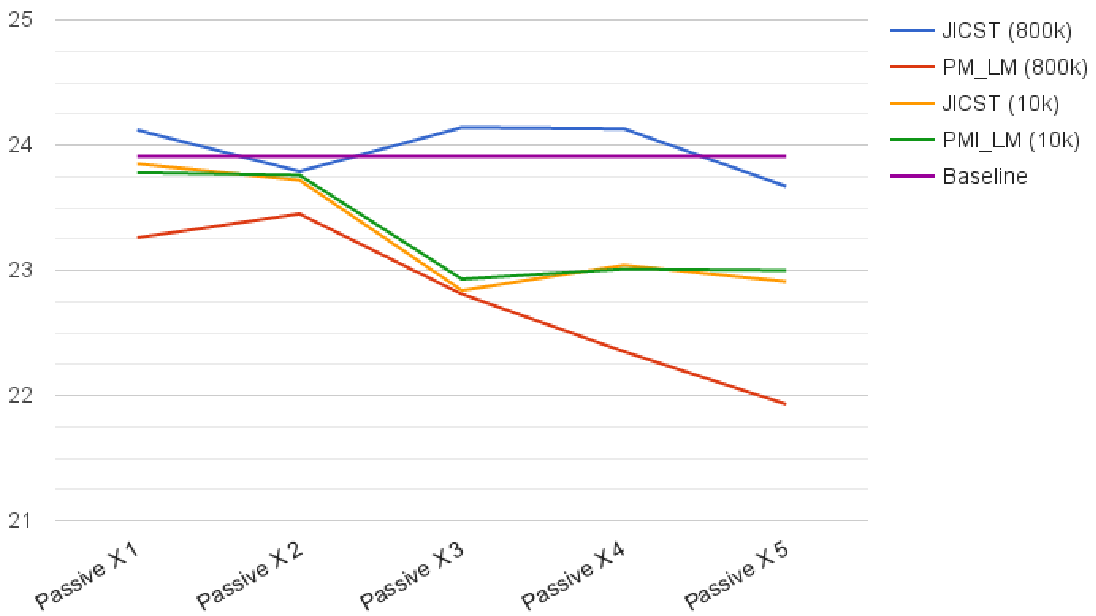

Figure 5.4: Passively Adding Manually Crafted Dictionary vs Automatically Extracted Terminology (Japanese-English)

In contrast to passively adding 800k automatically extracted terms when translation from EN-JA, we see a slight increase in BLEU 23.26 -> 23.45 when the 800k terms were added twice. However, the same linearly degenerate behavior happens as we continue to passively increase the effective multiplier.

As we reduce to the automatically extracted terminology size and the manually crafted dictionary size to 10k, we see that we achieved better results but still lower than the baseline (23.91 BLEU). But this does not come as a surprise since the best possible results from adding lexical information comes from adding the full JICST dictionary and it scantily improves upon the baseline system 23.91 -> 24.14.

5.2 Summary

In this chapter, we explore the multiple facade to integrate lexical information to phrase-based statistical machine translation from (i) comparing passive and pervasive techniques, to understanding (ii) the synergy in using both mono-lexical and multi-word entries in passively adding lexicon to the SMT process and (iii) the difference in adding manually crafted versus automatically extracted lexicon to the SMT training data.

In general, passively adding additional lexical information to statistical machine translation performs better than pervasively forcing the machine translation decoder to use the bilingual lexicon entries. To achieve optimal 'effective multiplier' effect in passively addition lexicon to machine translation, we found that adding the lexicon more than once improves the performance; in our experiments, we found that adding 3-4 times yields the best result. Additionally, adding manually crafted dictionaries outperforms automatically extracted dictionaries.

Chapter 6

Measuring the Goodness of Machine Translation

The discussion up till now on evaluating translation quality is based on BLEU scores (Papineni et al., 2002) and its statistical significance based on bootstrap resampling (Koehn, 2004). In this Chapter, we propose a different way to look at machine translation evaluation focusing on (i) evaluating the semantic ‘goodness’ of machine translated texts by casting it as a semantic textual similarity task and (ii) highlighting an awkward disparity between BLEU / RIBES[1] and human evaluation of translated text.

[1] An extension of BLEU

6.1 Evaluating Semantic ‘Goodness’ of Machine Translated Texts

Translation is becoming an utility in everyday life. The increased availability of real-time machine translation services relying on Statistical Machine Translation (SMT) allows users who do not understand the language of the source text to quickly gist text in a foreign language and understand its general meaning. For these users, often the accurate meaning of translated words is more important than the fluency of the translated sentence.

However, SMT can suffer from poor lexical choices. Fluent but inadequate translations are commonly produced due to the strong bias towards the language model component that prefers consecutive words based on the data that the system is trained on.

Current state of art MT evaluation metrics are generally able to identify problems with the grammaticality of the translation but less evidently the accuracy of translated semantics, e.g. incorrect translation of ambiguous words or wrong assignment of semantic roles. In the example below, the ideal Machine Translation (MT) evaluation metric should appropriately penalise poor lexical choice, such as ‘*braked*’, and reward or at least allow leeway for semantically similar translations, such as ‘*external trade*’.

Source (DE):
Auch der Aussenhandel bremste die Konjunktur.
Phrase-based MT:
The foreign trade braked the economy.
Neural MT:
External trade also slowed the economy.
Reference (EN):
Foreign goods trade had slowed, too.

The German word *bremste* is commonly used as ‘*braked*’ in the context of driving, but the appropriate translation should have been ‘*slowed*’ in the example mentioned above. Although the phrase *external trade* differs from f*oreign goods trade* in the reference sentence, it should be considered as an acceptable translation.

To pursue a semantically motivated measure of goodness for machine translation, we pursue an evaluation metric that takes into account both fluency (grammaticality) and adequacy (semantics) to evaluate whether the machine translation output has the same meaning as its reference translation through the Semantic Textual Similarity (STS) task. Semantic Textual Similarity (STS) is the task of measuring the degree to which two texts have the same meaning (Agirre et al., 2014). For instance, given the two texts, "*the man is slicing the tape from the box.*" and "*a man is cutting open a box.*", an STS system predicts a real number similarity score on a scale of 0 (no relation) to 5 (semantic equivalence). In this case we can relate the pair of texts closely to what we are evaluating when measuring the goodness of machine translation output if we treat one of the sentence as the output and the other as the reference.

We propose a model that approaches the task by (i) combining existing machine translation evaluation metrics (that are good in determining fluency) and (ii) using linguistically motivated monolingual word alignments and neural embeddings (to add the semantic dimension to our new metric). We call our metric, `Stasis`.[2]

6.1.1 Measuing Grammatical Fluency with MT Evaluation Metrics Ensemble

Previous approaches have applied MT evaluation metrics for the STS task with progressively improving results (Agirre et al., 2012, 2013, 2014, 2015). We propose a single MT evaluation metric by ensembling a glut of MT evaluation metrics.[3]

Previous Usage of MT Metrics to Determine Textual Similarity

At the pilot English STS-2012 task, Rios et al. (2012) trained a Support Vector Regressor using the lexical overlaps between the surface strings, named entities and semantic role labels and the BLEU (Papineni et al., 2002a) and METEOR (Banerjee & Lavie, 2005a; Denkowski & Lavie, 2010) scores between the text snippets and their best system scored a Pearson correlation mean of 0.3825. The system underperformed compared to the organizers' baseline system[5] which scored 0.4356.

For the English STS-2013 task, Barrón-Cedeño et al. (2013) also used a Support Vector Regressor with an larger array of machine translation metrics (BLEU, METEOR, ROUGE (Lin & Och, 2004a), NIST (Doddington, 2002), TER (Snover et al., 2006)) with measures that compute similarities of dependency and constituency parses (Liu & Gildea, 2005) and semantic roles, discourse representation and explicit semantic analysis (Gabrilovich & Markovitch, 2007) annotations of the text snippets. These similarity measures are packaged in the Asiya toolkit (Giménez & Màrquez, 2010). They scored 0.4037 mean score and performed better than the Takelab baseline (Šarić et al., 2012) at 0.3639.

[2] Part of the research in this chapter has been published in Tan et al. (2015c, 2016)

[3] This section presents a collaborative work between Saarland University and University of Sheffield[4] in developing the MT evaluation metric ensemble.

[5] Refers to the token cosine baseline system (`baseline-tokencos`) in STS-2012.

At the SemEval-2014 Cross-level Semantic Similarity task (Jurgens et al., 2014, 2015), participating teams submitted similarity scores for text of different granularity. Huang & Chang (2014) used a linear regressor solely with MT evaluation metrics (BLEU, METEOR, ROUGE) to compute the similarity scores between paragraphs and sentences. They scored 0.792 beating the lowest common substring baseline which scored 0.613.

In the SemEval-2015 English STS and Twitter similarity tasks, (Bertero & Fung, 2015) trained a neural network classifier using (i) lexical similarity features based on WordNet (Miller, 1995a), (ii) neural auto-encoders (Socher et al., 2011), syntactic features based on parse tree edit distance (Zhang & Shasha, 1989; Wan et al., 2006) and (iii) MT evaluation metrics, viz. BLEU, TER, SEPIA (Habash & Elkholy, 2008), BADGER (Parker, 2008) and MEANT (Lo et al., 2012).

For the classic English STS task in SemEval-2015, Tan et al. (2015c) used a range of MT evaluation metrics based on lexical (surface n-gram overlaps), syntactic (shallow parsing similarity) and semantic features (METEOR variants) to train a Bayesian ridge regressor. Their best system achieved 0.7275 mean Pearson correlation outperforming the `token-cos` baseline which scored 0.5871 while the top system (Sultan et al., 2015b) achieved 0.8015.

Another notable mention of MT technology in the STS tasks is the use of referential translation machines to predict and derive features instead of using MT evaluation metrics (Biçici & van Genabith, 2013; Biçici & Way, 2014; Bicici, 2015).

Feature Matrix

Following the success of STS systems that use MT evaluation metrics, we train three regression models using an array of MT metrics based on lexical, syntactic and semantic features.

Machine translation evaluation metrics utilize various degrees of lexical, syntactic and semantic information. Each metric considers several features that compute the translation quality by comparing a translation against one or several reference translations.

We trained our system using the follow feature sets: (i) n-gram, shallow parsing and named entity overlaps (`Asiya`), (ii) `BEER`, (iii) `METEOR` and (iv) `ReVal`. Although `BEER` and `METEOR` metrics provided mecahnisms to fine-tuned the metric with respect to the training data, we did not tune these metrics because they will be used as input features that will be fed into an ensemble which will automatically learn their feature weights which effectively fine-tuned the metric to the training data too.

`Asiya` **Features**: Gonzàlez et al. (2014) introduced a range of language independent metrics relying on n-gram overlaps similar to the modified $n-$-gram precisions of the BLEU metric (Papineni et al., 2002a). Different from BLEU, Gonzàlez et al. (2014) computes n-gram overlaps using similarity coefficients instead of proportions. We use the `Asiya` toolkit (Giménez & Màrquez, 2010) to annotate the dataset with the similarity coefficients of n-gram overlap features described in this section. We use 16 features from both cosine similarity and Jaccard Index coefficients of the character-level and token-level n-grams from the order of bigrams to 5-grams. Additionally, we use the Jaccard similarity of the pseudo-cognates and the ratio of n-gram length as the 17th and 18th features. Adding a syntactic dimension to our feature set, we use 52 shallow parsing features described in Tan et al. (2015c); they measure the similarity coefficients from the n-gram overlaps of the lexicalized shallow parsing (aka chunking) annotations. As for semantics, we use 44 similarity coefficients from Named Entity (NE) annotation overlaps between two texts. After some feature analysis, we found that 22 out of the 44 NE n-gram overlap features and 1 of the shallow parsing features have extremely low variance across all sentence pairs in the training data. We removed these features before training our models.

`BEER` **Features**: Stanojevic & Sima'an (2014) presents an MT evaluation metric that uses character n-gram overlaps, the Kendall tau distance of the monotonic word order (Isozaki et al., 2010; Birch & Osborne, 2010) and abstract ordering patterns from tree factorization of permutations (Zhang & Gildea, 2007). While Asiya features are agnostic to word classes, BEER differentiates between function words and non-function words when calculating its adequacy features.

`METEOR` **Features**: METEOR first aligns the translation to its reference, then it uses the unigram mapping to see whether they match based on their surface forms, word stems, synonyms and paraphrases (Banerjee & Lavie, 2005a; Denkowski & Lavie, 2010). Similar to BEER features, METEOR makes a distinction between content words and function words and its recall mechanism weights them differently. We use all four variants of METEOR: exact, stem, synonym and paraphrase.

`ReVal` **Features**: ReVal (Gupta et al., 2015a) is a deep neural net based metric which uses the cosine similarity score between the Tree-based Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997; Cho et al., 2014; Tai et al., 2015a) dense vector space representations of two sentences.

The MT Metric Ensemble

The STS 2012 to 2015 datasets are made up of sentence pairs with manually annotated scores of the similarity between each pair of sentences. We annotated the STS 2012 to 2015 datasets with the features as described in the previous section and trained three models using (i) a linear regressor (`Linear`), (ii) boosted tree regressor (`Boosted`) (Friedman, 2001) and (iii) eXtreme Gradient Boosted tree regressor (`XGBoost`) (Chen & He, 2015; Chen & Guestrin, 2015).

6.1.2 Measuing Semantic Adequacy With Monolingual Word Alignments and Neural Network Embeddings

For the 2014 and 2015 editions of the STS task, the top performing submissions are from the DLS@CU team (Sultan et al., 2014b, 2015a).

Their STS2014 submission is based on the proportion of overlapping content words between the two sentences treating semantic similarity as a monotonically increasing function of the degree to which two sentences contain semantically similar units and these units occur in similar semantic contexts (Sultan et al., 2014b). Essentially, their semantic metric is based on the proportion of aligned content words between two sentences, formally defined as:

$$prop_{Al}^{(1)} = \frac{|\{i : [\exists j : (i,j) \in Al] \; and \; w_i^{(1)} \in C\}|}{|\{i : w_i^{(1)} \in C\}|} \tag{6.1}$$

where $prop_{Al}^{(1)}$ is the monotonic proportion of the semantic unit alignment from a set of alignments Al that maps the positions of the words (i, j) between sentences $S^{(1)}$ and $S^{(2)}$, given that the aligned units belong to a set of content words, C. Since the proportion is monotonic, the equation above only provides the proportion of semantic unit alignments for $S^{(1)}$. The Al alignments pairs are automatically annotated by a monolingual word aligner (Sultan et al., 2014a) that uses word similarity measures based on contextual evidence from the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) and syntactic dependencies.

The same computation needs to be made for $S^{(2)}$. An easier formulation of Equation (6.1) without the formal logic symbols is:

$$prop_{Al}^{(1)} = \frac{sum(1 \; for \; w_i, w_j \; in \; Al^{(1,2)} \; if \; w_i \; in \; C)}{sum(1 \; for \; w_i \; in \; S^{(1)} \; if \; w_i \; in \; C)} \tag{6.2}$$

Since the semantic similarity between $(S^{(1)}, S^{(2)})$ should be a single real number, Sultan et al. (2014b) combined the proportions using harmonic mean:

$$sim(S^{(1)}, S^{(2)}) = \frac{2 * prop_{Al}^{(1)} * prop_{Al}^{(2)}}{prop_{Al}^{(1)} + prop_{Al}^{(2)}} \tag{6.3}$$

Instead of simply using the alignment proportions, Sultan et al. (2015a) extended their hypothesis by leveraging pre-trained neural net embeddings (Baroni et al., 2014). Sultan et al. (2015a) posited that the semantics of the sentence can be captured by the centroid of its content words[6] computed by the element-wise sum of the content word embeddings normalized by the number of content words in the sentence. Together with the similarity scores from Equation (6.3) and the cosine similarity between two sentence embeddings, Sultan et al. (2015a) trained a Bayesian ridge regressor to learn the similarity scores between text snippets.

Replicating of DLS

To replicate the success of Sultan et al. (2014b), we use the monolingual word aligner from Sultan et al. (2014a) to annotate the STS-2012 to STS-2015 datasets and computed the alignment proportions as in Equation (6.1) and (6.2).

In duplicating Sultan et al. (2015a) work, we first have to tokenize and lemmatize text. The details of preprocessing choices was undocumented in their paper, thus we lemmatized the datasets with the NLTK tokenizer (Bird et al., 2009) and PyWSD lemmatizer (Tan, 2014). We use the lemmas to retrieve the word embeddings from the COMPOSES vector space (Baroni et al., 2014). Similar to Equation (6.2) (changing only the numerator), we sum the sentence embedding's centroid as follows:

$$v(S^{(1)}) = \frac{sum(v(w_i)\ for\ w_i\ in\ S^{(1)}\ if\ w_i\ in\ C)}{sum(1\ for\ w_i\ in\ S^{(1)}\ if\ w_i\ in\ C)} \tag{6.4}$$

where $v(S^{(1)})$ refers to the dense vector space representation of the sentence $S^{(1)}$ and $v(w_i)$ refers to the word embedding of word i provided by the COMPOSES vector space. The same computation has to be done for $S^{(2)}$.

Intuitively, if either of the sentences contains more or less content words than the other, we can see the numerator changing but the denominator changes with it. The difference between $v(S^{(1)})$ and $v(S^{(2)})$ contributes to *distributional semantic distance*.

[6] In the implementation, they have used lemmas instead of words to reduce sparsity when looking up the pre-trained embeddings (personal communication with Arafat Sultan).

To calculate a real value similarity score between the sentence vectors, we take the dot product between the vectors to compute the cosine similarity between the sentence vectors:

$$sim(S^{(1)}, S^{(2)}) = \frac{v(S^{(1)}) \cdot v(S^{(2)})}{|v(S^{(1)})|\,|v(S^{(2)})|} \tag{6.5}$$

There was no clear indication of which vector space Sultan et al. (2015a) have chosen to compute the similarity score from Equation 5.5. Thus we compute two similarity scores using both COMPOSES vector spaces trained with these configurations:

- 5-word context window, 10 negative samples, subsampling, 400 dimensions
- 2-word context window, PMI weighting, no compression, 300K dimensions

In this way, we extracted two similarity features for every sentence pair. With the harmonic proportion feature from Equation 5.3 and the similarity scores from Equation 5.5, we trained a boosted tree ensemble on the 3 features using the STS 2012 to 2015 datasets and submitted the outputs from this model as our baseline submission in the English STS Task in SemEval 2016.

