DEEP SYNTAX IN STATISTICAL MACHINE TRANSLATION

Yvette Graham
Dublin City University

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This work investigates the current performance capabilities of LFG f-structure based transfer machine translation. Our empirical evaluation compares transfer-based machine translation performance to state of the art machine translation. Our investigation reveals that although the LFG-based approach under-performs compared to state of the art method in general, when the evaluation is restricted to translations where the target language f-structure falls within coverage of the generation grammar, the LFG-based system can in fact achieve higher coverage of unseen data in addition to improvements in translation quality.

1 Introduction

Essentially, machine translation (MT) systems need to accomplish two things: translate the source language (SL) word into the target language (TL) and produce these words in the correct order for the TL (Koehn, 2009). Approaches to MT use different levels of linguistic analysis for translation and divide the tasks involved in the translation of words and word order between analysis and generation components and a transfer component. The shallowest approach translates a SL surface form sentence directly into the TL, assigning the tasks of translating both words and word order to the transfer component, as in Phrase-Based Statistical Machine Translation (PB-SMT) (Koehn et al., 2003) for example. At a slightly deeper level of analysis, such as Phrase-Based Factored Models (Koehn and Hoang, 2007), transfer involves translating the lemma form, morpho-syntactic information and word order to the TL. Deep syntactic analysis goes a level deeper and transfer now involves translating SL syntactic representations such as dependency relations, lemmas and morpho-syntactic information to the TL. Even deeper again we have semantic analysis, with transfer translating between SL and TL context and meaning representations, relations, roles and (possibly) morpho-syntactic information. Finally, an interlingual analysis assigns the entire translation task to the analysis and generation components, with no transfer required, since the representation itself is entirely language independent.

Although increasing the depth of analysis can potentially *decrease the difficulty of translation*, there is the inevitable trade-off as a deeper analysis *increases the difficulty of analysis and generation*. In addition, when we divide the task of translation into separate components in a pipeline architecture, we need to consider how well each step in the pipeline fits together. The output of the parser used for analysis must be the input expected by the transfer decoder, and likewise the transfer decoder output must provide good input for generation. In addition, the use of parsers and generators to a deep level of analysis can also restrict the number of translation hypotheses reached by the search. For example, if generation is only

[†]This work was partly funded by a Science Foundation Ireland PhD studentship P07077-60101.

possible on the sentence level, as opposed to the word level, significantly more pruning of translation options may be necessary.

2 Deep Syntax for Transfer

Deep syntax, such as the Lexical Functional Grammar f-structure (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001), has been used in transfer-based machine translation (Riezler and Maxwell, 2006; Bojar and Hajič, 2008; Graham et al., 2009) as it provides a good level of linguistic analysis for machine translation, for several reasons:

- The reordering model, required by PB-SMT and shallow-syntax approaches is one of the most challenging models to devise (Koehn, 2009; Chen et al., 2006) and is not required for deep syntactic transfer. Source language (SL) word order is eliminated from the translation process since translation happens at the deep syntax level, abstracting away from surface form word order differences.

- The number of nodes in a deep syntactic representation is linear in sentence length, avoiding complexity problems encountered with shallow syntax based approaches (Deneefe et al., 2007; Deneefe and Knight., 2009).

- Non-terminals are allowed in transfer rules to map pieces of SL structure to the correct position in the TL but in a much more constrained way than in, for example, Hierarchical Models (Chiang, 2007) avoiding the severe pruning necessary for decoding in other parsing-based approaches (Li et al., 2009).

- Decoding can be carried out via a top-down application of contiguous transfer rules, so there are no gaps between TL words, eliminating the need for sophisticated heuristic language modeling techniques such as cube-pruning (Chiang, 2007), for example.

- Morpho-syntactic information for source and target sentences is present in deep syntactic representations, enabling the use of statistically richer Factored Models (Koehn and Hoang, 2007) also increasing coverage of inflections of lemmas not observed in bilingual training data.

This work focuses on investigating the current feasibility of deep syntactic transfer by comparing performance of two publicly available machine translation systems and to provide as meaningful a comparison as possible, we use two systems that are trained fully automatically. As such, we compare English translations of German text using the publicly available state-of-the-art phrase-based statistical machine translation (PB-SMT) system, Moses (Koehn et al., 2007) with translations produced by Sulis (Graham, 2010), a (also publicly available) transfer-based SMT system that uses the LFG f-structure as the intermediate representation for transfer. Although Sulis is in fact linguistic theory independent, with the only restriction being that input and output structures are deep syntax, the system was initially developed for LFG f-structure transfer, and therefore is fit for the purpose of our empirical comparison. In addition to investigating just how far off state-of-the-art performance the LFG f-structure transfer system currently is, we are also interested to know if the deep syntax SMT system produces the same kinds of translations as a PB-SMT system, examining one syntactic construction in particular, compound nouns. We investigate if for this particular syntactic construction, if the LFG-based system achieves state-of-the-art performance by providing a human evaluation of translations of a sample of compound nouns occurring in the test data.

3 LFG-based Transfer and PB-SMT Comparison

3.1 Experimental Set-up

German and English Europarl (Koehn, 2005) and Newswire sentences length 5-15 words were parsed using using LFG Grammars (Kaplan et al., 2004; Riezler et al., 2002), resulting in approx. 360K parsed sentence pairs, applying a disambiguation model to select the single best parse for each input. A trigram deep syntax language model was trained on the LFG-parsed English side of the Europarl corpus, with approximately 1.26M English f-structures (again using only the single-best parse) by extracting all unigram, bigram and trigrams from the f-structures before running SRILM (Stolcke, 2002). The surface-form language model, used after generation, consisted of the English side of the Europarl, also computed using SRILM. Word alignment was run on the training data yielding an alignment between local f-structures for each f-structure pair in the bilingual training data. All transfer rules consistent with this alignment were extracted. Minimum Error Rate Training (MERT) (Och, 2003) was carried out on 1000 development sentences for each configuration using Zmert (Zaidan, 2009).[1] Parsing and generation were carried out using XLE (Kaplan et al., 2002) and LFG Grammars (Kaplan et al., 2004; Riezler et al., 2002). We restrict our evaluation to short sentences and use the test set of Koehn et al. (2003), which includes 1755 German-English translations.

We compare the performance of a state-of-the-art PB-SMT system, Moses (Koehn et al., 2007), with the LFG f-structure transfer-based system (Graham, 2010). In our investigation, we examine if the LFG-based system produces the same kinds of translations as the Phrase-Based system, focusing on one specific syntactic construction, the compound noun (CN), to observe if, for this particular syntactic construction, the f-structure system can achieve state-of-the-art performance in a human evaluation of the first 100 CNs in the test data. The same data

[1] Settings for MERT training were as follows: beam=20, m=100, k=1, k-option=shortest. MERT was carried out separately for each method of word alignment. In all other experiments weights for the LFG-INT configuration were used.

	Bleu	Correct CNs	Fuzzy CNs	Precision Grammar Coverage
LFG	17.29 %	56 %	25 %	38%
PB	30.70 %	54 %	22 %	n/a

Table 1: LFG f-structure transfer and PB-SMT comparison

as in previous experiments was used for training and testing of both systems. For training the LFG-based system, we use technologies described in (Graham and van Genabith, 2008; Graham et al., 2009; Graham and van Genabith, 2009, 2010a,b). Configuration settings for the LFG-based system were as follows: word alignment – deep syntax intersection, no rule size limit , beam size of 100, m-best list size of 100 and non-deterministic generation (*allstrings* XLE option).[2]

3.1.1 Results

Table 1 contains automatic evaluation results for the f-structure transfer (LFG) system (17.29 Bleu) compared to the Phrase-Based (PB) system (30.7 Bleu) showing the degree to which the LFG-based system currently under-performs compared to state-of-the-art.[3] For CNs, however, the LFG-based system performs at least as well as the PB system by translating 56% CNs correctly and 25% in a way that contributes at least some correct meaning to the translation (labeled *fuzzy correct*), while the PB system translates 54% correctly and 22% as a fuzzy translation, in our human evaluation.[4]