Replacing COMPOSES with GloVe

Pennington et al. (2014a) handles semantic regularities (Levy et al., 2014) explicitly by using a global log-bilinear regression model which combines the global matrix factorization and the local context vectors when training word embeddings.

Instead of using the COMPOSES vector space, we experimented with replacing the $v(w_i)$ component in Equation 5.4 with the GloVe vectors,[7] $v_{glove}(w_i)$ such that:

$$sim_{glove}(S^{(1)}, S^{(2)}) = \frac{v_{glove}(S^{(1)}) \cdot v_{glove}(S^{(2)})}{|v_{glove}(S^{(1)})|\,|v_{glove}(S^{(2)})|} \tag{6.6}$$

The novelty lies in the usage of the global matrix to capture corpus wide phenomena that might not be captured by the local context window. The model leverages on both the non-zero elements in the word-word co-occurence matrix (not a sparse bag-of-words matrix) and the individual context window vectors similar to the word2vec model (Mikolov et al., 2013b).

[7] We use the 300 dimensions vectors from the GloVe model trained on the Commoncrawl Corpus with 840B tokens, 2.2M vocabulary.

Similarity Using Tree LSTM

Recurrent Neural Nets (RNNs) allow arbitrarily sized sentence lengths (Elman, 1990) but early work on RNNs suffered from the vanishing/exploding gradients problem (Bengio et al., 1994). Hochreiter & Schmidhuber (1997) introduced multiplicative input and output gate units to solve the vanishing gradients problem. While RNN and LSTM process sentences in a sequential manner, Tree-LSTM extends the LSTM architecture by processing the input sentence through a syntactic structure of the sentence. We use the ReVal metric (Gupta et al., 2015a) implementation of Tree-LSTM (Tai et al., 2015b) to generate the similarity score.

ReVal represents both sentences (h_1, h_2) using Tree-LSTMs and predicts a similarity score $\hat{y}$ based on a neural network which considers both distance and angle between h_1 and h_2:

$$\begin{aligned} h_\times &= h_1 \odot h_2 \\ h_+ &= |h_1 - h_2| \\ h_s &= \sigma\left(W^{(\times)} h_\times + W^{(+)} h_+ + b^{(h)}\right) \\ \hat{p}_\theta &= \text{softmax}\left(W^{(p)} h_s + b^{(p)}\right) \\ \hat{y} &= r^T \hat{p}_\theta \end{aligned} \tag{6.7}$$

where, σ is a sigmoid function, $\hat{p}_\theta$ is the estimated probability distribution vector and $r^T = [1\ 2 ... K]$. The cost function $J(\theta)$ is defined over probability distributions p and $\hat{p}_\theta$ using regularised Kullback–Leibler (KL) divergence.

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{KL}\left(p^{(i)} \middle|\middle| \hat{p}_\theta^{(i)}\right) + \frac{\lambda}{2} ||\theta||_2^2 \tag{6.8}$$

In Equation 6.8, i represents the index of each training pair, n is the number of training pairs and p is the sparse target distribution such that $y = r^T p$ is defined as follows:

$$p_j = \begin{cases} y - \lfloor y \rfloor, & j = \lfloor y \rfloor + 1 \\ \lfloor y \rfloor - y + 1, & j = \lfloor y \rfloor \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq j \leq K$, where, $y \in [1, K]$ is the similarity score of a training pair. This gives us a similarity score between [1, K] which is mapped to [0, 1].[8]

A Semantically Adequate Metric

Similar to our approach in the MT metric ensemble, we annotated the STS 2012 to 2015 datasets with the semantic similarity features as described in the previous section and trained three models using (i) a linear ridge regressor (`L`) that uses the similarity score from Equations 6.3 and 6.5 as features, (ii) extending the linear regression, we included the similarity score from Equations 6.6 and 6.8 to the feature set and trained a boosted tree ensemble (Friedman, 2001) (`B`) (iii) we use the same feature set as our B system to train an eXtreme Gradient Boosted tree regressor (`XGBoost`) (Chen & He, 2015; Chen & Guestrin, 2015), we refer to it as (`X`).

6.1.3 Experimental Setup and Results

We evaluated the two components of our `Stasis` metric using the STS2016 dataset that consist of text snippets from the following domains, (i) forum answers, (ii) forum questions, (iii) news headlines, (iv) plagiarism checking samples from student essays and (v) post-editing texts.

	answer-answer	headlines	plagiarism	postediting	question-question	Overall
`MT Ensemble (L)`	0.31539	0.76551	0.82063	0.83329	**0.73987**	0.68923
`MT Ensemble (B)`	0.37717	0.77183	0.81529	0.84528	0.66825	0.69259
`MT Ensemble (X)`	*0.47716*	**0.78848**	**0.83212**	*0.84960*	0.69815	*0.72693*
`Semantic (L)`	0.48799	0.71043	0.80605	0.84601	0.61515	0.69244
`Semantic (B)`	0.49415	0.71439	*0.79655*	0.83758	*0.63509*	0.69453
`Semantic (X)`	*0.49947*	*0.72410*	0.79076	*0.84093*	0.62055	*0.69471*
`Stasis (X)`	**0.50628**	0.77824	0.82501	**0.84861**	0.70424	**0.73050**
Median	0.48018	0.76439	0.78949	0.81241	0.57140	0.68923
Best	0.69235	0.82749	0.84138	0.86690	0.74705	0.77807

Table 6.1: `Stasis` Metric Evaluated on the STS-2016 Task; *Best* is the best performing results per domain from various systems participating in the English STS-2016 shared task and *Median* is the median scores from all participating systems in the English STS-2016 shared task.

Table 6.1 presents the results for evaluating our `Stasis` metric on the English STS-2016 dataset. The top part

[8] score = (score-1)/K Please refer to Gupta et al. (2015a) for training details.

of the table presents the results of the component that ensembles the MT evaluations metrics and the second portion of the table presents the results of the semantic adequacy component that uses linguistically motivated monolingual word alignment and neural network embeddings. The bottom-most part of the table presents the median and the best correlation results across all participating teams in the STS-2016 task.

Our MT metric ensemble system, our baseline linear model outperforms the median scores for all domains except the *answer-answer* domain. Our boosted tree model performs better than the linear model and the extreme gradient boosted tree model performs the best of the three. We note that our correlation scores for all three models is lower than the median for the *answer-answer* domain.

As for our semantic adequacy component, the median and best scores are computed across all participating teams in the task. Our baseline system performs reasonably well, outperforming the median scores in most domains. Our extended variant of the baseline using boosted tree ensemble performs better in the answer-answer, headlines and postediting domains but performed worse in others. Comparatively, it improves the overall correlation score marginally by 0.002. The system using XGBoost performs the best of the 3 models but it underperforms in the headlines and plagiarism domain when compared to the median scores.

When the MT evaluation ensemble and the semantic adequacy is combined using the XGBoost regressor, we achieved a higher score for every domain leading to an overall Pearson correlation score of 0.73050 (`Stasis (X)` in Table 6.1).

Inadequacy in Current MT Evaluation Metrics

In this section, we look closer at the inadequacy of the MT metric ensemble system. Figure 6.1 shows the bubble chart of the L1 error analysis of our XGBoost model against the gold standard similarity scores for the answer-answer domain. The colored lines correspond to the integer annotations, e.g. the yellow line represents the data points where the gold-standard annotations are 1.0. The span of the line represents the span of predictions our model made for these texts. The size of the bubble represents the effect size of our predictions' contribution to the Pearson correlation score, i.e. how close our predictions are to the gold standards.

As we see from Figure 6.1, the centroids of the bubbles represents our model's best predictions. Our predictions for texts that are annotated at 1 to 4 similarity scores are reasonably close to the gold standards but the model performs poorly for texts annotated with the 0 and 5 similarity scores.

Looking at the texts that are rated 0, we see that there are cases where the n-grams within these texts are

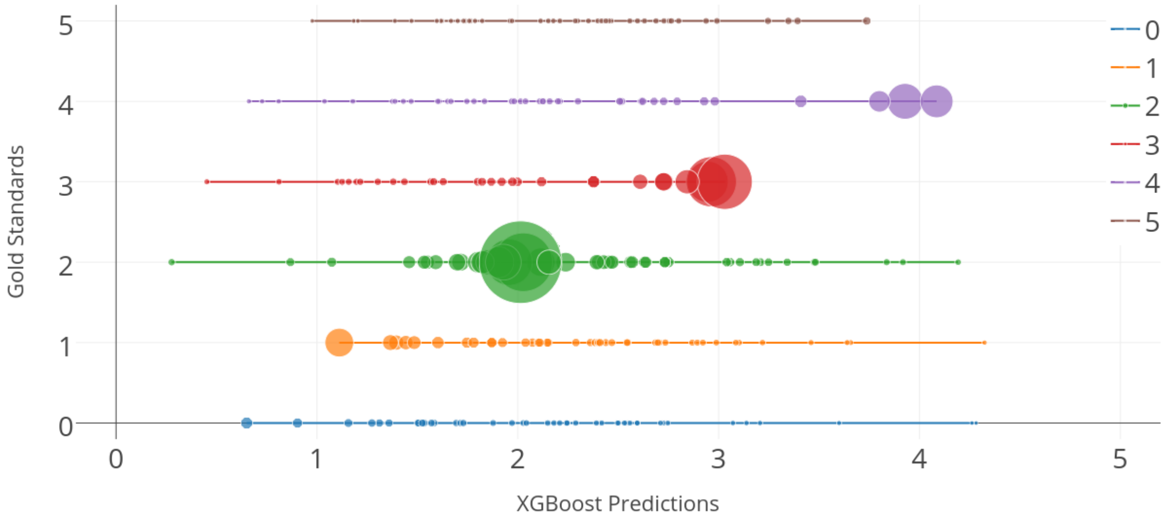

Figure 6.1: L1 Error Analysis on the answer-answer domain

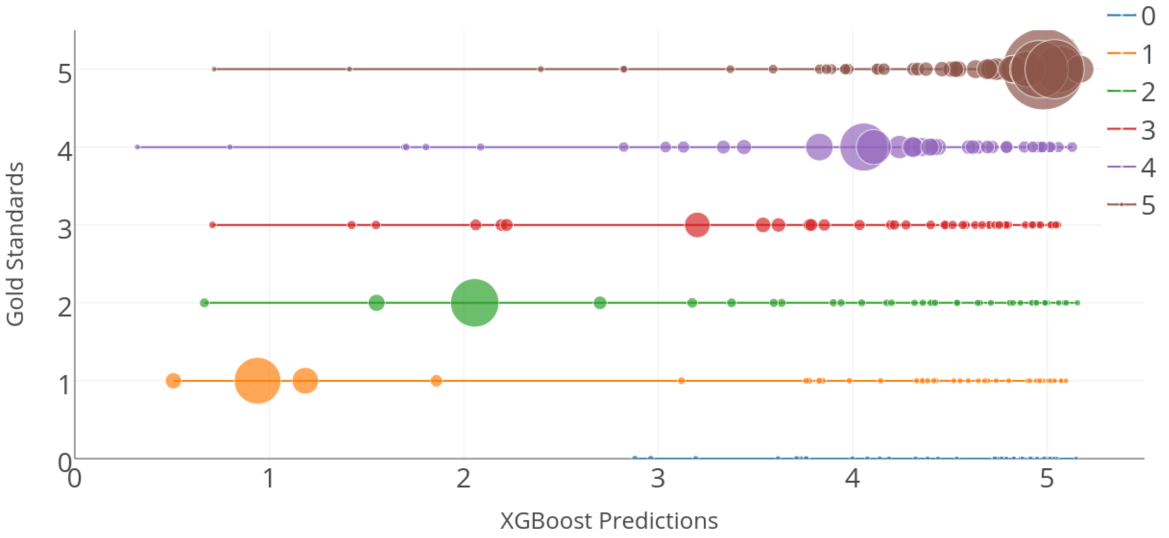

Figure 6.2: L1 Error Analysis on the post-editing domain

lexically / syntactically similar but the meaning of the texts are disparate. For example, this pair of text snippets, *'You don't have to know'* and *'You don't have equipments/facilities'* are rated 0 in the gold standards but from a machine translation perspective, a translator would have to do little work to change *'to know'* to *'equipments/facilities'*.

Because of this, machine translation metrics would rate the texts as being similar and even suitable for post-editing. However, the STS task focuses only on the meaning of the text which corresponds more to the adequacy aspect of the machine translation metrics. Semantic adequacy is often overlooked in machine translation because our mass reliance on BLEU scores to measure the goodness of translation with little considerations for penalizing semantic divergence between the translation and its reference.

On the other end of the spectrum, machine translation metrics remain skeptical when text snippets are annotated with a score of 5 for being semantically analogous but syntactically the texts are expressed in a different form. For example, given the text snippets, *'There's not a lot you can do about that'* and *'I'm afraid there's not really a lot you can do'*, most machine translation metrics will not allocate full similarity scores due to the difference in lexical and stylistic ways in which the sentences are expressed.

Machine translation metrics' failure to capture similarity score extremes is evident in Figure 6.1 where there are no 0 and 5.0 predictions.

Naturally the XGBoost predictions based on the MT evaluation metric features fit the postediting domain. Figure 6.2 shows that the centroids in the L1 Figure for the postediting domain is more centered than in the answer-answer domain.

6.1.4 Summary

In this section, we pursue a combination machine translation evaluation metric `Stasis` to evaluate the grammaticality and semantic adequacy between pairs of sentences. We combined existing machine translation evaluation metrics (that are good in determining fluency) and used linguistically motivated monolingual word alignments and neural embeddings to add the semantic dimension to our new metric.

6.2 The Awkward Disparity between BLEU / RIBES Scores and Human Judgments

Automatic evaluation of machine translation (MT) quality is essential in developing high quality MT systems. The relatively consistent correlation of higher BLEU scores (Papineni et al., 2002b) and better human judgements in major machine translation shared tasks has led to the conventional wisdom that translations with significantly higher BLEU scores generally suggests better translation than its lower scoring counterparts (Bojar et al., 2014, 2015; Nakazawa et al., 2014; Cettolo et al., 2014).

However, automatic MT evaluation metrics have been criticized for a variety of reasons (Babych & Hartley, 2004; Callison-Burch et al., 2006). Callison-Burch et al. (2006) has anecdotally presented possible failures of BLEU by showing examples of translations with the same BLEU score but of different translation quality. Through meta-evaluation[9] of BLEU scores and human judgements scores of the 2005 NIST MT Evaluation exercise, they have also showed high correlations of $R^2 = 0.87$ (for adequacy) and $R^2 = 0.74$ (for fluency) when an outlier rule-based machine translation system with poor BLEU score and high human score is excluded; when included the correlations drops to 0.14 for adequacy and 0.74 for fluency.

Although Callison-Burch et al. (2006) showed poor correlation between BLEU and human scores, they had only empirically meta-evaluated a scenario where low BLEU score does not necessary result in a poor human judgement score.

In this section, we demonstrate a real-world example of machine translation that yielded high automatic evaluation scores but failed to obtain a good score on manual evaluation in an MT shared task submission. In addition to the BLEU metric, we also evaluated our experiments results using the RIBES metric which has previously shown to have better correlations with human judgements due to its sensitivity to reordering (Isozaki et al., 2010).

[9] Meta-evaluation refers to the measurement of the Pearson correlation R^2 between an automatic evaluation metrics and human judgment scores. The meta-evaluation involves the calculation using other correlation measures such as the Spearman's rank correlation ρ (Callison-Burch et al., 2007) or the Kendall's Tau τ (Stanojević et al., 2015; Graham et al., 2015)

6.2.1 BLEU

Papineni et al. (2002) originally define BLEU n-gram precision p_n by summing the n-gram matches for every hypothesis sentence S in the test corpus C:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)} \tag{6.9}$$

BLEU is a precision based metric; to emulate recall, the brevity penalty (BP) is introduced to compensate for the possibility of high precision translations that are too short. The BP is calculated as:

$$BP = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-r/c} & \text{if} \quad c \leq r \end{cases} \tag{6.10}$$

where c and r respectively refers to the length of the hypothesis translations and the reference translations. The resulting system BLEU score is calculated as follows:

$$\text{BLEU} = \text{BP} \times \exp(\sum_{n=1}^{N} w_n \log p_n) \tag{6.11}$$

where n refers to the orders of n-gram considered for p_n and w_n refers to the weights assigned for the n-gram precisions; in practice, the weights are uniformly distributed.

A BLEU score can range from 0 to 1 and the closer to 1 indicates that a hypothesis translation is closer to the reference translation[10].

BLEU is used as the *de facto* standard automatic evaluation metric for major machine translation shared tasks. And BLEU continues to show high correlations primarily for n-gram -based machine translation systems (Bojar et al., 2015; Nakazawa et al., 2014).

However, the fallacy of BLEU-human correlations can be easily highlighted with the following example:

[10] Alternatively, researchers would choose to scale the BLEU score to a range between 0 to 100 to improve readability of the scores without the decimal prefix.

Source:

이러한작용을발휘하기위해서는, 각각 0.005%이상함유하는것이바람직하다.

Hypothesis:

このような作用を発揮するためには、夫々０．００５％以上含有することが好ましい。

Baseline:

このような作用を発揮するためには、それぞれ０．００５％以上含有することが好ましい。

Reference:

このような作用を発揮させるためには、夫々０．００５％以上含有させることが好ましい。

Source/Reference English Gloss:

"So as to achieve the reaction, it is preferable that it contains more 0.005% of each [chemical]"

The unigram, bigram, trigrams and fourgrams (p_1, p_2, p_3, p_4) precision of the hypothesis translation are 90.0, 78.9, 66.7 and 52.9 respectively. The p_n score for the hypothesis sentence precision score for the reference is 70.75. When considering the brevity penalty of 0.905, the overall BLEU is 64.03. Comparatively, the n-gram precisions for the baseline translations are p_1=84.2, p_2=66.7, p_3=47.1 and p_4=25.0 and the overall BLEU is 43.29 with a BP of 0.854. In this respect, one would consider the baseline translation inferior to the hypothesis with a >10 BLEU difference. However, there is only a subtle difference between the hypothesis and the baseline translation (それぞれvs 夫々, which both has the same meaning).

This is an actual example from the 2$^{\text{nd}}$ Workshop on Asian Translation (WAT 2015) MT shared task evaluation, and five crowd-sourced evaluators consider the baseline translation a better translation. For this particular example, the human evaluators preferred the natural translation from Korean 각각 *gaggag* to Japanese それぞれ *sorezore* instead of the patent document usage of 夫々 *sorezore*, both それぞれ and 夫々 can be loosely translated as '*respectively*' or '*(for) each*' in English.

The big difference in BLEU for a single lexical difference in translation is due to the geometric averaged scores for the individual n-gram precisions. It assumes the independence of n-gram precisions and accentuates the precision disparity by involving the single lexical difference in all possible n-gram s that capture the particular position in the sentence. This is clearly indicated by the growing precision difference in the higher order

n-grams.

6.2.2 RIBES

Another failure of BLEU is the lack of explicit consideration for reordering. Callison-Burch et al. (2006) highlighted that since BLEU only takes reordering into account by rewarding the higher *n*-gram orders, freely permuted unigrams and bigrams matches are able to sustain a high BLEU score with little penalty caused by tri/fourgram mismatches. To overcome reordering, the RIBES score was introduced by adding a rank correlation coefficient[11] prior to unigram matches without the need for higher order *n*-gram matches (Isozaki et al., 2010).

To account for word ordering, the RIBES metric first determines the word rank correlation of the word alignments by computing the Kendall's τ coefficient. Given the reference and a hypothesis translation:

Reference: *Alice kisses Bob yesterday*

Hypothesis: *Bob kisses Alice yesterday*

The first word 'Alice' in the reference shifted to the third word in the hypothesis and the third word 'Bob' becomes the first. The second and fourth word remains intact. From the original [0, 1, 2, 3] word order of the reference, we get the word order list [2, 1, 0, 3] in the hypothesis; where 0 = '*Alice*', 1 = '*kisses*', 2 = '*Bob*' and 3 = '*yesterday*'.

Given the [2, 1, 0, 3] word order list of the hypothesis, we extract all possible pair of words: [(2, 1), (2, 0), (2, 3), (1, 0), (1, 3), (0, 3)]

The number of all pairs of words can be determined as ${}^4C_2$ = 6. Then, we extract the pairs of words with increasing order, i.e. [(2,3), (1,3), (0,3)].

And the Kendall's τ coefficient for the word order list is compute as such:

$$\tau = 2 \times \frac{no.\ of\ increasing\ pairs}{no.\ of\ all\ pairs} - 1 \tag{6.12}$$

The τ for the particular hypothesis in the example above is τ is 2 x 3/6 - 1 = 0.0. The τ coefficient ranges between [-1, 1].

[11] normalized Kendall τ of all *n*-gram pairs between the hypothesis and reference translations

To ensure that the final RIBES score ranges between 0.0 to 1.0, th e τ in the above formulation refers to the normalized Kendall τ computed as such:

$$\tau_{norm} = (\tau + 1)/2 \tag{6.13}$$

Simplifying BLEU's n-th order n-grams precision, the RIBES score only considers the unigram precision, p_1 using the same Equation (6.9) with n=1. Similarly, the RIBES brevity penalty, BP, follows the BLEU formulation in Equation (6.10)

The final RIBES computation scales the unigram precision and brevity penalty by the Kendall's τ coefficient.

$$RIBES = \tau_{norm} * \ \alpha(p_1) * \ \beta(BP) \tag{6.14}$$

where, α and $beta$ are hyperparameter used as a prior for the unigram precision and brevity penalty. They are set as α=0.25 and β=0.10 based on the correlation with human evaluation in (Isozaki et al., 2010).

Let us consider another example:

Source:

T용융(DSC) = 89.9℃; T결정화(DSC) = 72℃(5℃/ 분에서DSC 로측정) .

Hypothesis:

Ｔｍｅｌｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣ測定（ＤＳＣ）＝89.9結晶化度（Ｔ）。

Baseline:

Ｔ溶融（ＤＳＣ）＝８９．９℃；Ｔ結晶化（ＤＳＣ）＝７２℃（５℃／分でＤＳＣで測定）。

Reference:

Ｔｍｅｌｔ（ＤＳＣ）＝８９．９℃；Ｔｃｒｙｓｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣを用いて測定）。

Source/Reference English Gloss:

Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)

The example above shows the marginal effectiveness of RIBES when penalizing wrongly ordered phrases in the hypothesis. The baseline translation accurately translates the meaning of the sentence with a minor partial translation of the technical variables (i.e. *Tmelt* -> Ｔ溶融 and T결정화 -> Ｔ結晶化. However, the hypothesis translation made serious adequacy errors when inverting the values of the technical variables but the hypothesis translation was minimally penalized in RIBES and also BLEU.

The RIBES score for the hypothesis and baseline translations are 94.04 and 86.33 respectively whereas their BLEU scores are 53.3 and 58.8. In the WAT 2015 evaluation, five evaluators unanimously voted in favor for the baseline translation. Although the RIBES score presents a wider difference between the hypothesis and baseline translation than BLEU, it is insufficient to account for the arrant error that the hypothesis translation made.

6.2.3 Other Shades of BLEU / RIBES

It is worth noting that there are other automatic MT evaluation metrics that depend on the same precision-based score with primary differences in how the $\mathit{Count}_{match}\mathit{(ngram)}$ is measured; Gimenez & Marquez (2007) described other linguistic features that one could match in place of surface *n*-grams , such as lexicalized syntactic parse features, semantic entities and roles annotations, etc. As such, the modified BLEU-like metrics can present other aspects of syntactic fluency and semantic adequacy complementary to the string-based BLEU.

A different approach to improve upon the BLEU scores is to allow paraphrases or gappy variants and replace the proportion of $\mathit{Count}_{match}\mathit{(ngram)}$ */* Count(ngram) by a lexical similarity measure. Banerjee & Lavie (2005b) introduced the METEOR metric that allows hypotheses' *n*-gram s to match paraphrases and stems instead of just the surface strings. Lin & Och (2004b) presented the ROUGE-S metrics that uses skip-gram matches. More recently, pre-trained regression models based on semantic textual similarity and neural network-based similarity measures trained on skip-grams are applied to replace the *n*-gram matching (Vela & Tan, 2015b; Gupta et al., 2015b).

While enriching the surface *n*-gram matching allows the automatic evaluation metric to handle variant translations, it does not resolve the “prominent crudeness" of BLEU (Callison-Burch, 2006) involving (i) the omission of content-bearing materials not being penalized, and (ii) the inability to calculate recall despite the brevity penalty.

6.2.4 Experimental Setup

We describe our experiment setup to evaluate Korean to Japanese patent translation using the Japan Patent Office (JPO) Patent Corpus. The JPO Patent Corpus is the official resource provided for the WAT 2015 shared task. The training dataset is made up of 1 million sentences (250k each from the chemistry, electricity, mechanical engineering and physics domains). Two development datasets[12] and one test set each comprises 2000 sentences with 500 sentences from each of the training domains. The Korean and Japanese texts were tokenized using KoNLPy (Park & Cho, 2014) and MeCab (Kudo et al., 2004) respectively.

We used the phrase-based SMT implemented in the Moses toolkit Koehn et al. (2003, 2007) with the configurations as described in Chapter 2 (Section 2.4).

A More Vanilla Baseline System

Parameters	**Organizers**	**Ours**
Input document length	40	80
Korean tokenizer	MeCab	KoNLPy
Japanese tokenizer	Juman	MeCab
LM *n*-gram order	5	5
Distortion limit	0	20
Quantized & binarized LM	no	yes
`devtest.txt` in LM	no	yes
Binarized phrase tables	no	yes
MERT Runs	1	2

Table 6.2: Differences between Organizer's and our Phrase-based SMT system

Human evaluations were conducted as pairwise comparisons between translations from our system and the WAT organizers' phrase-based statistical MT baseline system. Table 6.2 highlights the parameter differences between the organizers and our phrase-based SMT system.

Human Evaluation (Pairwise Comparison)

The human judgment scores for the WAT evaluations were acquired using the Lancers crowdsourcing platform. Human evaluators were randomly assigned documents from the test set. They were shown the source document, the hypothesis translation and a baseline translation generated by the phrase-based MT system. Five evaluators were asked to judge each document.

[12] `dev.txt` and `devtest.txt`

The crowdsourced evaluators were non-experts, thus their judgements were not necessary precise, especially for patent translations. The evaluators were asked to judge whether the hypothesis or the baseline translation was better, or they were tied. The translation that was judged better constituted a *win* and the other a *loss*. For each, the majority vote between the five evaluators for the hypothesis decided whether the hypothesis *won*, *lost* or *tied* the baseline. The final human judgment score, *HUMAN*, is calculated as follows:

$$\text{HUMAN} = 100 \times \frac{W - L}{W + L + T} \tag{6.15}$$

By definition, the *HUMAN* score ranges from -100 to $+100$, where higher is better.

6.2.5 Results

Moses' default parameter tuning method, MERT, is non-deterministic, and hence it is advisable to tune the phrase-based model more than once (Clark et al. 2011). We repeated the tuning step and submitted the system translations that achieved the higher BLEU score on the development set for manual evaluation.

As a sanity check we also replicated the organizers' baseline system and submitted it for manual evaluation. We expect this system to score close to zero. We submitted a total of three sets of output to the WAT 2015 shared task, two of which underwent manual evaluation.

Systems	RIBES	BLEU	HUMAN
Organizers' PBMT baseline	94.13	69.22	0.0
Our replica baseline	94.29	70.23	**+3.50**
Ours (MERT 1)	95.03	84.26	-
Ours (MERT 2)	**95.15**	**85.23**	-17.75

Table 6.3: BLEU and HUMAN scores for WAT 2015

Table 6.3 presents the BLEU scores achieved by our phrase-based MT system in contrast to the organizers' baseline phrase-based system. The difference in BLEU between the organizers' system and ours may be due to our inclusion of the second development set in building our language model and the inclusion of more training data by allowing a maximum of 80 tokens per document as compared to 40 (see Table 6.2).

However, the puzzling fact is that our system being 15 BLEU points better than the organizers' baseline begets a terribly low human judgement score. We discuss this next.

6.2.6 Segment Level Meta-Evaluation

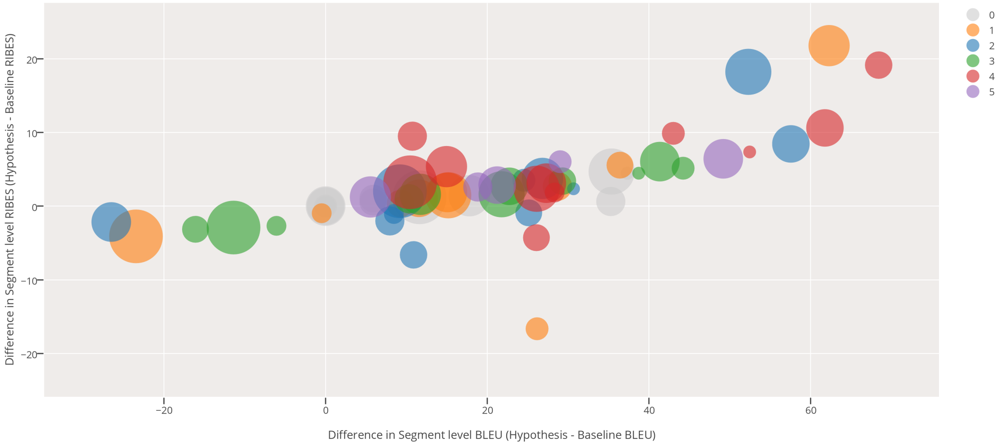

Figure 6.3: Correlation between BLEU, RIBES differences and Positive HUMAN Judgements (HUMAN Scores of 0, +1, +2, +3, +4 and +5 represented by the colored bubbles: *grey, orange, blue, green, red and purple*; larger area means more segments with the respective HUMAN Scores)

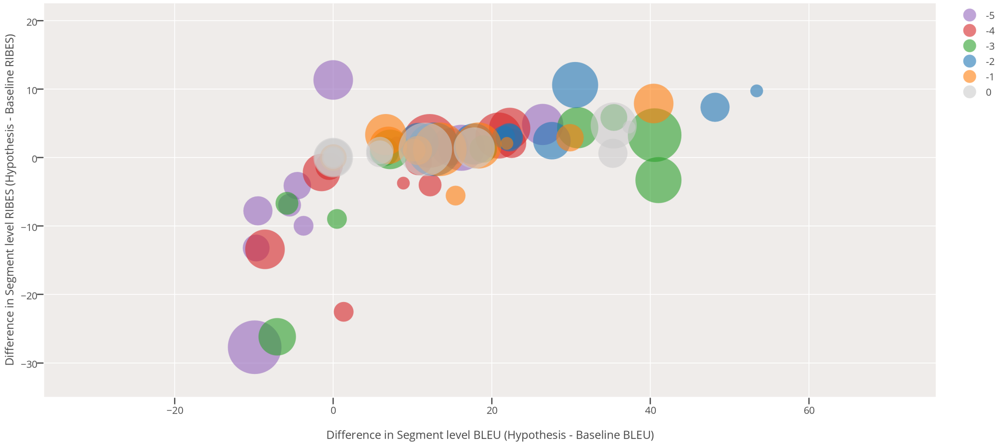

Figure 6.4: Correlation between BLEU, RIBES differences and Negative HUMAN Judgements (HUMAN Scores of 0, -1, -2, -3, -4 and -5 represented by the colored bubbles: *grey, orange, blue, green, red and purple*; larger area means more segments with the respective HUMAN Scores)

We perform a segment level meta-evaluation by calculating the BLEU and RIBES score difference for each hypothesis-baseline translation. Figures 6.3 and 6.4 show the correlations of the BLEU and RIBES score difference against the positive and negative human judgements score for every sentence.

Figure 6.3 presents the considerable incongruity between our system's high BLEU improvements (>+60 BLEU) being rated marginally better than the baseline translation, indicated by the orange and blue bubbles on the top right corner. There were even translations from our system with >+40 BLEU improvements that tied with the organizer's baseline translations, indicated by the grey bubbles at around the +40 BLEU and +5 RIBES region. Except for a portion of segments that scored worse than the baseline system (lower right part of the graph where BLEU and RIBES falls below 0), the overall trend in Figure 6.3 presents the conventional wisdom that the BLEU improvements from our systems reflects positive human judgement scores.

However, Figure 6.4 presents the awkward disparity where many segments with BLEU improvements were rated strongly as poorer translations when compared against the baseline. Also, many segments with high BLEU improvements were tied with the baseline translations, indicated by the grey bubbles across the positive BLEU scores.

As shown in the examples in Section 2, a number of prominent factors contribute to this disparity between high BLEU / RIBES improvements and low HUMAN judgement scores:

- Minor lexical differences causing a huge difference in *n*-gram precision
- Crowd-sourced *vs*. expert preferences on terminology, especially for patents
- Minor MT evaluation metric differences not reflecting major translation inadequacy

Each of these failures contributes to an increased amount of disparity between the automatic translation metric improvements and human judgement scores.

6.2.7 Summary

In this section, we have demonstrated a real-world case where high BLEU and RIBES scores do not correlate with better human judgement. We presented several factors that might contribute to the poor correlation, and also performed a segment level meta-evaluation to identify segments where our system's high BLEU / RIBES improvements were deemed substantially worse than the baseline translations. We hope our results and analysis will lead to improvements in automatic translation evaluation metrics.

6.3 Summary

Taking a short detour from terminology and ontology, Section 6.1 and 6.2 in this chapter presented our first attempt in mitigating the inadequacy of current state-of-art machine translation evaluation metrics and propose a semantically motivated metric by casting the machine translation task as a semantic textual similarity task. Then we explored the discrepancy between automatic machine translation evaluation metrics against the human judgments of the quality of the translated text and we found an awkward disparity between them. We hope that the findings in this chapter can help in future studies in the measurement of machine translation quality and eventually improve machine translation in turn.

Chapter 7

Ontology Induction using Neural Vector Space

Semantic ontologies provide structured world knowledge to Artificial Intelligence (AI) and Natural Language Processing (NLP) systems. Traditional broad-coverage taxonomies such as CYC (Lenat, 1995), SUMO (Pease et al., 2002b; Miller, 1995b), YAGO (Suchanek et al., 2007) and Freebase (Bollacker et al., 2008) have been manually created or curated with much effort and yet they suffer from coverage sparsity. This motivated the move towards unsupervised approaches to extract structured relational knowledge from texts (Lin & Pantel, 2001; Snow et al., 2006; Velardi et al., 2013).

With the rapid technological evolution, it is more feasible to automatically construct a domain-specific taxonomy that caters to sector or company specific terminology (Lefever, 2015). This motivated the move towards unsupervised approaches to taxonomy extraction (Berland & Charniak, 1999; Lin & Pantel, 2001; Snow et al., 2006) and specifically focused towards particular domains (Velardi et al., 2013; Bordea et al., 2015b).

Previous work in taxonomy extraction focused on rule-based, clustering and graph-based approaches. The hierarchical structure of domain concepts is made up of hypo-hypernymy relations between terms. Different approaches have been proposed to induce these relations automatically ranging from pattern/rule-based approaches (Hearst, 1992; Girju, 2003; Kozareva et al., 2008; Ceesay & Hou, 2015) to clustering and frequency based approaches (Lin, 1998; Caraballo, 2001; Pantel & Ravichandran, 2004; Grefenstette, 2015), classification approaches (Snow et al., 2004; Ritter et al., 2009; Espinosa Anke et al., 2015) and graph-based approaches (Kozareva & Hovy, 2010; Navigli et al., 2011; Fountain & Lapata, 2012; Tuan et al., 2014; Cleuziou et al., 2015) (See Chapter 2, Section 2.5.3).

More recently, there is a resurgence of vector space or distributional approaches (Van Der Plas, 2005; Lenci & Benotto, 2012; Santus et al., 2014) primarily because of the renaissance of deep learning and network networks. Semantic knowledge can be thought of as a vector space where each word is presented by a point and the proximity between words in this space quantifies their semantic association. The vector space is usually constructed from the distribution of words across context such that similar meanings tend to be found close to each other within the vector space (Mitchell & Lapata, 2010).