Table 2 contains results for the 38% of translations that were within coverage of the precision grammar used for generation, showing the PB system (32.69% Bleu) outperforming the LFG-based system (27.85% Bleu), by almost 5 Bleu points absolute. Due to the possibility of (ngram-based) Bleu unfairly biasing in favor of the PB system, we include results for human-targeted Bleu, NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006, 2005) automatic evaluation metrics using reference translations produced by post-editing the first 150 translations from each MT system (Snover et al., 2006). Results for this evaluation show that the LFG system (73.12% Bleu) in fact outperforms the PB system (70.8%) by a little over 2 Bleu points absolute for translations within coverage of the precision grammar used for generation. We also include the number of untranslated words for the LFG-based system (2 words) and the PB system (34 words), showing that for translations in-coverage (by in-coverage we mean the input sentence achieves a full parse by the source language precision grammar)

[2] See http://www2.parc.com/isl/groups/nltt/xle/doc/xle.html for further details of available options with XLE.

[3] The unfair bias of ngram-based Bleu metric in favor of Moses should be noted, and is discussed later.

[4] It is worth noting that it is highly likely that the LFG-based system would not perform as well on a test set of unrestricted sentences length due lower parser coverage of long sentences.

	Bleu	HBleu	HNIST	HTER	HMETEOR	Untrans. Words
LFG	27.85	73.12	8.3602	20.74	82.80	2
PB	32.69	70.80	8.1710	23.63	86.00	34

Table 2: Precision grammar in-coverage comparison with state-of-the-art. Note: H-Bleu = human targeted Bleu for 150 post-edited reference translations (similar to HTER (Snover et al., 2006))

of the precision grammar, the LFG-based system also achieves higher coverage of unseen data.

3.1.2 Discussion

Automatic evaluation results for the entire test set suggest that the LFG-based system under-performs significantly in comparison with state-of-the-art (Table 1). However, the results are unfairly biased in favor of the PB system, due to a combination of the Bleu evaluation metric being ngram-based with legitimate syntactic variations in the LFG system output. The difference in results is, however, too large to claim that this is entirely due to this bias. Table 4 shows a random selection of translations produced by the LFG-based system from the entire test set.

Human evaluation of 100 CNs shows that the LFG system does in fact achieve state-of-the-art performance for this particular syntactic construction, however. Interestingly, the intersection of the CNs that both the LFG and PB systems translate correctly is quite small, with the LFG-based system correctly translating 30% of those not translated correctly by Moses, and Moses correctly translating 23% of those not translated correctly by the LFG-based system, suggesting the possibility of a hybrid MT system (similar to (Eisele et al., 2008; Chen et al., 2007; Eisele, 2005)) or that deep syntax parsing could be used to improve translation of CNs for PB-SMT. Table 3 shows a selection of CNs taken from the entire test set for the PB and LFG systems. The LFG system achieves coverage of CNs not observed in training data where component nouns were observed in training. For example, the CN, *Hafenpolitik*, was not observed in the German training data, but *Hafen* appears combined with other nouns a total of approximately 80 times and *politik* also appears in the German training data approximately 3,400 times combined with another noun. This CN is translated correctly by the LFG-based system but not the PB system.

For translations within coverage of the precision grammar, i.e. where the transfer decoder manages to produce a combination of lemmas, dependency relations and morpho-syntactic information in TL structures that do not clash with constraints during TL generation, human-targeted evaluation results show the LFG system achieves state-of-the-art performance for these translations, in addition to achieving higher translation coverage of unseen data, mainly due to its ability to learn how to translate new unseen CNs from CNs in the training data that con-

CN	PB Translation	LFG Translation
Wiederaufnahme	Resumption	
Tagesordnung	agenda	
Rechnungsführung	accounts	
Unternehmensneugründungen		company start-ups
Vorsichtsmassnahmen		measures precautionary*
Asien-Europa-Stiftung		Asia Europe Foundation
Osttimors		East Timor
ASEM-Gesprächen		ASEM talks
Hafenpolitik		port policy
Schwerpunkt	Emphasis	
Hauptsache	reason*	
Eigenkapital	capital*	invested capital*
Arbeitsrecht	labour law*	employment legislation*
Küstenstaaten		coastal states
Subsidiaritätsprinzip	principle of subsidiarity*	
Bewerberländer	candidate countries	applicant countries*
Parlamentswahlen	parliamentary elections*	general elections
Standpunkt	position*	question*
Ostsee	Baltic	
Änderungsantrag	Amendment	
Dioxinskandal	dioxin scare*	dioxin scandal
Einteilung	classification*	division
Futtermittelsicherheit		feed safety
Futtermittelkette		feed chain
Futtermitteln	feed*	means of feed*
Gemeinschaftsebene	Community level	Community scale*
Weltanschauung	World view*	world like mindedness*
weltweit	in the world*	worldwide*
Gemeinderatswahlen		elections local*
Richtlinien	directives*	directive*
Kernstück	heart*	lifeblood*
Ausnahmemöglichkeiten		opportunity for exceptions*
Änderungsanträgen	amendments	
Änderungsanträge		amendments
Vertragseinhaltung		Treaty compliance*
Entschliessungsantrags	resolution*	
Forschungsraum	research area	period of Research*
Endkontrolle		final*
Gegenprüfung		counter examination