With the present advancement in neural nets and word embeddings (Mikolov et al., 2013b; Pennington et al., 2014b; Levy et al., 2014; Shazeer et al., 2016), neural space models are gaining popularity in taxonomy induction and relation extraction tasks (Saxe et al., 2013; Fu et al., 2014; Tan et al., 2015b).

Fu et al. (2014) proposed a vector space approach to hypernym identification using word embeddings that trains a projection matrix that converts a hyponym vector to its hypernym. However, their approach requires an existing hypernym-hyponym pairs for training before discovering new pairs.

Instead of learning a supervised transition matrix Φ, we capitalize on the fact that hypernym-hyponym pairs often occur in a sentence with an ‘is a’ phrase, e.g. “The goldfish (Carassius auratus auratus) is a freshwater fish”.[1]

We propose a simpler unsupervised approach where we learn a vector for the phrase “*is-a*". We single-tokenize the adjacent “is" and “a" tokens and learn the word embeddings with *is-a* forming part of the vocabulary in the input matrix.

7.1 Inducing a Hypernym with *is-a* Vector

Effectively, we hypothesize that Φ can be replaced by the “*is-a*" vector. To achieve the piecewise projection effects of Φ, we trained a different deep neural net model for each Taxonomy Extraction Evaluation (TaxEval) domain (Bordea et al., 2015a) and assume that the “*is-a*" scales automatically across domains. For instance, the multiplication of the $v(\texttt{tiramisu})$ and the $v(\texttt{is-a}_{food})$ vectors yields a proxy vector and we consider the top ten word vectors that are most similar to this proxy vector as the possible hypernyms, i.e. $v(\texttt{tiramisu}) \times v(\texttt{is-a}_{food}) \approx v(\texttt{cake})$.

There is little or no previous work that manipulates non-content word vectors in vector space models for natural

[1] http://en.wikipedia.org/wiki/Goldfish

language processing. Often, non-content words[2] were implicitly incorporated into the vector space models by means of syntactic frames or syntactic parses (Sarmento et al., 2009).

Our main contribution for ontological induction using vector space models are primarily (i) the use of non-content word vectors and (ii) simplifying a previously complex process of learning a hyper-hyponym transition matrix[3].

7.1.1 Experimental Setup

Similar to Fountain & Lapata (2012), the SemEval-2015 Taxonomy Extraction Evaluation (TaxEval) task addresses taxonomy learning without the term discovery step, i.e. the terms for which to create the taxonomy are given (Bordea et al., 2015a). The focus is on creating the hypernym-hyponym relations. We will be using this dataset to evaluate our hypernym induction system using the *'is-a'* vector.

In the TaxEval task, ontologies are evaluated through comparison with gold standard taxonomies. There is no training corpus provided by the organisers of the task and the participating systems are to generate hyper-hyponyms pairs using a list of terms from four different domains, viz. chemicals, equipment, food and science.

The gold standards used in evaluation are the *ChEBI ontology* for the chemical domain (Degtyarenko et al., 2008), the *Material Handling Equipment taxonomy*[4] for the equipment domain, the *Google product taxonomy*[5] for the food domain and the *Taxonomy of Fields and their Different Sub-fields*[6] for the science domain. In addition, all four domains are also evaluated against the sub-hierarchies from the WordNet ontology that subsumes the Suggested Upper Merged Ontology (Pease et al., 2002a).

There is no specified training corpus released for the SemEval-2015 TaxEval task. To produce a domain specific corpus for each of the given domains in the task, we used the Wikipedia dump and preprocessed it using WikiExtractor[7] and then extracted documents that contain the terms for each domain individually.

We trained a skip-gram model phrasal word2vec neural net (Mikolov et al., 2013a) using gensim (Řehůřek & Sojka, 2010). The neural nets were trained for 100 epochs with a window size of 5 for all words in the corpus.

[2] Words that are not noun (entities/arguments), verbs (predicates), adjectives or adverbs (adjuncts).
[3] Parts of the research reported in this chapter has been published in Tan et al. (2015b)
[4] http://www.ise.ncsu.edu/kay/mhetax/index.htm
[5] http://www.google.com/basepages/producttype/taxonomy.en-US.txt
[6] http://sites.nationalacademies.org/PGA/Resdoc/PGA_044522
[7] We use the same Wikipedia dump to text extraction process from the SeedLing - Human Language Project (Emerson et al., 2014).

7.1.2 Evaluation Metrics

For the TaxEval task, the multi-faceted evaluation scheme presented in Navigli (2013) was adopted to compare the overall structure of the taxonomy against a gold standard, with an approach used for comparing hierarchical clusters. The multi-faceted evaluation scheme evaluates (i) the structural measures of the induced taxonomy (left columns of Table 6.1), (ii) the comparison against gold standard taxonomy (right columns of Table 6.1 and leftmost column of Table 6.2) and (iii) manual evaluation of novel edges precision (last row of Table 6.2).

Regarding the two types of automatic evaluation measures, the structural measures provide a gauge of the system's coverage and the ontology structural integrity, i.e. "tree-likeness" of the ontology produced by the hypernym-hyponym pairs, and the comparison against the gold standards gives an objective measure of the "human-likeness" of the system in producing a taxonomy that is similar to the manually-crafted taxonomy.

7.2 Results

	|V|	**|E|**	**#c.c**	**cycles**	**#VC**	**%VC**	**#EC**	**%EC**	**:NE**
Chemical	13785	30392	302	YES	13784	**0.7838**	2427	0.0977	1.1268
Equipment	337	548	28	YES	336	0.549	227	0.3691	0.5219
Food	1118	2692	23	YES	948	0.6092	428	0.2696	1.4265
Science	355	952	14	YES	354	0.7831	173	**0.3720**	1.6752
WN Chemical	1173	3107	31	YES	1172	**0.8675**	532	**0.3835**	1.8566
WN Equipment	354	547	43	YES	353	0.7431	149	0.3072	0.8206
WN Food	1200	3465	23	YES	1199	0.8068	549	0.3581	1.9021
WN Science	307	892	8	YES	306	0.7132	156	0.3537	1.6689

Table 7.1: Structural Measures and Comparison against Gold Standards for `USAAR-WLV`. The labels of the columns refer to no. of distinct vertices and edges in induced taxonomy (|**V**| and |**E**|), no. of connected components (**#c.c**), whether the taxonomy is a Directed Acyclic Graph (**cycles**), vertex and edge coverage, i.e. proportion of gold standard vertices and edges covered by system (**%VC** and **%EC**), no. of vertices and edges in common with gold standard (**#VC** and **#EC**) and ratio of novel edges (**:NE**).

	INRIASAC	LT3	NTNU	QASSIT	TALN-UPF	USAAR
Avg. F&M	0.3270	**0.4130**	0.0580	0.3880	0.2630	0.0770
Avg. Precision	0.1721	**0.3612**	0.1754	0.1563	0.0720	0.2014
Avg. Recall	0.4279	**0.6307**	0.2756	0.1588	0.1165	0.3139
Avg. F-Score	0.2427	**0.3886**	0.2075	0.1575	0.0798	0.2377
Avg. Precision of NE	0.4800	**0.5960**	0.3530	0.2470	0.1020	0.4200

Table 7.2: Averaged F&M Measure, Precision, Recall, F-score for All Systems Outputs when Compared to Gold Standard and Manually Evaluated Average Precision of Novel Edges.

Table 6.2 presents the evaluation scores for our system (USAAR) in the TaxEval task, the %VC and %EC scores summarize the performance of the system in replicating the gold standard taxonomies.

In terms of vertex coverage, our system performs best in the chemical and WordNet chemical domain. Regarding edge coverage, our system achieves highest coverage for the science domain and WordNet chemical domain. Having high edge and vertex coverage significantly lowers false positive rate when evaluating hypernym-hyponyms pairs with precision, recall and F-score.

We also note that the wikipedia corpus extracted that we used to induce the vectors lacks coverage for the food domain. In the other domains, we discovered all terms in the wikipedia corpus plus the domains' root hypernym (i.e. $|\mathbf{V}| = \mathbf{\#VC} + 1$).

Table 6.2 presents the comparative results between the participating teams in the TaxEval task averaged over all domains. We performed reasonably well as compared to the other systems in all measures. While our system's Fowlkes and Mallows measure (F&M) is low, it is only representative of the clusters we have induced as compared to the gold standard. To improve our F&M measure, we could reduce the number of redundant novel edges by pruning our system outputs and achieve comparable results to the other teams given our relatively precision of novel edges.

7.3 Hyponym Endocentricity

Early research in theoretical linguistics discussed the idea of *endocentric* vs. *exocentric* constructions (Brugmann, 1886; Aleksandrov, 1886; Brockelmann, 1908; Bloomfield, 1983).

A grammatical construction is *endocentric* when it fulfils the same linguistic function as one of its parts. For instance, the word *goldfish* is an endocentric compound noun that shares the syntactic noun properties of *fish* and semantically the compound denotes a type of *fish*.

Conversely, when a grammatical construction made up of two or more parts is not *endocentric*, the construction would be exocentric such that no one part of the construction contains the main meaning of the word. Intuitively, we can perceive that there are many endocentric hyponyms in a taxonomy where part of the term conveys its main meaning and usually that part of term would be its hypernym.

While experimenting with ways to weight a term for information retrieval, Jones (1979) observed that compound nouns follow the head-modifier principle where the meaning of the term can be conveyed by part(s) of the compound. In the first TExeval task in SemEval-2015, both Lefever (2015) and Tan et al. (2015b)[8] independently developed string-based systems that exploit the endocentric nature of hyponyms.

In this section, we seek to answer the question of exactly ***"how many of hyponyms within a taxonomy are endocentric?"***. Additionally, we exploit the endocentric nature of the hyponyms to extend the taxonomy by trawling Wikipedia *List of Lists of Lists*[9]. Often these lists of terms are found in Wikipedia marked up tables or in bullet forms.

7.3.1 Identifying Endocentric Parts

The main implementation of the rule-based identifier[10] checks ***if a term T1 is a substring of T2*** **and if so,** ***assign T1 as a hypernym of T2***. Examples of hypo-hypernym pairs captured by this rule includes are (*linguistics, psycholinguistics*), (*beef, kobe beef*), (*sauce, sauce gribiche*).

Our implementation is a little simpler than the three part morpho-syntactic analyzer component of the multi-modular taxonomy constructor in Lefever (2015). She implemented rules for three different syntactic constructions where they check for suffix and treat single-word terms and multi-word terms differently while our implementation is agnostic to the single and multi-word distinction.

In addition to first, ***if a term contains the "of" preposition, we swap the assignment and check that T2 starts with T1 then assign T2 as a hypernym of T1***. Examples of hypo-hypernym pairs captured by this swap rule are (*elixir of life, elixir*), (*sociology of education, sociology*).

[8] https://github.com/alvations/USAAR-SemEval-2015/tree/master/task17-USAAR-WLV

[9] https://en.wikipedia.org/wiki/List_of_lists_of_lists

[10] Our open-source implementation can be found at https://github.com/alvations/Endrocentricity

To improve the precision of the identifier, we set a threshold of a minimum character length of three when identifying a term as a hypernym.

7.4 Extending Taxonomy with Wikipedia List of Lists of Lists

The Wikipedia List of Lists of Lists (LOLOL) is a crowd-sourced list of lists of terms that belong to their respective categories. We adapted the a customized crawler[11] (Tan et al., 2014; Tan & Ordan, 2015) to crawl for tables or bullet points in Wikipedia the subpages of the LOLOL for the food domain. We started the crawl from these seed pages under the bullet point of `https://en.wikipedia.org/wiki/List_of_lists_of_lists#Food_and_drink`.

When the crawler lands on each List of List (LOL) page, it will ***treat the URL suffix as the hypernym*** and ***find words in the bullet points or tables that contains endocentric hyponyms***.

If an endocentric hyponym exists, it will extract either extract (i) all the bolded terms if the LOL page is bulletined or (ii) all terms in the first column if the LOL page is in table form. The choice of the first column is based on the fact that often LOL tables are bi-column, one containing the terms and the other the gloss or/and description of the term.

7.4.1 Limitations of LOLOL Trawler

However, there are a couple issues with this ***trawling*** (*crawl+clean*) approach to extend the taxonomy.

LOL Pages are Not Standardized: The way the crawler cleans the bullets or tables on each LOL page is not standardized because there is not constraint put on the format of the Wikipedia's LOL page, our crawler only managed to crawl and clean less than 10 LOL pages when extracting the new terms for the food domain.

LOL Pages are Inceptive: The depth of how nested the LOLs are is undefined. Our crawler can start with a `List_of_foods` page and it leads to the `List_of_breads` page and then the `List_of_American_breads` page and it continues. For sanity, we had to break our trawler at the second page depth and return to the main LOLOL page to move on to the next LOL that we have not previously trawled.

[11] It was built for crawling translations and diachronic texts in previous SemEval tasks

7.4.2 Results

	Environment	**Food**	**Food**	**Science**	**Science**	**Science**
	(Eurovoc)	(WordNet)		(Eurovoc)	(WordNet)	
#Terms	261	1486	1555	370	125	452
#Relations	261	1576	1587	452	124	465
#Correct / Identified	38 / 47	381 / 540	– / 4347*	66 / 104	25 / 30	119 / 312
Precision	**0.8085**	**0.7056**	0.0603	**0.6333**	**0.8173**	0.3814
Recall	0.1456	0.2418	0.1651	0.1532	0.1881	0.2559
F-score	0.2468	**0.3601**	0.0883	0.2468	0.3058	0.3063
F&M	0.0007	0.0021	0.0	0.0023	0.0008	0.0020

Table 7.3: Results of Our Endocentric Hypo-Hypernym Identifier Against the Gold Standard Taxonomy (**#Terms** refers to the no. of terms in the domain and **#Relations** refers to the no. of hypo-hypernym pairs found in the gold-standard taxonomy. **#Correct / #Identified** refers to the proportion of hypo-hypernym pairs our system has correctly identified. **Bold** items indicates that it is highest score among the competing teams in TExEval-2. The asterisk * indicates that the trawler was used to produce submissions for this domain.)

Domain	**JUNLP**	**TAXI**	**NUIG-UNLP**	**QASSIT**	**USAAR**
Environment (Eurovoc)	0.02	0.11	0.08	0.07	**0.22**
Food	0.2	0.36	–	–	**0.73***
Food (Wordnet)	0.18	0.32	–	–	**0.81**
Science	0.06	0.14	0.09	0.07	**0.71**
Science (Eurovoc)	0.02	0.02	**0.04**	0.05	0.00
Science (Wordnet)	0.06	0.22	0.05	0.22	**0.47**

Table 7.4: Results of a Manual Evaluation on 100 Random Novel Hypo-Hypernym Pairs for Competing Teams In TExEval-2

Table 6.3 presents the overview results of our submissions (USAAR) to the TExEval-2 task. Only the results for the food domain contains the hypo-hypernym pairs extracted by our trawler. The rest of the domains comprise of the outputs solely generated by our endocentric hypo-hypernym identifier.

Although it's counter-intuitive to think that endocentric hypo-hypernym pairs can be wrong, the following example aptly demonstrates the limitations of our approach: (*honey bunches of oats, honey*). In this case, neither '*honey bunches of oats*' can be a hypernym of '*honey*' or vice versa.

When compared against the gold standard taxonomies, our submission achieved the highest precision in the enviornment, food (WordNet), science (Eurovoc) and science (WordNet) domains.

As for the Food domain, we have expected the fall in precision due to the additional terms that we have introduced from the Wikipedia LOLOL outside of the gold standard taxonomy. Thus, we are also unable to determine the true “correctness" of these terms (indicated by the dash in Table 1).

Looking at the proportion of the number of hypo-hypernym pairs that our system correctly identified, we can empirically claim that *15-25% of the hypernyms in a taxonomy can be easily identified through their endocentric hyponyms* by taking the ratio of #Correct / #Terms.

However, the proportions presented in Table 6.3 exclude the correct hypo-hypernym pairs that are identified but are not currently in the gold-standard taxonomy. Table 6.4 presents the results of the manual evaluation for the precision of 100 randomly selected hypo-hypernym pairs that are not in the gold standard taxonomy. Our system achieved top precision in all domains other than the science (Eurovoc).

If we consider the precision scores from Table 6.4 as the precision of the remaining identified but not correct hypo-hypernym pairs in Table 6.3, we might be able to add to the empirical claim of 15-25% hypononym endocentricity in taxonomies. However, the aggregation of the manual evaluation results should only be considered if the novel hypo-hypnym relations are curated and added to the standard taxonomies.

		TExEval-2 (Baseline)	**Lefever** (2015)	**Ours**
Food	P	0.5000	0.602	**0.7056**
(WordNet)	R	**0.2576**	0.176	0.2418
Science	P	0.6897	0.696	**0.8173**
(WordNet)	R	**0.2655**	0.270	0.1881

Table 7.5: Comparison of String-based Methods

Comparing against the TExEval-2 organizers baseline string-based method and Lefever's (2015) morpho-syntactic module for the WordNet taxonomies, our system achieved highest precision but underperfomed in recall as shown in Table 6.5.

Since our main implementation of our hypernym identifier is language independent, in retrospect, we can easily remove the swap rule that is attached to the English ‘*of*’ and apply it to other languages in the TExEval-2 task.

7.5 Summary

Our vector space hypernym generator achieved modest results when compared against other participating teams. Given the simple approach to hypernym-hyponym relations, it is possible that future research can apply the method to other non-content word vectors to induce other relations between entities.

In exploring hypernym endocentricity, we have empirically shown that 15-25% of the hypernyms in an ontology can be easily identified through their endocentric hyponyms and we briefly discuss the intuitions and limitations of the approach. We have achieved competitive results in taxonomy construction and achieved top precisions for hypernym identification in most domains involved in the task.

Chapter 8

Using Sub-Ontological Knowledge to Improve Machine Translation

Words can be grouped together into equivalence classes to help reduce data sparsity and better generalize data. Word clusters can be seen as an intermediate representation of knowledge that is more descriptive than a flat list of terms and less expressive than a full ontology. As such, we consider word clusters to be ***sub-ontological knowledge*** that can be easily incorporated into many NLP and MT applications without the overhead of considering the full graphical representation of an ontology.

Within machine translation word classes are used in word alignment (Brown et al., 1993a; Och & Ney, 2000), factored machine translation (Koehn & Hoang, 2007b) and translation models (Koehn & Hoang, 2007b; Wuebker et al., 2013), reordering (Cherry, 2013), preordering (Stymne, 2012), target-side inflection (Chahuneau et al., 2013), SAMT (Zollmann & Vogel, 2011), sparse word features (Haddow et al., 2015), and OSM (Durrani et al., 2014), among many others.

Word clusterings have also found utility in parsing (Koo et al., 2008; Candito & Seddah, 2010; Kong et al., 2014), semantic parsing (Zhao et al., 2009), chunking (Turian et al., 2010), NER (Miller et al., 2004; Liang, 2005; Ratinov & Roth, 2009; Ritter et al., 2011), tweet tagging (Owoputi et al., 2013; Nooralahzadeh et al., 2014), structure transfer (Täckström et al., 2012), and discourse relation discovery (Rutherford & Xue, 2014).

Word clusters are useful in training neural and MaxEnt language models. Word clusters also speed up normalization in training neural network and MaxEnt language models, via class-based decomposition (Goodman, 2001b). This reduces the normalization time from $\mathcal{O}(|V|)$ (the vocabulary size) to $\approx \mathcal{O}(\sqrt{|V|})$. Further improvements to $\mathcal{O}(\log(|V|))$ are found using hierarchical softmax (Morin & Bengio, 2005; Mnih & Hinton, 2009).[1]

8.1 Exchange-based Word Clustering

Word clustering partitions a vocabulary *V*, grouping together words that function similarly. This helps generalize language and alleviate data sparsity. We discuss flat clustering in this section of the thesis. Flat, or strict partitioning clustering maps word types onto a smaller set of clusters.

The **exchange algorithm** (Kneser & Ney, 1993) is an efficient technique that exhibits a general time complexity of $\mathcal{O}(|V| \times |C| \times I)$, where $|V|$ is the number of word types, $|C|$ is the number of classes, and I is the number of training iterations, typically < 20. This omits the specific method of exchanging words, which adds further complexity. Words are exchanged from one class to another until convergence or I.

One of the oldest and still most popular exchange algorithm implementations is `mkcls` (Och, 1995), which adds various metaheuristics to escape local optima. Botros et al. (2015) introduce their implementation of three exchange-based algorithms (Martin et al., 1998; Müller & Schütze, 2015)[2]. Clark (2003) adds an orthotactic bias.

The previous algorithms use an unlexicalized (two-sided) language model: $P(w_i|w_{i-1}) = P(w_i|c_i)\, P(c_i|c_{i-1})$, where the class c_i of the predicted word w_i is conditioned on the class c_{i-1} of the previous word w_{i-1}. Goodman (2001a) altered this model so that c_i is conditioned directly upon w_{i-1}, hence: $P(w_i|w_{i-1}) = P(w_i|c_i)\, P(c_i|w_{i-1})$. This new model fractionates the history more, but it allows for a large speedup in hypothesizing an exchange since the history doesn't change. The resulting partially lexicalized (one-sided) class model gives the accompanying **predictive exchange algorithm** (Goodman, 2001a; Uszkoreit & Brants, 2008) a time complexity of $\mathcal{O}((B + |V|) \times |C| \times I)$ where B is the number of unique bigrams in the training set.

Dehdari et al. (2016b) developed a *bidirectional, interpolated, refining, and alternating* (`BIRA`) predictive exchange algorithm. The goal of `BIRA` is to produce better clusters by using multiple, changing models to escape local optima. This uses both forward and reversed bigram class models to improve cluster quality by evaluating log-likelihood on two different models. Unlike using trigrams, bidirectional bigram models only linearly increase time and memory requirements, and in fact some data structures can be shared. The two

[1] Part of the research presented in this chapter has been previously published in Dehdari et al. (2016a), Dehdari et al. (2016b) and Tan (2016d)

[2] use trigrams within the exchange algorithm.

directions are interpolated to allow softer integration of these two models:

$$P(w_i|w_{i-1}, w_{i+1}) = P(w_i|c_i) \cdot (\lambda P(c_i|w_{i-1}) + (1-\lambda)P(c_i|w_{i+1})) \tag{8.1}$$

The interpolation weight λ for the forward direction alternates to $1-\lambda$ every a iterations (i):

$$\lambda_i := \begin{cases} 1-\lambda_0 & \text{if } \; i \bmod a = 0 \\ \lambda_0 & \text{otherwise} \end{cases} \tag{8.2}$$

The time complexity is $\mathcal{O}(2 \times (B + |V|) \times |C| \times I)$. The original predictive exchange algorithm can be obtained by setting $\lambda = 1$ and $a = 0$.[3]

8.2 Experimental Setup

We evaluated the word clusters from the `BIRA` predictive exchange algorithm extrinsically in machine translation. As discussed in the previous section, word clusters are employed in a variety of ways within machine translation systems, the most common of which is in word alignment where `mkcls` is widely used. As training sets get larger every year, `mkcls` struggles to keep pace, and is a substantial time bottleneck in MT pipelines with large datasets. We compare time and BLEU scores of using `mkcls` vs `BIRA` in word alignment.

Similar to the experiments in Chapter 3 (Section 3.5), we used data from the Workshop on Machine Translation 2015 (WMT15) Russian-English dataset and the Workshop on Asian Translation 2014 (WAT14) Japanese-English dataset using the phrase-based machine translation configurations as described in Chapter 2 (Section 2.4).

[3] The time complexity is $\mathcal{O}((B + |V|) \times |C| \times I)$ if $\lambda = 1$.

8.3 Results

$\lvert C \rvert$	**EN-RU**	**RU-EN**	**EN-JA**	**JA-EN**
10	20.8→20.9*	26.2→26.0	23.5→23.4	16.9→16.8
50	21.0→21.2*	25.9→25.7	24.0→23.7*	16.9→16.9
100	20.4→21.1	25.9→25.8	23.8→23.5	16.9→17.0
200	21.0→20.8	25.8→25.9	23.8→23.4	17.0→16.8
500	20.9→20.9	25.8→25.9*	24.0→23.8	16.8→17.1*
1000	20.9→21.1	25.9→26.0**	23.6→23.5	16.9→17.1

Table 8.1: BLEU scores mkcls→BIRA and significance across cluster sizes ($\lvert C \rvert$)

#Clusters	**EN-RU**	**RU-EN**	**EN-JA**	**JA-EN**
10	$+0.1$*	-0.2	-0.14	-0.1
50	$+0.2$*	-0.2	-0.32*	-0.04
100	$+0.7$	-0.1	-0.27	$+0.12$
200	-0.2	$+0.1$	-0.35	-0.15
500	0	$+0.1$*	-0.29	$+0.27$*
1000	$+0.2$	$+0.1$**	-0.07	$+0.13$

Table 8.2: BLEU score changes and significance across cluster sizes.

Table 8.1 presents the absolute BLEU score changes and their statistical significance across cluster size ($\lvert C \rvert$). Table summarizes the change in BLEU scores when replacing word clusters generated by mkcls with clusters from BIRA[4]. The BLEU score differences between using mkcls and our BIRA implementation are small but there are a few statistically significant changes[5], using bootstrap resampling Koehn (2004b).

The maximum difference in BLEU score in any configuration is 0.7 absolute points (EN-RU, $\lvert C \rvert = 100$, BIRA $=$ 21.1, mkcls $=$ 20.4, p $=$ 0.32). We ran the hypothesis tests twice per configuration, and there were no disagreements between each hypothesis testing. Nonetheless, MERT tuning is quite erratic, and some of the BLEU differences could be affected by noise in the tuning process in obtaining quality weight values.

Figure 8.1 shows that BIRA enables MT experiments to explore high cluster sizes without the time-consuming overhead of mkcls. Using the BIRA clustering algorithm reduces the translation model training time with 500

[4] Positive values indicate improvements made by replacing mkcls with BIRA

[5] (*: p-value $<$ 0.05, **: p-value $<$ 0.01)

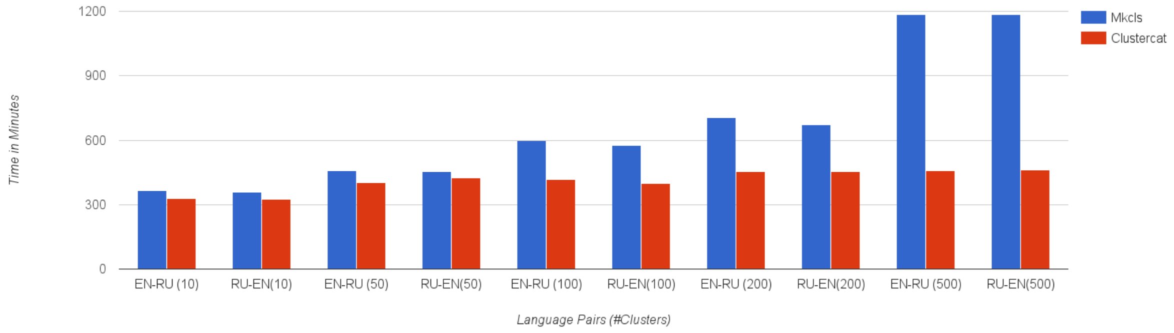

Figure 8.1: End-to-end translation model training times for English-Russian and Russian-English for various cluster sizes using `mkcls` and `BIRA`.

clusters from 20 hours using `mkcls` (of which 60% of the time is spent on clustering) to just 8 hours (of which 5% is spent on clustering). It takes 8 hours to complete the full translation model training with 500 clusters, as compared to 20 hours using `mkcls`. We bring the time for clustering from 60% of the total training time to just 5% .This reduces 5% of the total training time for the `BIRA` clusterer, compared to 60% for `mkcls`. Using our `BIRA` implementation it takes 9 hours to complete the full translation model training with 1000 clusters, as compared to 30 hours using `mkcls` – 7% of the total training time for the `BIRA` clusterer, compared to 74% for `mkcls`.

8.4 Summary

In this Chapter, we incorporated sub-ontological structures (i.e. word clusters) into machine translation through improving the predictive exchange algorithm that address longstanding drawbacks of the original algorithm compared to other clustering algorithms and showed that word alignment models using the `BIRA` implementation fully match those using `mkcls` in BLEU scores, with time savings found by using our improvements.[6]

[6] The software is freely available at `https://github.com/jonsafari/clustercat`.

Chapter 9

Conclusion

I don't know whether machine translation will eventually get good enough to allow us to browse people's websites in different languages so you can see how they live in different countries.

—Tim Berners-Lee

This thesis contributes novel algorithms for terminology extraction and ontology induction with the aim of improving machine translation with terminological and ontological knowledge.

We proposed the Language Model Pointwise Mutual Information PMI_{LM} measure for terminology extraction. PMI_{LM} leverages the robust language model probabilities to estimate the co-occurrence statistics of the individual words within a term using pointwise mutual information. We explored supervised and unsupervised methods of applying the PMI_{LM} measure. When evaluated on the Disease Names and Adverse Effects (DNAE) corpus, our unsupervised approach with PMI_{LM} achieves similar F-score to a rule-based term extracted in the bio-medical domain; our supervised approach also achieved better F-scores compared to off-the-shelf named-entity recognizer trained on the same corpus. Beyond monolingual term extraction, we extended our term extractor using word alignments to extract bilingual terms (Chapter 3).

We have empirically verified that adding lexical information from an automatically extracted terminology or a manually crafted dictionary in machine translation in some cases is able to provide statistically significant but marginal BLEU score improvements. However, the number of times to add the lexicon to MT becomes an additional hyperparameter that does not justify the marginal gains (Chapter 4).

We introduced a novel unsupervised method to induce an ontology using neural vector space. By capturing the vectorial representation of the non-content phrase, 'is-a', we project the cross-product of the hyponym and the 'is-a' vector as a proxy vector and perform a similarity search across candidate hypernyms to extract the top ranking hypernym(s). Our method achieved competitive results in an ontology induction shared task and the unsupervised and parsimonious nature of the approach makes it easy to scale across multiple knowledge domains (Chapter 6).

While terminological information can be easily added passively as additional data to train an SMT system, sub-ontological information (i.e. word clusters) can be incorporated in the word alignment step in SMT. We have introduced an exchange based clustering algorithm that provides substantial speed gains in training a phrase-based SMT system without affecting the translation quality (Chapter 7).

Beyond the terminology and ontology scope of this thesis, we find that measuring the 'goodness' of translation automatically is crucial in machine translation. We experimented with ensembles of machine translation evaluation metrics and created a hybrid system with state-of-art neural net embeddings and machine translation metrics. Our system achieved state-of-art performance in the semantic textual similarity task where one of the sub-task is to determine the similarity between the reference and post-edited sentences. Additionally, we have also highlighted the disparity between BLEU scores and human judgments of machine translation output through extensive meta-evaluation (Chapter 5).

9.1 Future Work

9.1.1 Terminology Extraction

The novelty in the PMI_{LM} measure is to leverage on well-studied language model based statistics in place of traditional co-occurrence. In fact, all statistical co-occurrence based measures[1] used for term extraction can be extended using language model probabilities. As the neural tsunami (Manning, 2016) hits the NLP field, a natural progression is to use neural nets to train supervised term extraction systems. However, the bottleneck might be the lack of term annotated data. Possibly, using high ranking term candidates from existing term extraction statistics can bootstrap the annotation process.

[1] e.g. Contrastive weights (C-Value) and Non-Contrastive weights (NC-value) (Frantzi et al., 1998), Discriminative Weights (DW) (Bonin et al., 2010; Enkhsaikhan et al., 2007), etc.

9.1.2 Ontology Induction

While a general upper ontology remains an infinite expansion of human knowledge, we see the trend of developing domain specific ontologies designed specifically to provide knowledge for specific tasks (Bordea et al., 2015a, 2016; Jurgens & Pilehvar, 2016).

Similar to any other NLP task, neural approaches seem to be a good way to achieving state-of-art performance (Shwartz et al., 2016; Shwartz & Dagan, 2016) in ontology induction. Alternatively, combining linguistically motivated rules, existing ontologies would give mileage to improving existing ontology induction system (Lefever, 2016).

We can easily extend our unsupervised ontology induction method based on the 'is-a' vector by including more phrasal vectors using the patterns that relate hyper-hyponyms as listed in (Hearst, 1992). Additionally, to verify the efficacy of the phrasal vector in our approach, we can conduct experiments to retro-fit the phrasal vector (Faruqui et al., 2015).

9.1.3 Using Terminology and Ontology in Machine Translation

As the dusk of phrase-based machine translation draws closer, neural machine translation (Luong et al., 2015; Sennrich et al., 2015b; Wu et al., 2016; Crego et al., 2016; Kalchbrenner et al., 2016) shows similar tendencies of the '*effective multiplier*'(Brown et al., 1993a). Arthur et al. (2016) showed that automatically extracted word alignment probabilities can bootstrap neural machine translation, leading to improved BLEU scores and faster convergence time. There remains a lack of study in the gain in BLEU from incorporating different lexical knowledge resources in neural machine translation as we have explored in this thesis, i.e. *automatic vs. manual lexicon*, *mono-lexical vs. multi-word*, *varying lexicon sizes*, *etc*.)

Bibliography

Agirre, Eneko, Carmen Banea, Claire Cardic, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau & Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 81–91. Dublin, Ireland.

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria & Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 252–263. Denver, Colorado.

Agirre, Eneko, Daniel Cer, Mona Diab & Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Main Conference and the Shared Task*, 385–393. Montréal, Canada.

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre & Weiwei Guo. 2013. SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, 32–43. Atlanta, Georgia.

Ahmad, Khurshid, Lee Gillam, Lena Tostevin et al. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *Trec*, .

Al-Onaizan, Yaser, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A Smith & David Yarowsky. 1999. Statistical machine translation. In *Final report, jhu summer workshop*, vol. 30, .

Al-Onaizan, Yaser & Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Pro-*

ceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics, 529–536.

Aleksandrov, Alexander. 1886. *Sprachliches aus dem nationaldichter litauens donalitius. i. zur semasiologie. inaugural-dissertation... von alexander aleksandrow,...* Schnakenburg.

Arcan, Mihael, Claudio Giuliano, Marco Turchi & Paul Buitelaar. 2014. Identification of bilingual terms from monolingual documents for statistical machine translation. *COLING 2014* 22.

Arthur, Philip, Graham Neubig & Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1557–1567. Austin, Texas. `https://aclweb.org/anthology/D16-1162`.

Babych, Bogdan & Anthony Hartley. 2004. Extending the bleu mt evaluation method with frequency weightings. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, 621.

Baldwin, Timothy & Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In Takaaki Tanaka, Aline Villavicencio, Francis Bond & Anna Korhonen (eds.), *Second acl workshop on multiword expressions: Integrating processing*, 24–31. Barcelona, Spain: Association for Computational Linguistics.

Banerjee, Satanjeev & Alon Lavie. 2005a. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 65–72. Ann Arbor, Michigan.

Banerjee, Satanjeev & Alon Lavie. 2005b. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for mmachine translation and/or summarization*, vol. 29, 65–72.

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 238–247. Baltimore, Maryland. `http://www.aclweb.org/anthology/P14-1023`.

Barrón-Cedeño, Alberto, Lluís Màrquez, Maria Fuentes, Horacio Rodríguez & Jordi Turmo. 2013. UPC-CORE: What Can Machine Translation Evaluation Metrics and Wikipedia Do for Estimating Semantic Textual Similarity? In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, 143–147. Atlanta, Georgia.

Bechara, Hanna, Rohit Gupta, Liling Tan, Constantin Orasan, Ruslan Mitkov & Josef van Genabith. 2016a. Wolvesaar: Replicating the success of monolingual word alignment and neural embeddings for semantic textual similarity. In *Proceedings of the 10th international workshop on semantic evaluation (semeval 2016)*, San Diego, California.

Bechara, Hannah, Rohit Gupta, Liling Tan, Constantin Orasan, Ruslan Mitkov & Josef van Genabith. 2016b. Wolvesaar at semeval-2016 task 1: Replicating the success of monolingual word alignment and neural embeddings for semantic textual similarity. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, 634–639. San Diego, California: Association for Computational Linguistics. `http://www.aclweb.org/anthology/S16-1096`.