Table 3: German Compound Noun translations for the Phrase-Based SMT system and the LFG-based system, translations evaluated as a fuzzy translation are marked with an asterisk

SRC:	Dies kann nicht hingenommen werden.
REF:	This is an unacceptable situation.
LFG:	Not one that can allow continue
SRC:	Herr Präsident! Die Sicherheit verschiedener Verkehrsarten steht ernsthaft auf dem Spiel.
REF:	Mr President, safety is a serious issue for various forms of transport.
LFG:	Mr President. Die of different forms of transport safety is at stake seriously.
SRC:	Das ist die politische Position.
REF:	That is the political position.
LFG:	That is the political position.
SRC:	Natürlich ist sich auch die türkische Gesellschaft dieses Gegensatzes bewusst.
REF:	Turkish society obviously perceives this contradictory attitude.
LFG:	Of course ist sich the Turkish society also of this contradiction bewusst
SRC:	Solche Gewalttätigkeit potenziert die Hassgefühle nur noch weiter.
REF:	That sort of violence only stirs up feelings of hatred.
LFG:	This violation potenzieren only hate emotions further

Table 4: Randomly selected translations, original reference translations provided (not human-targeted)

tain component nouns, in addition to achieving coverage of inflections of words not seen in bilingual training, since we use Factored Models (Koehn and Hoang, 2007). Table 5 shows a random selection of translations for the PB and LFG systems for translations in coverage of the precision generation grammar and Table 6 shows German words that were not translated by the LFG and PB systems for translations in coverage of the precision grammar.

4 Summary

Compared to state-of-the-art PB-SMT the LFG-based system under-performs, but for sentences in-coverage of the precision grammar used for generation, state-of-the-art performance and higher coverage of unseen data is achieved. Some practical challenges still need to be overcome before reaching state-of-the-art performance for all input. One challenge is parser coverage: depending on the parsing technologies used, coverage of long sentences can be low, resulting in a much smaller sized bilingual corpus used for training in comparison to a phrase-based system. A similar challenge occurs for generator coverage: technologies for generation from deep syntactic structures are usually tested on gold-standard input, and even with adaptation to allow more robust generation, generator coverage can still be low. In addition, even when generation succeeds, a fluent sentence of the target language is not guaranteed. LFG f-structures contain a large amount of information, such as dependency relations between words and morpho-syntactic features and in order for TL generation to produce good quality output, the particular combination of lemmas, dependency relations and morpho-syntactic information in the TL structure must comply with constraints within the generation grammar. If the TL structure contains morpho-syntactic and dependency information that clash when constraints are solved during generation, a fragment grammar can be used, but the quality of output severely deteriorates. Constructing TL deep syntactic structures that do not cause clashes in generation constraints remains a major challenge for f-structure transfer.

References

Banerjee, Satanjeev and Lavie, Alon. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *Proceedings of Workshop on "Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, at the 43rd Annual Meeting of the Association for Computational Linguistics"*, pages 65–73, Ann Arbor, Michigan.

Bojar, Ondřej and Hajič, Jan. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation at the 46th Annual Meeting of the Association for Computational Linguistics*, pages 143–146, Columbus, OH.