Becher, Johann Joachim. 1962. *Zur mechanischen sprachübersetzung: ein programmierungsversuch aus dem jahre 1661: Jj becher allgemeine verschlüsselung der sprachen (character, pro notitia linguarum universalis, deutsch-lateinisch. mit einer interpretierenden einleitung von... wg waffenschmidt.[stellungnahme von pater roberto busa.].* W. Kohlhammer.

Bengio, Yoshua, Patrice Simard & Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on* 5(2). 157–166.

Berland, Matthew & Eugene Charniak. 1999. Finding Parts in Very Large Corpora. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics*, 57–64.

Bertero, Dario & Pascale Fung. 2015. Hltc-hkust: A neural network paraphrase classifier using translation metrics, semantic roles and lexical similarity features. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 23–28. Denver, Colorado. `http://www.aclweb.org/anthology/S15-2004`.

Biçici, Ergun & Josef van Genabith. 2013. CNGL-CORE: Referential Translation Machines for Measuring Semantic Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, 234–240. Atlanta, Georgia.

Biçici, Ergun & Andy Way. 2014. RTM-DCU: Referential Translation Machines for Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 487–496. Dublin, Ireland.

Bicici, Ergun. 2015. Rtm-dcu: Predicting semantic similarity with referential translation machines. In *Pro-*

ceedings of the 9th international workshop on semantic evaluation (semeval 2015), 56–63. Denver, Colorado. http://www.aclweb.org/anthology/S15-2010.

Biggs, Norman, E. Keith Lloyd & Robin J. Wilson. 1976. *Graph theory 1736-1936*. Clarendon Press.

Birch, Alexandra & Miles Osborne. 2010. Lrscore for evaluating lexical and reordering quality in mt. In *Proceedings of the joint fifth workshop on statistical machine translation and metricsmatr*, 327–332.

Birch, Alexandra, Miles Osborne & Phil Blunsom. 2010. Metrics for mt evaluation: evaluating reordering. *Machine Translation* 24(1). 15–26.

Birch, Alexandra, Miles Osborne & Philipp Koehn. 2007. Ccg supertags in factored statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, 9–16. Association for Computational Linguistics.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with python*. " O'Reilly Media, Inc.".

Bloomfield, Leonard. 1983. *An introduction to the study of language* 2. Benjamins. https://books.google.de/books?id=ymAYoabOG5YC.

Bojar, Ondřej. 2007. English-to-czech factored machine translation. In *Proceedings of the second workshop on statistical machine translation*, 232–239. Association for Computational Linguistics.

Bojar, Ondrej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, 12–58.

Bojar, Ondřej & Jan Hajič. 2008. Phrase-based and deep syntactic english-to-czech statistical machine translation. In *Proceedings of the third workshop on statistical machine translation*, 143–146. Association for Computational Linguistics.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the eighth workshop on statistical machine translation*, 1–44. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-2201.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia & Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the tenth workshop on statistical machine translation*, 1–46. Lisbon, Portugal.

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge & Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 acm sigmod international conference on management of data* SIGMOD '08, 1247–1250. New York, NY, USA: ACM. doi:10.1145/1376616.1376746. http://doi.acm.org/10.1145/1376616.1376746.

Bond, Francis & Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Gwc 2012 6th international global wordnet conference*, vol. 8 4, 64.

Bonin, Francesca, Felice Dell'Orletta, Giulia Venturi & Simonetta Montemagni. 2010. A contrastive approach to multi-word term extraction from domain corpora. In *Proceedings of the 7th international conference on language resources and evaluation*, .

Bordea, Georgeta, Paul Buitelaar, Stefano Faralli & Roberto Navigli. 2015a. Semeval-2015 task 17: Taxonomy Extraction Evaluation. In *Proceedings of the 9th international workshop on semantic evaluation*, .

Bordea, Georgeta, Paul Buitelaar, Stefano Faralli & Roberto Navigli. 2015b. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 902–910. Denver, Colorado. http://www.aclweb.org/anthology/S15-2151.

Bordea, Georgeta, Els Lefever & Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th international workshop on semantic evaluation*, .

Botros, R., K. Irie, M. Sundermeyer & H. Ney. 2015. On efficient training of word classes and their application to recurrent neural network language models. In *Proc. interspeech*, 1443–1447. https://www-i6.informatik.rwth-aachen.de/publications/download/985/Botros--2015.pdf.

Bouma, Gerlof. ???? Normalized (pointwise) mutual information in collocation extraction .

Bourigault, Didier, Isabelle Gonzalez-Mullier & Cécile Gros. 1996. Lexter, a natural language processing tool for terminology extraction. In *Proceedings of the 7th euralex international congress*, 771–779.

Brockelmann, Carl. 1908. *Grundriss der vergleichenden grammatik der semitischen sprachen*, vol. 1. Reuther & Reichard.

Brown, Peter E., Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer. 1993a. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2). 263–311. http://aclweb.org/anthology/J/J93/J93-2003.pdf.

Brown, Peter F, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer & Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics* 16(2). 79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer & Surya Mohanty. 1993b. But dictionaries are data too. In *Proceedings of the workshop on human language technology* HLT '93, 202–205. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1075671.1075716. http://dx.doi.org/10.3115/1075671.1075716.

Brown, Peter F, Vincent J Della Pietra, Stephen A Della Pietra & Robert L Mercer. 1993c. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19(2). 263–311.

Brugmann, Karl. 1886. *Vergleichende grammatik der indogermanischen sprachen*. Walter de Gruyter.

Busche, Hubertus. 2009. *Gottfried wilhelm leibniz: Monadologie*, vol. 34. Walter de Gruyter.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz & Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the second workshop on statistical machine translation*, 136–158. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki & Omar F Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the joint fifth workshop on statistical machine translation and metricsmatr*, 17–53. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the seventh workshop on statistical machine translation*, 10–51. Montréal, Canada: Association for Computational Linguistics. http://www.aclweb.org/anthology/W12-3102.

Callison-Burch, Chris, Philipp Koehn, Christof Monz & Omar F Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the sixth workshop on statistical machine translation*, 22–64. Association for Computational Linguistics.

Callison-Burch, Chris, Miles Osborne & Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *Eacl*, vol. 6, 249–256.

Candito, Marie & Djamé Seddah. 2010. Parsing word clusters. In *Proceedings of the naacl hlt 2010 first workshop on statistical parsing of morphologically-rich languages*, 76–84. Los Angeles, CA, USA. `http://www.aclweb.org/anthology/W10-1409`.

Cap, Fabienne, Alexander Fraser, Marion Weller & Aoife Cahill. 2014. How to produce unseen teddy bears: Improved morphological processing of compounds in smt. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics*, 579–587. Gothenburg, Sweden: Association for Computational Linguistics. `http://www.aclweb.org/anthology/E14-1061`.

Caraballo, Sharon A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* ACL '99, 120–126. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/1034678.1034705. `http://dx.doi.org/10.3115/1034678.1034705`.

Caraballo, Sharon Ann. 2001. *Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text*. Providence, RI, USA: dissertation. AAI3006696.

Ceesay, Bamfa & Wen Juan Hou. 2015. Ntnu: An unsupervised knowledge approach for taxonomy extraction. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 938–943. Denver, Colorado. `http://www.aclweb.org/anthology/S15-2156`.

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli & Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the eleventh international workshop on spoken language translation (iwslt), lake tahoe, ca*, 2–17.

Chahuneau, Victor, Eva Schlinger, Noah A. Smith & Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 conference on empirical methods in natural language processing (emnlp)*, 1677–1687. Seattle, WA, USA. `http://www.aclweb.org/anthology/D13-1174`.

Chang, Pi-Chuan, Michel Galley & Christopher D Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, 224–232. Association for Computational Linguistics.

Chang, Pi-Chuan, Dan Jurafsky & Christopher D Manning. 2009. Disambiguating de for chinese-english machine translation. In *Proceedings of the fourth workshop on statistical machine translation*, 215–223. Association for Computational Linguistics.

Chelba, Ciprian, Johan Schalkwyk & Michiel Bacchiani. 2010. Challenges in automatic speech recognition. In *Interspeech 2010*, ISCA Student panel presentation slides.

Chelba, Ciprian, Peng Xu, Fernando Pereira & Thomas Richardson. 2012. Distributed acoustic modeling with back-off n-grams. In *Acoustics, speech and signal processing (icassp), 2012 ieee international conference on*, 4129–4132. IEEE.

Chen, Stanley F & Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on association for computational linguistics*, 310–318. Association for Computational Linguistics.

Chen, Stanley F & Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13(4). 359–393. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.4604`.

Chen, Tianqi & Carlos Guestrin. 2015. Xgboost: Reliable large-scale tree boosting system .

Chen, Tianqi & Tong He. 2015. xgboost: extreme gradient boosting. *R package version 0.4-2* .

Cherny, Julius. 2000. Translation system and method in which words are translated by a specialized dictionary and then a general dictionary. US Patent 6,085,162.

Cherry, Colin. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt)*, 22–31. Atlanta, GA, USA. `http://www.aclweb.org/anthology/N13-1003.pdf`.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 263–270. Association for Computational Linguistics.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, Doha, Qatar.

Cholakov, Kostadin & Valia Kordoni. 2014. Better statistical machine translation through linguistic treatment of phrasal verbs. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 196–201. Doha, Qatar: Association for Computational Linguistics. http://www.aclweb.org/anthology/D14-1024.

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1). 22–29.

Clark, Alexander. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th conference of the european chapter of the association for computational linguistics (eacl-2003)*, 59–66. http://aclweb.org/anthology/E/E03/E03-1009.pdf.

Cleuziou, Guillaume, Davide Buscaldi, Gaël Dias, Vincent Levorato & Christine Largeron. 2015. Qassit: A pretopological framework for the automatic construction of lexical taxonomies from raw texts. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 955–959. Denver, Colorado. http://www.aclweb.org/anthology/S15-2159.

ClientSideNews. 2006. A new mid-level solution for terminology management .

Crego, Josep, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng et al. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540* .

Daiber, Joachim & Khalil Sima'an. 2015. Machine translation with source-predicted target morphology. *Proceedings of MT Summit XV, Miami, Florida* .

Daille, Beatrice. 1996. Study and implementation of combined techniques for automatic extraction of terminology 49–66.

Degtyarenko, Kirill, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan Mcnaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj & Michael Ashburner. 2008. ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. *Nucleic acids research* 36(suppl 1). D344–D350.

Dehdari, J., L. Tan & J. van Genabith. 2016a. BIRA: Improved predictive exchange word clustering. In *Proc. naacl*, San Diego, CA, USA. http://www.aclweb.org/anthology/N/N16.

Dehdari, Jon, Liling Tan & Josef van Genabith. 2016b. Bira: Improved predictive exchange word clustering. In *Proceedings of the 2016 conference of the north american chapter of the association for computational lin-*

guistics: Human language technologies, 1169–1174. San Diego, California: Association for Computational Linguistics. http://www.aclweb.org/anthology/N16-1139.

Dehdari, Jon, Liling Tan & Josef van Genabith. 2016c. Scaling up word clustering. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations*, 42–46. San Diego, California. http://www.aclweb.org/anthology/N16-3009.

Denkowski, Michael & Alon Lavie. 2010. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Proceedings of the HLT: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 250–253. Los Angeles, California.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on human language technology research*, 138–145. Morgan Kaufmann Publishers Inc.

Durrani, Nadir, Philipp Koehn, Helmut Schmid & Alexander Fraser. 2014. Investigating the usefulness of generalized word representations in SMT. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers*, 421–432. Dublin, Ireland. http://www.aclweb.org/anthology/C14-1041.

Dyer, Chris, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman & Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the association for computational linguistics (acl)*, .

Eichler, Kathrin & Günter Neumann. 2010. Dfki keywe: Ranking keyphrases extracted from scientific articles. In *Proceedings of the 5th international workshop on semantic evaluation*, 150–153.

Eisner, Jason. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the 41st annual meeting on association for computational linguistics-volume 2*, 205–208. Association for Computational Linguistics.

Ellendorff, Tilia, Liling Tan, Giuseppe Rizzo, Francesca Frontini & Rodrigo Agerri. 2014. Biopener project. In *Come hack with opener! workshop hackathon presentation*, https://www.dropbox.com/s/s0xhx4461cv85yw/BiOpeNER_presentation.pdf.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive science* 14(2). 179–211.

Emerson, Guy, Liling Tan, Susanne Fertmann, Alexis Palmer & Michaela Regneri. 2014. SeedLing: Building and Using a Seed corpus for the Human Language Project. In *Proceedings of the 2014 workshop on the use of computational methods in the study of endangered languages*, 77–85.

Enkhsaikhan, Majigsuren, Wilson Wong, Wei Liu & Mark Reynolds. 2007. Measuring data-driven ontology changes using text mining. In *Proceedings of the sixth australasian conference on data mining and analytics-volume 70*, 39–46. Australian Computer Society, Inc.

Espinosa Anke, Luis, Horacio Saggion & Francesco Ronzano. 2015. Taln-upf: Taxonomy learning exploiting crf-based hypernym extraction on encyclopedic definitions. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 949–954. Denver, Colorado. `http://www.aclweb.org/anthology/S15-2158`.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy & Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons .

Felber, Helmut. 1984. *Terminology Manual*. International Information Centre for Terminology.

Fountain, Trevor & Mirella Lapata. 2012. Taxonomy Induction using Hierarchical Random Graphs. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 466–476.

Frantzi, Katerina, Sophia Ananiadou & Hideki Mima. 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries* 3(2). 115–130.

Frantzi, Katerina T, Sophia Ananiadou & Junichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *Research and advanced technology for digital libraries*, 585–604. Springer.

Friedman, Jerome H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Fu, Ruiji, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang & Ting Liu. 2014. Learning Semantic Hierarchies via Word Embeddings. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1199–1209.

Gabrilovich, Evgeniy & Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of twentieth international joint conference on artificial intelligence*, .

Galley, Michel, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang & Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics*, 961–968. Association for Computational Linguistics.

Galley, Michel, Mark Hopkins, Kevin Knight & Daniel Marcu. 2004. What's in a translation rule? In *Hlt-naacl 2004: Main proceedings*, 273–280. Boston, Massachusetts, USA: Association for Computational Linguistics. `http://research.microsoft.com/apps/pubs/default.aspx?id=151142`.

Galley, Michel & Christopher D Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the conference on empirical methods in natural language processing*, 848–856.

Ganitkevitch, Juri, Benjamin Van Durme & Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 758–764. Atlanta, Georgia. `http://www.aclweb.org/anthology/N13-1092`.

Gao, Qin & Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software engineering, testing, and quality assurance for natural language processing*, 49–57.

Gimenez, Jesus & Lluis Marquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of the second workshop on statistical machine translation*, 256–264. Stroudsburg, PA, USA.

Giménez, Jesús & Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics* (94). 77–86.

Girju, Roxana. 2003. Automatic Detection of Causal Relations for Question Answering. In *Proceedings of the acl 2003 workshop on multilingual summarization and question answering-volume 12*, 76–83.

Gómez Guinovart, Xavier & Alberto Simoes. 2009. Parallel corpus-based bilingual terminology extraction .

Gonzàlez, Meritxell, Alberto Barrón-Cedeño & Lluís Màrquez. 2014. Ipa and stout: Leveraging linguistic and source-based features for machine translation evaluation. In *Ninth workshop on statistical machine translation*, 8.

Goodman, J. 2001a. A bit of progress in language modeling, extended version. Tech. Rep. MSR-TR-2001-72 Microsoft Research. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.8929`.

Goodman, Joshua. 2001b. Classes for fast maximum entropy training. In *Proceedings of the ieee international conference on acoustics, speech, and signal processing (icassp '01)*, vol. 1, 561–564. doi:10.1109/ICASSP.2001.940893. http://arxiv.org/pdf/cs/0108006.

Graham, Yvette, Timothy Baldwin & Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 1183–1191. Denver, Colorado: Association for Computational Linguistics. http://www.aclweb.org/anthology/N15-1124.

Grefenstette, Gregory. 2015. Inriasac: Simple hypernym extraction methods. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 911–914. Denver, Colorado. http://www.aclweb.org/anthology/S15-2152.

Gruber, Thomas R. 1993. A translation approach to portable ontology specifications 199–220.

Gupta, Rohit, Constantin Orasan & Josef van Genabith. 2015a. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1066–1072. Lisbon, Portugal.

Gupta, Rohit, Constantin Orasan & Josef van Genabith. 2015b. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1066–1072. Lisbon, Portugal. http://aclweb.org/anthology/D15-1124.

Gurulingappa, Harsha, Roman Klinger, Martin Hofmann-Apitius & Juliane Fluck. 2010. An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature. In *2nd workshop on building and evaluating resources for biomedical text mining tuesday, 18 th march 2010*, 15.

Habash, Nizar & Ahmed Elkholy. 2008. Sepia: surface span extension to syntactic dependency precision-based mt evaluation. In *Proceedings of the nist metrics for machine translation workshop at the association for machine translation in the americas conference, amta-2008. waikiki, hi*, .

Haddow, B., M. Huck, A. Birch, N. Bogoychev & P. Koehn. 2015. The Edinburgh/JHU phrase-based machine translation systems for WMT 2015. In *Proc. wmt*, 126–133. http://aclweb.org/anthology/W15-3013.

Hanisch, Daniel, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer & Juliane Fluck. 2005. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics* 6(1). 1–9.

Hasegawa, Takaaki, Satoshi Sekine & Ralph Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, 415. Association for Computational Linguistics.

Heafield, Kenneth. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the EMNLP 2011 sixth workshop on statistical machine translation*, 187–197. Edinburgh, Scotland. `http://aclweb.org/anthology/W/W11/W11-2123.pdf`.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark & Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)*, 690–696. Sofia, Bulgaria. `http://aclweb.org/anthology/P/P13/P13-2121.pdf`.

Hearst, Marti A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th conference on computational linguistics-volume 2*, 539–545.

Hermjakob, Ulf, Kevin Knight & Hal Daumé III. 2008. Name translation in statistical machine translation - learning when to transliterate. In *Proceedings of acl-08: Hlt*, 389–397. Columbus, Ohio: Association for Computational Linguistics. `http://www.aclweb.org/anthology/P/P08/P08-1045`.

Hippisley, Andrew, David Cheng & Khurshid Ahmad. 2005. The head-modifier principle and multilingual term extraction. *Natural Language Engineering* 11(02). 129–157.

Hoang, Hieu. 2011. Improving statistical machine translation with linguistic information .

Hoang, Hieu & Philipp Koehn. 2008. Design of the moses decoder for statistical machine translation. In *Software engineering, testing, and quality assurance for natural language processing*, 58–65. Association for Computational Linguistics.

Hoang, Hieu & Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *In proceedings of the international workshop on spoken language translation (iwslt*, 152–159.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8). 1735–1780.

Hoffmeister, Ana. 2014. Terminology processes and quality assurance .

Huang, Liang, Kevin Knight & Aravind Joshi. 2006. A syntax-directed translator with extended domain of locality. In *Proceedings of the workshop on computationally hard problems and joint inference in speech and language processing* CHSLP '06, 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1631828.1631829.

Huang, Pingping & Baobao Chang. 2014. SSMT:A Machine Translation Evaluation View To Paragraph-to-Sentence Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 585–589. Dublin, Ireland.

Hutchins, W John. 2000. *Early years in machine translation: memoirs and biographies of pioneers*, vol. 97. John Benjamins Publishing.

Isozaki, Hideki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh & Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, 944–952.

Jones, K Sparck. 1979. Experiments in relevance weighting of search terms. *Information Processing & Management* 15(3). 133–144.

Junczys-Dowmunt, Marcin. 2012. Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *The Prague Bulletin of Mathematical Linguistics* 98. 63–74.

Jurgens, David & Mohammad Taher Pilehvar. 2016. Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, 1092–1102. San Diego, California. http://www.aclweb.org/anthology/S16-1169.

Jurgens, David, Mohammad Taher Pilehvar & Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)*, 17–26. Dublin, Ireland: Association for Computational Linguistics and Dublin City University. http://www.aclweb.org/anthology/S14-2003.

Jurgens, David, Mohammad Taher Pilehvar & Roberto Navigli. 2015. Cross level semantic similarity: an evaluation framework for universal measures of similarity. *Language Resources and Evaluation* 50(1). 5–33.

Justeson, John S & Slava M Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering* 1(01). 9–27.

Kageura, Kyo & Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology* 3(2). 259–289.

Kalchbrenner, Nal, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves & Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099* .

Khanh, Vo Ho Bao & Shun Ishizaki. ???? Japanese–vietnamese compound noun translation .

Khashabi, Daniel. 2013. On the Recursive Neural Networks for Relation Extraction and Entity Recognition. Tech. rep.

Kirchhoff, Katrin & Mei Yang. 2005. Improved language modeling for statistical machine translation. In *Proceedings of the acl workshop on building and using parallel texts*, 125–128. Association for Computational Linguistics.

Kitamura, Mihoko & Yuji Matsumoto. 1996. Automatic extraction of word sequence correspondences in parallel corpora. In *Proceedings of the 4th workshop on very large corpora*, .

Kneser, Reinhard & Hermann Ney. 1993. Improved clustering techniques for class-based statistical language modelling. In *Third european conference on speech communication and technology (eurospeech'93)*, 973–976. Berlin, Germany.

Kneser, Reinhard & Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, speech, and signal processing, 1995. icassp-95., 1995 international conference on*, vol. 1, 181–184. IEEE.

Knight, Kevin. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics* 25(4). 607–615.

Koehn, Philipp. 2004a. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Machine translation: From real users to research*, 115–124. Springer.

Koehn, Philipp. 2004b. *Proceedings of the 2004 conference on empirical methods in natural language processing* chap. Statistical Significance Tests for Machine Translation Evaluation.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of mt summit, vol. 5, pp. 79-86.*, .

Koehn, Philipp. 2009. *Statistical machine translation.* Cambridge University Press.

Koehn, Philipp, Barry Haddow, Philip Williams & Hieu Hoang. 2010. More linguistic annotation for statistical machine translation. In *Proceedings of the joint fifth workshop on statistical machine translation and metricsmatr* WMT '10, 115–120. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1868850.1868865.

Koehn, Philipp & Hieu Hoang. 2007a. Factored translation models. In *Emnlp-conll*, 868–876.

Koehn, Philipp & Hieu Hoang. 2007b. Factored translation models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (emnlp-conll)*, 868–876. Prague, Czech Republic. `http://www.aclweb.org/anthology/D/D07/D07-1091.pdf`.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin & Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the acl on interactive poster and demonstration sessions*, 177–180. `http://www.aclweb.org/anthology/P07-2045.pdf`.

Koehn, Philipp, Franz Josef Och & Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1*, 48–54.

Koehn, Philipp & Josh Schroeder. 2007a. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation* StatMT '07, 224–227. Stroudsburg, PA, USA. `http://dl.acm.org/citation.cfm?id=1626355.1626388`.

Koehn, Philipp & Josh Schroeder. 2007b. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, 224–227.

Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer & Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 1001–1012. Doha, Qatar. `http://www.aclweb.org/anthology/D14-1108`.

Koo, Terry, Xavier Carreras & Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of acl-08: Hlt*, 595–603. Columbus, OH, USA. `http://www.aclweb.org/anthology/P/P08/P08-1068`.

Kordoni, Valia & Iliana Simova. 2004. Multiword expressions in machine translation. In *Proceedings of lrec2014*, .

Kozareva, Zornitsa & Eduard Hovy. 2010. A Semi-Supervised Method to Learn and Construct Taxonomies using the Web. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, 1110–1118.

Kozareva, Zornitsa, Ellen Riloff & Eduard Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In *Proceedings of acl-08: Hlt*, 1048–1056. Columbus, Ohio. `http://www.aclweb.org/anthology/P/P08/P08-1119`.

Kudo, Taku, Kaoru Yamamoto & Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the conference on empirical methods in natural language processing*, 230–237.

Lambert, Patrik & Rafael Banchs. 2006. Grouping multi-word expressions according to part-of-speech in statistical machine translation. In *Proceedings of the workshop on multi-word-expressions in a multilingual context*, 9–16.

Lambert, Patrik & Nuria Castell. 2004. Alignment of parallel corpora exploiting asymmetrically aligned phrases. In *Proceedings of lrec 2004 workshop*, 26.

Lefever, Els. 2015. Lt3: A multi-modular approach to automatic taxonomy construction. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 944–948. Denver, Colorado. `http://www.aclweb.org/anthology/S15-2157`.

Lefever, Els. 2016. A hybrid approach to domain-independent taxonomy learning. *Applied Ontology* (Preprint). 1–24.

Lembersky, Gennadi, Noam Ordan & Shuly Wintner. 2012. Adapting translation models to translationese improves smt. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics*, 255–265. Association for Computational Linguistics.

Lenat, Douglas B. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38(11). 33–38.

Lenci, Alessandro & Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the first joint conference on lexical and computational semantics-volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation*, 75–79.

Levy, Omer, Yoav Goldberg & Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *Conll*, 171–180.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren NG Thornton, Jonathan Weese & Omar F Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the fourth workshop on statistical machine translation*, 135–139. Association for Computational Linguistics.

Liang, Percy. 2005. *Semi-supervised learning for natural language*. The Massachusetts Institute of Technology MA thesis. `https://cs.stanford.edu/~pliang/papers/meng-thesis.pdf`.

Lin, Chin-Yew & Franz Josef Och. 2004a. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd meeting of the association for computational linguistics (acl'04), main volume*, 605–612. Barcelona, Spain. doi:10.3115/1218955.1219032. `http://www.aclweb.org/anthology/P04-1077`.

Lin, Chin-Yew & Franz Josef Och. 2004b. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, Stroudsburg, PA, USA.

Lin, Dekang. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th international conference on computational linguistics-volume 2*, 768–774.

Lin, Dekang. 2004. A path-based transfer model for machine translation. In *Proceedings of the 20th international conference on computational linguistics*, 625. Association for Computational Linguistics.

Lin, Dekang & Patrick Pantel. 2001. Discovery of Inference Rules for Question-Answering. *Natural Language Engineering* 7(04). 343–360.

Lionbridge. 2010. Standardizing business terminology through localization management .

Lita, Lucian Vlad, Abraham Ittycheriah, Salim Roukos & Nanda Kambhatla. 2003. tRuEcasIng. In Erhard Hinrichs & Dan Roth (eds.), *Proceedings of the 41st annual meeting of the association for computational linguistics*, 152–159. `http://www.aclweb.org/anthology/P03-1020.pdf`.

Liu, Ding & Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, .

Liu, Ding & Daniel Gildea. 2008. Improved tree-to-string transducer for machine translation. In *Proceedings of the third workshop on statistical machine translation*, 62–69. Association for Computational Linguistics.

Lo, Chi-kiu, Anand Karthik Tumuluru & Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of the seventh workshop on statistical machine translation*, 243–252.

Lossio-Ventura, Juan Antonio, Clement Jonquet, Mathieu Roche & Maguelonne Teisseire. 2013. Combining c-value and keyword extraction methods for biomedical terms extraction. In *Lbm'2013: 5th international symposium on languages in biology and medicine*, .

Luong, Minh-Thang, Hieu Pham & Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

Manning, Christopher D. 2016. Computational linguistics and deep learning. *Computational Linguistics* .

Marcu, Daniel & William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the acl-02 conference on empirical methods in natural language processing-volume 10*, 133–139. Association for Computational Linguistics.

Martin, Sven, Jörg Liermann & Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. *Speech Communication* 24(1). 19–37. doi:10.1016/S0167-6393(97)00062-9. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.2354.

Menezes, Arul & Chris Quirk. 2005. Dependency treelet translation: The convergence of statistical and example-based machine-translation? In *Mt summit x*, 99.

Meng, Fandong, Deyi Xiong, Wenbin Jiang & Qun Liu. 2014. Modeling term translation for document-informed machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 546–556. Doha, Qatar.

Mi, Haitao, Liang Huang & Qun Liu. 2008. Forest-based translation. In *Proceedings of acl-08: Hlt*, 192–199. Columbus, Ohio: Association for Computational Linguistics. http://www.aclweb.org/anthology/P/P08/P08-1023.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Miller, George A. 1995a. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11). 39–41.

Miller, George A. 1995b. WordNet: A Lexical Database for English. *Commun. ACM* 38(11). 39–41. doi: 10.1145/219717.219748. `http://doi.acm.org/10.1145/219717.219748`.

Miller, Scott, Jethran Guinness & Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In Susan Dumais, Daniel Marcu & Salim Roukos (eds.), *Hlt-naacl 2004: Main proceedings*, 337–342. Boston, MA, USA. `https://www.aclweb.org/anthology/N/N04/N04-1043.pdf`.

Miñarro-Giménez, Jose Antonio, Johannes Hellrich & Stefan Schulz. 2015. Acquisition of character translation rules for supporting snomed ct localizations. *Studies in health technology and informatics* 210. 597.

Mitchell, Jeff & Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science* 34(8). 1388–1439.

Mnih, Andriy & Geoffrey Hinton. 2009. A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio & L. Bottou (eds.), *Advances in neural information processing systems 21 (nips)*, vol. 21, 1081–1088. `http://papers.nips.cc/paper/3583-a-scalable-hierarchical-distributed-language-model.pdf`.

Morin, Frederic & Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *Proceedings of the 10th international workshop on artificial intelligence and statistics (aistats 2005)*, vol. 5, 246–252. Society for Artificial Intelligence and Statistics. `http://zot2wpv.googlecode.com/svn/trunk/zot2wpv/papers/Hierarchical%20Probabilistic%20Neural%20Network%20Language%20Model.pdf`.

Müller, T. & H. Schütze. 2015. Robust morphological tagging with word representations. In *Proc. naacl*, 526–536. `http://www.aclweb.org/anthology/N15-1055`.

Musen, Mark A. 1992. Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research* 25(5). 435 – 467. doi:http://dx.doi.org/10.1016/0010-4809(92)90003-S. `http://www.sciencedirect.com/science/article/pii/001048099290003S`.

Nakazawa, Toshiaki, Hideya Mino, Isao Goto, Sadao Kurohashi & Eiichiro Sumita. 2014. Overview of the first workshop on Asian translation. In *Proceedings of the first workshop on asian translation (wat2014)*, .

Nakazawa, Toshiaki, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi & Eiichiro Sumita. 2015. Overview of the 2nd workshop on asian translation. In *Proceedings of the 2nd workshop on asian translation (wat2015)*, .

Nakazawa, Toshiaki, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi & Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on language resources and evaluation (lrec 2016)*, 2204–2208. Portorož, Slovenia: European Language Resources Association (ELRA).

Nanba, Hidetsugu, Hideaki Kamaya, Toshiyuki Takezawa, Manabu Okumura, Akihiro Shinmori & Hidekazu Tanigawa. 2009. Automatic translation of scholarly terms into patent terms. In *Proceedings of the 2nd international workshop on patent information retrieval*, 21–24. ACM.

Navigli, Roberto & Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250.

Navigli, Roberto, Paola Velardi & Stefano Faralli. 2011. A Graph-Based Algorithm for Inducing Lexical Taxonomies from Scratch. In *IJCAI 2011, proceedings of the 22nd international joint conference on artificial intelligence, barcelona, catalonia, spain, july 16-22, 2011*, 1872–1877.

Neubig, Graham. 2013. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proceedings of the acl demonstration track*, Sofia, Bulgaria.

Nooralahzadeh, Farhad, Caroline Brun & Claude Roux. 2014. Part of speech tagging for French social media data. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers*, 1764–1772. Dublin, Ireland: Dublin City University and Association for Computational Linguistics. `http://www.aclweb.org/anthology/C14-1166`.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy & James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artif. Intell.* 194. 151–175. doi:10.1016/j.artint.2012.03.006. `http://dx.doi.org/10.1016/j.artint.2012.03.006`.

Och, Franz-Josef. 1995. *Maximum-Likelihood-Schätzung von Wortkategorien mit Verfahren der kombinatorischen Optimierung*. Germany Friedrich-Alexander-Universität Erlangen-Nürnburg Bachelor's thesis (Studienarbeit).

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting on association for computational linguistics-volume 1*, 160–167. http://www.aclweb.org/anthology/P03-1021.pdf.

Och, Franz Josef, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin & Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In Daniel Marcu Susan Dumais & Salim Roukos (eds.), *Hlt-naacl 2004: Main proceedings*, 161–168. Boston, Massachusetts, USA: Association for Computational Linguistics.

Och, Franz Josef & Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th international conference on computational linguistics (coling)*, 1086–1090. Saarbrücken, Germany. https://www.aclweb.org/anthology/C/C00/C00-2163.pdf.

Och, Franz Josef & Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics, july 6-12, 2002, philadelphia, pa, USA.*, 295–302.

Och, Franz Josef & Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics* 29(1). 19–51.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider & Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 380–390. Atlanta, GA, USA. http://www.aclweb.org/anthology/N13-1039.

Pal, Santanu, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay & Andy Way. 2010. Handling named entities and compound verbs in phrase-based statistical machine translation. In *23rd international conference of computational linguistics (coling 2010), beijing, chaina*, 46–54.

Palermo, David S. & James J. Jenkins. 1964. *Word association norms*. University of Minnesota Press. http://www.jstor.org/stable/10.5749/j.cttttpz4.

Panchenko, Stefano Ruppert Eugen Remus Steffen Naets Hubert Fairon Cedrick Ponzetto Simone Paolo, Alexander Faralli & Chris Biemann. 2016. Taxi: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th international workshop on semantic evaluation*, Association for Computational Linguistics.

Pantel, Patrick & Deepak Ravichandran. 2004. Automatically Labeling Semantic Classes. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, .

Papineni, Kishore, Salim Roukos, Todd Ward & Wei jing Zhu. 2002a. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Park, Eunjeong L. & Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *Proceedings of the 26th annual conference on human and cognitive language technology*, Chuncheon, Korea.

Parker, Steven. 2008. Badger: A new machine translation metric 21–25.

Paul, Michael. 2006. Overview of the iwslt 2006 evaluation campaign. In *In proceedings of the international workshop on spoken language translation*, 1–15.

Pease, Adam, Ian Niles & John Li. 2002a. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working notes of the aaai-2002 workshop on ontologies and the semanticweb*, .

Pease, Adam, Ian Niles & John Li. 2002b. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *In working notes of the aaai-2002 workshop on ontologies and the semantic web*, .

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014a. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 1532–1543. Doha, Qatar. `http://www.aclweb.org/anthology/D14-1162`.

Pennington, Jeffrey, Richard Socher & Christopher D Manning. 2014b. Glove: Global vectors for word representation. In *Emnlp*, vol. 14, 1532–1543.

Pinkham, Jessie E & Martine MSA Smets. 2008. Machine translation using learned word associations without referring to a multi-lingual human authored dictionary of content words. US Patent 7,356,457.

Pocostales, Joel. 2016. Nuig-unlp: A simple word embedding-based approach for taxonomy extraction. In *Proceedings of the 10th international workshop on semantic evaluation*, .

Popel, Martin & Zdeněk Žabokrtskỳ. 2010. Tectomt: modular nlp framework. In *Advances in natural language processing*, 293–304. Springer.

Popović, Maja, Daniel Stein & Hermann Ney. 2006. Statistical machine translation of german compound words. In *Advances in natural language processing*, 616–624. Springer.

Porsiel, Jörg. 2008. Machine translation at volkswagen: a case study 19(8). 58.

Porsiel, Jörg. 2011. Machine translation at volkswagen .

Pu, Xiao, Laura Mascarell, Andrei Popescu-Belis, Mark Fishel, Ngoc-Quang Luong & Martin Volk. 2015. Leveraging compounds to improve noun phrase translation from chinese and german. In *Proceedings of the acl-ijcnlp 2015 student research workshop*, 8–15. Beijing, China: Association for Computational Linguistics. `http://www.aclweb.org/anthology/P15-3002`.

Rackow, Ulrike, Ido Dagan & Ulrike Schwall. 1992. Automatic translation of noun compounds. In *Proceedings of the 14th conference on computational linguistics-volume 4*, 1249–1253. Association for Computational Linguistics.

Ramanathan, Ananthakrishnan, Hansraj Choudhary, Avishek Ghosh & Pushpak Bhattacharyya. 2009. Case markers and morphology: addressing the crux of the fluency problem in english-hindi smt. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp: Volume 2-volume 2*, 800–808. Association for Computational Linguistics.

Ratinov, Lev & Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (conll-2009)*, 147–155. Boulder, CO, USA. `http://www.aclweb.org/anthology/W09-1119`.