SRC:	Auf Gesetzesebene gibt es allgemeine Texte, in denen Diskriminierung weltweit verurteilt wird.
REF:	**Legally speaking, there are general texts condemning discrimination everywhere.**
LFG:	General texts condemning worldwide discrimination have been given to any legislative level.
PB:	There is general provisions on gesetzesebene where discrimination is condemned in the world.
SRC:	Das soll sich hier hoffentlich nicht wiederholen!
REF:	**I hope we will not see a repeat performance here!**
LFG:	hopefully that should not be repeated.
PB:	I hope it will not repeat here!
SRC:	Der BSE-Skandal war das schlechteste, bekannteste Beispiel.
REF:	**The BSE scandal was the worst and most notorious example.**
LFG:	The BSE scandal is the worst and most known case.
PB:	The BSE scandal was the worst and most well-known example.
SRC:	In Erwartung von mehr Klarheit haben wir uns deshalb der Stimme enthalten.
REF:	**Pending further clarification, we therefore abstain from the vote.**
LFG:	Therefore I abstained in expectation of greater clarity for.
PB:	In expectation of greater clarity, we have therefore abstain from voting.
SRC:	Der Wiederaufbau Osttimors ist noch im Gange.
REF:	**The rebuilding of East Timor is still an ongoing process.**
LFG:	The reconstruction of East Timor is still taking place.
PB:	The reconstruction osttimors is still in progress.
SRC:	Möchte sich jemand für diesen Antrag aussprechen?
REF:	**Is there a speaker to support this request?**
LFG:	Does anyone wish to speak in support of this motion?
PB:	Does anyone wish to speak in favor of this request?
SRC:	Vielen Dank für diese Klarstellung, Herr Kommissar.
REF:	**Thank you very much for that clarification, Commissioner.**
LFG:	I would like to thank the Commissioner for that clarification.
PB:	Thank you for that clarification, Commissioner.
SRC:	In diesem Punkt sind wir einer Meinung.
REF:	**On this point we are in agreement.**
LFG:	We will be agreement on point about this.
PB:	In this regard, we are in agreement.
SRC:	Gibt es Einwände?
REF:	**Are there any comments?**
LFG:	Are there any objections?
PB:	Are there any comments?
SRC:	Verhaltenskodex für Waffenausfuhren
REF:	**Arms trade code of conduct**
LFG:	Code of Conduct on Arms Exports
PB:	Code of conduct on arms exports

Table 5: Randomly selected sample of translations in-coverage of precision grammar, original reference translations provided.

Phrase-Based System	LFG-based system
interparlamentarischer	liegen
asien-europa-stiftung	vorsichtshalber
osttimors	
interparlamentarischer	
europäers	
zu	
lehnten	
spielzeugbomben	
erfahrenen	
kompetenzverteilung	
marktposition	
enttäuschte	
selbstbewertung	
gegenwert	
kostengünstiges	
bleibenden	
geldverkehrs	
reformpläne	
eindämmungsmassnahmen	
regem	
auslanLFGdiplomatie	
kompetenzabgrenzung	
planungssicherheit	
papua-führer	
dominiert	
ersuchten	
neuzuteilung	
eu-lärmindizes	
zusatzstoffes	
klimafrage	
vorsichtshalber	
sicherheitsspielraum	
un-flüchtlingshilfswerk	
gesamtgesellschaftlichen	

Table 6: German words not translated in translations within coverage of the TL generation precision grammar for the Phrase-Based and LFG-based systems

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Chen, Boxing, Cettolo, Mauro and Federico, Marcello. 2006. Reordering rules for phrase-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 53–58, Kyoto, Japan.

Chen, Yu, Eisele, Andreas, Federmann, Christian, Hassler, Eva, Jellinghaus, Michael and Theison, Silke. 2007. Multi-engine machine translation with an open-source SMT deocder. In *Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association for Computational Linguistics*, pages 193–196, Prague, Czech Republic.

Chiang, David. 2007. Hierarchical Phrase-based Models of Translation. *Computational Linguistics* 2(33), 201âĂŞ228.

Dalrymple, Mary. 2001. *Lexical-Functional Grammar*. Academic Press.

Deneefe, Steve and Knight., Kevin. 2009. Synchronous Tree Adjoining Machine Translation. In *Proceedings of Empirical Methods in Natural Language Processing Conference*, pages 727–736, Edinburgh, United Kingdom.

Deneefe, Steve, Knight, Kevin, Wang, Wei and Marcu, Daniel. 2007. What can Syntax-based MT learn from Phrase-based MT? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processsing and Computational Natural Language Learning*, pages 755–763, Prague, Czech Republic.