Řehůřek, Radim & Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu & Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the workshop on multiword expressions: Identification, interpretation, disambiguation and applications* MWE '09, 47–54. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=1698239.1698249`.

Rios, Miguel, Wilker Aziz & Lucia Specia. 2012. UOW: Semantically Informed Text Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Main Conference and the Shared Task*, 673–678. Montréal, Canada.

Ritter, Alan, Sam Clark, Mausam & Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, 1524–1534. Edinburgh, Scotland. `http://www.aclweb.org/anthology/D11-1141.pdf`.

Ritter, Alan, Stephen Soderland & Oren Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. In *Aaai spring symposium: Learning by reading and learning to read*, 88–93.

Rutherford, Attapol & Nianwen Xue. 2014. Discovering implicit discourse relations through Brown cluster pair representation and coreference patterns. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics (eacl)*, 645–654. Gothenburg, Sweden. `http://www.aclweb.org/anthology/E14-1068`.

Sager, Juan C, David Dungworth, Peter F McDonald et al. 1980. *English special languages: principles and practice in science and technology*. John Benjamins Pub Co.

Sánchez Cartagena, Víctor Manuel, Felipe Sánchez Martínez, Juan Antonio Pérez Ortiz et al. 2011. Integrating shallow-transfer rules into phrase-based statistical machine translation, Machine Translation Summit.

Santus, Enrico, Alessandro Lenci, Qin Lu & Sabine Schulte im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th conference of the european chapter of the association for* *computational linguistics, volume 2: Short papers*, 38–42. Gothenburg, Sweden. http://www.aclweb.org/anthology/E14-4008.

Šarić, Frane, Goran Glavaš, Mladen Karan, Jan Šnajder & Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of the first joint conference on lexical and computational semantics-volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation*, 441–448. Association for Computational Linguistics.

Sarmento, Luís, Paula Carvalho & Eugénio Oliveira. 2009. Exploring the Vector Space Model for Finding Verb Synonyms in Portuguese. In *Proceedings of the recent advancement in nlp (ranlp-2009)*, 393–398.

Saxe, Andrew M., James L. McClelland & Surya Ganguli. 2013. Learning Hierarchical Category Structure in Deep Neural Networks 1271–1276.

Schwenk, Holger, Anthony Rousseau & Mohammed Attik. 2012. Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proceedings of the naacl-hlt 2012 workshop: Will we ever really replace the n-gram model? on the future of language modeling for hlt*, 11–19. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow & Alexandra Birch. 2015a. Neural machine translation of rare words with subword units. *CoRR* abs/1508.07909. http://arxiv.org/abs/1508.07909.

Sennrich, Rico, Barry Haddow & Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* .

Shannon, Claude Elwood. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1). 3–55.

Shazeer, Noam, Ryan Doherty, Colin Evans & Chris Waterson. 2016. Swivel: Improving embeddings by noticing what's missing. *arXiv preprint arXiv:1602.02215* .

Shen, Libin, Jinxi Xu & Ralph M Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Acl*, 577–585.

Shieber, Stuart M. 2007. Probabilistic synchronous tree-adjoining grammars for machine translation: The argument from bilingual dictionaries. In *Proceedings of the naacl-hlt 2007/amta workshop on syntax and structure in statistical translation*, 88–95. Association for Computational Linguistics.

Shieber, Stuart M & Yves Schabes. 1990. Synchronous tree-adjoining grammars. In *Proceedings of the 13th conference on computational linguistics-volume 3*, 253–258. Association for Computational Linguistics.

Shwartz, V. & I. Dagan. 2016. Path-based vs. Distributional Information in Recognizing Lexical Semantic Relations. *ArXiv e-prints* .

Shwartz, Vered, Yoav Goldberg & Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2389–2398. Berlin, Germany: Association for Computational Linguistics. `http://www.aclweb.org/anthology/P16-1226`.

Simova, Iliana & Valia Kordoni. 2013. Improving english-bulgarian statistical machine translation by phrasal verb treatment. In *Proceedings of mt summit xiv workshop on multi-word units in machine translation and translation technology*, Nice, France.

Skadiņš, Raivis, Mārcis Pinnis, Tatiana Gornostay & Andrejs Vasiļjevs. 2013. Application of online terminology services in statistical machine translation 281–286.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the americas*, .

Snow, Rion, Daniel Jurafsky & Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17* .

Snow, Rion, Daniel Jurafsky & Andrew Y Ng. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics*, 801–808.

Socher, Richard, Eric H Huang, Jeffrey Pennin, Christopher D Manning & Andrew Y Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in neural information processing systems*, 801–809.

Stanojevic, Miloš & Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the ninth workshop on statistical machine translation*, 414–419.

Stanojević, Miloš, Amir Kamran, Philipp Koehn & Ondřej Bojar. 2015. Results of the wmt15 metrics shared task. In *Proceedings of the tenth workshop on statistical machine translation*, 256–273. Lisbon, Portugal: Association for Computational Linguistics. `http://aclweb.org/anthology/W15-3031`.

Stymne, Sara. 2008. German compounds in factored statistical machine translation. In *Advances in natural language processing*, 464–475. Springer.

Stymne, Sara. 2012. Clustered word classes for preordering in statistical machine translation. In *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*, 28–34. Avignon, France. `http://www.aclweb.org/anthology/W12-0704.pdf`.

Stymne, Sara, Nicola Cancedda & Lars Ahrenberg. 2013. Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics* 39(4). 1067–1108.

Suchanek, Fabian M., Gjergji Kasneci & Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th international conference on world wide web* WWW '07, 697–706. New York, NY, USA: ACM. doi:10.1145/1242572.1242667. `http://doi.acm.org/10.1145/1242572.1242667`.

Sultan, Md Arafat, Steven Bethard & Tamara Sumner. 2014a. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics* 2. 219–230.

Sultan, Md Arafat, Steven Bethard & Tamara Sumner. 2014b. Dls@ cu: Sentence similarity from word alignment. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)*, 241–246.

Sultan, Md Arafat, Steven Bethard & Tamara Sumner. 2015a. Dls@ cu: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th international workshop on semantic evaluation*, 148–153.

Sultan, Md Arafat, Steven Bethard & Tamara Sumner. 2015b. Dls@cu: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 148–153. Denver, Colorado. `http://www.aclweb.org/anthology/S15-2027`.

Täckström, Oscar, Ryan McDonald & Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 477–487. Montréal, Canada. `http://www.aclweb.org/anthology/N12-1052`.

Tai, Kai Sheng, Richard Socher & Christopher D. Manning. 2015a. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd annual meeting of the associ-*

ation for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), 1556–1566. Beijing, China. `http://www.aclweb.org/anthology/P15-1150`.

Tai, Kai Sheng, Richard Socher & Christopher D. Manning. 2015b. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, 1556–1566. Beijing, China. `http://www.aclweb.org/anthology/P15-1150`.

Tan, Liling. 2013. *Examining crosslingual word sense disambiguation*. Nanyang Technological University. pages 17-21 MA thesis.

Tan, Liling. 2014. Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]. https://github.com/alvations/pywsd.

Tan, Liling. 2016a. D4. 2: Terminology and ontology. In *Technical report, expert (exploiting empirical approaches to translation) consortium*, .

Tan, Liling. 2016b. Expert innovations in terminology extraction and ontology induction. In *Proceedings of the expert scientific and technological workshop*, .

Tan, Liling. 2016c. Faster and lighter phrase-based machine translation baseline. In *Proceedings of the 3rd workshop on asian translation (wat2016)*, 184–193. Osaka, Japan.

Tan, Liling. 2016d. Faster and lighter phrase-based machine translation baseline. In *Proceedings of the 3rd workshop on asian translation (wat2016)*, 184–193. Osaka, Japan: The COLING 2016 Organizing Committee. `http://aclweb.org/anthology/W16-4618`.

Tan, Liling. 2016e. Usaar: Hyponym endocentricity. In *Proceedings of the 10th international workshop on semantic evaluation (semeval 2016)*, .

Tan, Liling, Josef van Genabith & Francis Bond. 2015a. Passive and pervasive use of bilingual dictionary in statistical machine translation. In *Proceedings of the fourth workshop on hybrid approaches to translation (hytra)*, 30–34. Beijing. `http://www.aclweb.org/anthology/W15-4105`.

Tan, Liling, Rohit Gupta & Josef van Genabith. 2015b. Usaar-wlv: Hypernym generation with deep neural nets. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 932–937. Denver, Colorado. `http://www.aclweb.org/anthology/S15-2155`.

Tan, Liling & Noam Ordan. 2015. Usaar-chronos: Crawling the web for temporal annotations. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 846–850. Denver, Colorado. http://www.aclweb.org/anthology/S15-2143.

Tan, Liling & Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, 201–206. Baltimore, Maryland, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W14-3323.

Tan, Liling, Carolina Scarton, Lucia Specia & Josef van Genabith. 2015c. Usaar-sheffield: Semantic textual similarity with deep regression and machine translation evaluation metrics. In *Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)*, 85–89. Denver, Colorado. http://www.aclweb.org/anthology/S15-2015.

Tan, Liling, Carolina Scarton, Lucia Specia & Josef van Genabith. 2016. Saarsheff at semeval-2016 task 1: Semantic textual similarity with machine translation evaluation metrics and (extreme) boosted tree ensembles. In *Proceedings of the 10th international workshop on semantic evaluation (semeval 2016)*, San Diego, California.

Tan, Liling, Anne Schumann, Jose Martinez & Francis Bond. 2014. Sensible: L2 translation assistance by emulating the manual post-editing process. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)*, 541–545. Dublin, Ireland. http://www.aclweb.org/anthology/S14-2094.

Tanaka, Takaaki & Timothy Baldwin. 2003. Noun-noun compound machine translation: a feasibility study on shallow processing. In *Proceedings of the acl 2003 workshop on multiword expressions: analysis, acquisition and treatment-volume 18*, 17–24. Association for Computational Linguistics.

Tezcan, Arda & Vincent Vandeghinste. 2011. Smt-cat integration in a technical domain: Handling xml markup using pre & post-processing methods. *Proceedings of EAMT 2011* .

Tillmann, Christoph. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of hlt-naacl 2004: Short papers*, 101–104.

Tran, Ke M., Arianna Bisazza & Christof Monz. 2014. Word translation prediction for morphologically rich languages with bilingual neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 1676–1688. Doha, Qatar: Association for Computational Linguistics. http://www.aclweb.org/anthology/D14-1175.

Tsuji, Keita & Kyo Kageura. 2006. Automatic generation of japanese–english bilingual thesauri based on bilingual corpora. *Journal of the American Society for Information Science and Technology* 57(7). 891–906.

Tsvetkov, Yulia & Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering* 18(04). 549–573.

Tuan, Luu Anh, Jung-jae Kim & Kiong See Ng. 2014. Taxonomy construction using syntactic contextual evidence. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 810–819.

Turian, Joseph, Lev-Arie Ratinov & Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 384–394. Uppsala, Sweden. `http://www.aclweb.org/anthology/P10-1040`.

Ueffing, Nicola, Gholamreza Haffari & Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, 25–32. Prague, Czech Republic: Association for Computational Linguistics. `http://www.aclweb.org/anthology/P07-1004`.

Uszkoreit, Jakob & Thorsten Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *Proceedings of acl-08: Hlt*, 755–762. Columbus, OH, USA. `http://www.aclweb.org/anthology/P/P08/P08-1086.pdf`.

Van Der Plas, Lonneke. 2005. Automatic acquisition of lexico-semantic knowledge for qa. In *Proceedings of the ijcnlp workshop on ontologies and lexical resources*, 76–84.

Čmejrek, Martin, Jan Cuřín & Jiří Havelka. 2003. Czech-english dependency-based machine translation. In *Proceedings of the tenth conference on european chapter of the association for computational linguistics - volume 1* EACL '03, 83–90. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/1067807.1067820. `http://dx.doi.org/10.3115/1067807.1067820`.

Vela, Mihaela & Liling Tan. 2015a. Predicting machine translation adequacy with document embeddings. In *Proceedings of the tenth workshop on statistical machine translation*, 402–410. Lisbon, Portugal. `http://aclweb.org/anthology/W15-3051`.

Vela, Mihaela & Liling Tan. 2015b. Predicting machine translation adequacy with document embeddings. In *Proceedings of the tenth workshop on statistical machine translation*, 402–410. Lisbon, Portugal. `http://aclweb.org/anthology/W15-3051`.

Velardi, Paola, Stefano Faralli & Roberto Navigli. 2013. OntoLearn Reloaded: A Graph-based Algorithm for Taxonomy Induction. *Computational Linguistics* 39(3). 665–707.

Vogel, Stephan & Christian Monson. 2004. Augmenting manual dictionaries for statistical machine translation systems. In *Proceedings of lrec 2004*, 1593–1596.

Vogel, Stephan, Hermann Ney & Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on computational linguistics-volume 2*, 836–841. Association for Computational Linguistics.

Wan, Stephen, Mark Dras, Robert Dale & Cécile Paris. 2006. Using dependency-based features to take the "para-farce" out of paraphrase. In *Proceedings of the australasian language technology workshop*, vol. 2006, .

Wang, Chao, Michael Collins & Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Emnlp-conll*, 737–745.

Wang, Wei, Kevin Knight & Daniel Marcu. 2006. Capitalizing machine translation. In *Proceedings of the human language technology conference of the naacl, main conference*, 1–8. New York City, USA. `http://www.aclweb.org/anthology/N/N06/N06-1001`.

Warburton, Kara. 2005. Terminology: Getting down to business, the globalization insider .

Weaver, Warren. 1955. Translation. *Machine translation of languages* 14. 15–23.

Weese, Jonathan, Juri Ganitkevitch, Chris Callison-Burch, Matt Post & Adam Lopez. 2011. Joshua 3.0: Syntax-based machine translation with the thrax grammar extractor. In *Proceedings of the sixth workshop on statistical machine translation*, 478–484. Association for Computational Linguistics.

Weischedel, Ralph, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Mitchell Marcus, Ann Taylor et al. 2010. Ontonotes release 4.0. *Linguistic Data Consortium, Philadelphia* .

Weller, Marion, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde & Alexander Fraser. 2014. Distinguishing degrees of compositionality in compound splitting for statistical machine translation. In *Proceedings of the first workshop on computational approaches to compound analysis (comacoma 2014)*, 81–90. Dublin, Ireland: Association for Computational Linguistics and Dublin City University. `http://www.aclweb.org/anthology/W14-5709`.

Whittaker, Edward W. D. & Bhiksha Raj. 2001. Quantization-based language model compression. In *Proceedings of interspeech*, 33–36.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics* 23(3). 377–403.

Wu, Dekai & Hongsing Wong Hkust. 1998. Machine translation with a stochastic grammatical channel. In *Proceedings of the 17th international conference on computational linguistics-volume 2*, 1408–1415. Association for Computational Linguistics.

Wu, Hua & Haifeng Wang. 2007. Comparative study of word alignment heuristics and phrase-based smt .

Wu, Hua, Haifeng Wang & Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd international conference on computational linguistics - volume 1* COLING '08, 993–1000. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=1599081.1599206`.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* .

Wuebker, Joern, Stephan Peitz, Felix Rietig & Hermann Ney. 2013. Improving statistical machine translation with word class models. In *Proceedings of the 2013 conference on empirical methods in natural language processing (emnlp)*, 1377–1381. Seattle, WA, USA. `http://www.aclweb.org/anthology/D13-1138.pdf`.

Wüster, Eugen. 1969. Die vier Dimensionen der Terminologiearbeit. *Mitteilungsblatt für Dolmetscher und Übersetzer* 2(15). 1–6.

Xiong, Deyi, Qun Liu & Shouxun Lin. 2007. A dependency treelet string correspondence model for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, 40–47. Association for Computational Linguistics.

Yamada, Kenji & Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th annual meeting on association for computational linguistics*, 523–530. Association for Computational Linguistics.

Žabokrtskỳ, Zdeněk, Jan Ptáček & Petr Pajas. 2008. Tectomt: Highly modular mt system with tectogrammatics used as transfer layer. In *Proceedings of the third workshop on statistical machine translation*, 167–170. Association for Computational Linguistics.

Zens, Richard, Franz Josef Och & Hermann Ney. 2002. Phrase-based statistical machine translation. In *Ki 2002: Advances in artificial intelligence*, 18–32. Springer.

Zhang, Hao & Daniel Gildea. 2007. Factorization of synchronous context-free grammars in linear time. In *Proceedings of the naacl-hlt 2007/amta workshop on syntax and structure in statistical translation*, 25–32.

Zhang, Hao, Liang Huang, Daniel Gildea & Kevin Knight. 2006. Synchronous binarization for machine translation. In *Proceedings of the main conference on human language technology conference of the north american chapter of the association of computational linguistics*, 256–263. Association for Computational Linguistics.

Zhang, Kaizhong & Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing* 18(6). 1245–1262.

Zhang, Min, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan & Sheng Li. 2008a. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of acl-08: Hlt*, 559–567. Columbus, Ohio: Association for Computational Linguistics. http://www.aclweb.org/anthology/P/P08/P08-1064.

Zhang, Ying, Stephan Vogel & Alex Waibel. 2004a. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of lrec 2004*, .

Zhang, Yongzheng, Evangelos Milios & Nur Zincir-Heywood. 2004b. A comparison of keyword-and keyterm-based methods for automatic web site summarization. In *Aaai04 workshop on adaptive text extraction and mining*, .

Zhang, Ziqi, Jose Iria & Christopher Brewster. 2008b. A comparative evaluation of term recognition algorithms. In *Sixth international conference on language resources and evaluation*, .

Zhao, Hai, Wenliang Chen, Chunyu Kity & Guodong Zhou. 2009. Multilingual dependency learning: A huge feature engineering method to semantic dependency parsing. In *Proceedings of the thirteenth conference on computational natural language learning (conll 2009): Shared task*, 55–60. Boulder, CO, USA. http://www.aclweb.org/anthology/W09-1208.

Zollmann, Andreas, Ashish Venugopal, Franz Och & Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of the 22nd international conference on computational linguistics-volume 1*, 1145–1152.

Zollmann, Andreas & Stephan Vogel. 2011. A word-class approach to labeling PSCFG rules for machine translation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (acl-hlt)*, 1–11. Portland, OR, USA. `http://www.aclweb.org/anthology/P11-1001.pdf`.