Doddington, George. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. In *Proceedings of Human Languages Technologies Conference*, San Diego, California.

Eisele, Andreas. 2005. First steps towards multi-engine machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Compuational Linguistics Workshop on Building and Using Parallel Texts*, pages 155–158, Ann Arbor, Michigan.

Eisele, Andreas, Federmann, Christian, Saint-Amand, Herve, Jellinghaus, Mochael, Herrmann, Teresa and Chen, Yu. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation at the 46th Annual Meeting of the Association for Compuational Linguistics*, pages 179–182, Columbus, Ohio.

Graham, Yvette. 2010. Sulis: An Open Source Transfer Decoder for Deep Syntactic Statistical Machine Translation. *In The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation* .

Graham, Yvette and van Genabith, Josef. 2008. Packed Rules for Automatic Transfer Rule Induction. In *Proceedings of the European Association of Machine Translation Conference 2008*, Hamburg, Germany.

Graham, Yvette and van Genabith, Josef. 2009. An Open Source Rule Induction Tool for Transfer-Based SMT. *The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation* pages 37–46.

Graham, Yvette and van Genabith, Josef. 2010a. Deep Syntax Language Models and Statistical Machine Translation. In *Proceedings of the Fourth International Workshop on Syntax and Structure in Statistical Translation at The 23rd International Conference on Computational Linguistics*, Beijing, China.

Graham, Yvette and van Genabith, Josef. 2010b. Factor Templates for Factored Machine Translation Models. In *Proceedings of the International Workshop on Spoken Language Translation*, Paris, France.

Graham, Yvette, van Genabith, Josef and Bryl, Anton. 2009. F-structure Transfer-Based Statistical Machine Translation. In *Proceedings of the 14th International Lexical Functional Grammar Conference*, Cambridge: CSLI Publications.

Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical Functional Grammar, a Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 173–281.

Kaplan, Ronald M., King, Tracy Holloway and Maxwell, John T. 2002. Adapting existing grammars: the XLE experience. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Kaplan, Ronald M., Riezler, Stefan, King, Tracy Holloway, Maxwell, John T. and Vasserman, Alexander. 2004. Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Meeting*, pages 97–104, Boston, MA.

Koehn, Philip, Och, Franz Josef and Marcu, Daniel. 2003. Statistical phrase-based translation. In *Proceedings of Human Language Technology - North American Chapter of the Association for Computational Linguistics Conference*, pages 48–54, Alberta, Canada.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand.

Koehn, Philipp. 2009. *Statistical Machine Translation*. Cambridge University Press.

Koehn, Philipp and Hoang, Hieu. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language*

Processing and Computational Natural Language Learning, pages 868–876, Prague, Czech Republic.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondrej, Constantin, Alexandra and Hoang, Evan HerbstHieu. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Li, Zhifei, Callison-Burch, Chris, Khundanpur, Sanjeev and Thorton, Wren. 2009. Decoding in Joshua, Open Source Parsing-Based Machine Translation. In *The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation*, volume 91, pages 47–56.

Och, Franz Josef. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Riezler, Stefan, King, Tracy H., Kaplan, Ronald M., Crouch, Richard, Maxwell, John T. and Johnson, Mark. 2002. Parsing the Wall Street Journal using Lexical Functional Grammar and discriminitive estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Philadelphia, PA.

Riezler, Stefan and Maxwell, John. 2006. Grammatical Machine Translation. In *Proceedings of Human Language Technologies and the 44th Annual Meeting of the Assciation for Computational Linguistics*, pages 248–255, New York City, NY.

Snover, Mathew, Dorr, Bonnie, Scwartz, Richard, Makhoul, John and Micciula, Linnea. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the 7th biennial Conference of the Association for Machine Translaiton in the Americas*, pages 223–231, Boston, MA.

Snover, Mathew, Dorr, Bonnie, Scwartz, Richard, Makhoul, John, Micciula, Linnea and Weischeidel, Ralphe. 2005. A Study of Translation Error Rate with Targeted Human Annotation. Technical Report, University of Maryland, College Park, MD.

Stolcke, Andreas. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Zaidan, Omar. 2009. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation* 91, 79–88.