A Dependency-Constrained Hierarchical Model with Moses

Yvette Graham[†‡]
[†]Department of Computing and Information Systems, The University of Melbourne
[‡]Centre for Next Generation Localisation, Dublin City University
ygraham@unimelb.edu.au

Abstract

This paper presents a dependency-constrained hierarchical machine translation model that uses Moses open-source toolkit for rule extraction and decoding. Experiments are carried out for the German-English language pair in both directions for projective and non-projective dependencies. We examine effects on SCFG size and automatic evaluation results when constraints are applied with respect to projective or non-projective dependency structures and on the source or target language side.

1 Introduction

A fundamental element of natural language syntax is the dependency structure encoding the binary asymmetric head-dependent relations captured in dependency grammar theory. A main criteria for determining the dependency structure of a given sentence is the following: *The linear position of dependent, D, is specified with reference to its head, H* (Kübler et al., 2009). This runs in parallel with that which hierarchical machine translation SCFG rules encode: *The linear position of a translated phrase, X_i, is specified with reference to the lexicalised words in the rule.* Figure 1 shows dependency structures for *She lives in a white house* and its Spanish translation, with example SCFG rules shown in Figure 2. Given the existence of initial rules (1), (2) and (3), hierarchical rules (4) and (5) can be created. Rule (4) specifies the linear position of the translation of the English phrase that precedes *house* with reference to lexicalised *casa*.

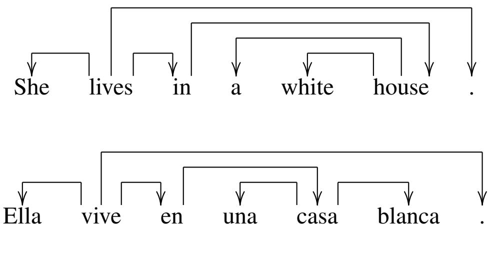

Figure 1: Projective Dependency Structures

(1) $X \rightarrow < \textit{white house , casa blanca} >$
(2) $X \rightarrow < \textit{white , blanca} >$
(3) $X \rightarrow < \textit{house , casa} >$
(4) $X \rightarrow < X_0 \textit{ house , casa } X_0 >$
(5) $X \rightarrow < \textit{white } X_0 , X_0 \textit{ blanca} >$

Figure 2: Initial rules (1), (2) and (3), with hierarchical rules (4) and (5)

For hierarchical machine translation models (Chiang, 2005), there is no requirement for a syntactic relationship to exist between the lexicalised words of a rule and the words replaced by non-terminals, the only requirement being that substituted words form an SMT phrase (Koehn et al., 2003). The dependency structure of either the source or target (or indeed both) can, however, be used to constrain rule extraction as to only allow hierarchical rules in which the linear position of dependents are specified with reference to the position of their lexicalised heads. For example, in the case of the hierarchical rules in Figure 2, rule (4) satisfies such a constraint according to both the source and target language dependency structures (since *white* is the dependent of *house* and *blanca* is the dependent of *casa*, and it is both *white* and *blanca* that are replaced by non-terminals while the heads remain lexicalised) and results in a synchronous grammar rule that positions a dependent relative to the position of its lexicalised head. Rule (5), on the other hand, does not satisfy such a constraint for either language dependency structure.

In this work, we examine a dependency-constrained model in which hierarchical rules are

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 464–471,
Sofia, Bulgaria, August 8-9, 2013 ©2013 Association for Computational Linguistics

only permitted in which lexicalised heads specify the linear position of missing dependents, and examine the effects of applying such constraints across a variety of settings for German to English and English to German translation.

2 Related Work

The increased computational complexity introduced by hierarchical machine translation models (Chiang, 2005), has motivated techniques of constraining model size as well as decoder search. Among such include the work of Zollmann et al. (2008) and Huang and Xiang (2010), in which rule table size is vastly reduced by means of filtering low frequency rules, while Tomeh et al. (2009), Johnson et al. (2007) and Yang and Zheng (2009) take the approach of applying statistical significance tests to rule filtering, with Lee et al. (2012) defining filtering methods that estimate translational effectiveness of rules.

Dependency-based constraints have also been applied in a variety of settings to combat complexity challenges. Xie et al. (2011) use source side dependency constraints for translation from Chinese to English, while Shen et al. (2010) apply target-side dependency constraints for the same language pair and direction in addition to Arabic to English, Peter et al. (2011) also apply dependency constraints on the target side, but rather soft constraints that can be relaxed in the case that an ill-formed structure does in fact yield a better translation. Gao et al. (2011) similarly apply soft dependency constraints but to the source side for Chinese to English translation, and Galley and Manning (2009) show several advantages to using maximum spanning tree non-projective dependency parsing decoding for Chinese to English translation. Li et al. (2012), although not constraining with dependency structure, instead create non-terminals with part-of-speech tag combinations for Chinese words identified as heads for translation into English.

In this paper, we apply the same dependency constraint to SCFG rule extraction in a variety of configurations to investigate effects of applying constraints on the source or target side, to the language with most or least free word order, as well as constraining with non-projective dependency structures.

Non-Projective Dependencies	
German	38%
English	11%

Table 1: WMT Parallel Training Data

3 Non-Projective Dependencies

A non-projectivity structure is defined as follows: *A non-projective dependency structure is a dependency structure in which at least one dependency relation exists between a head, H, and its dependent, D, in which the directed path from H to D does not include at least one word positioned linearly in the surface form between H and D.* Figure 3 shows an example non-projective dependency structure arising from English *Wh-fronting*.

Non-projective dependencies occur frequently for many languages, increasingly so for languages with high levels of free words order. An examination of Chinese treebanks, for example, reports that Chinese displays nine different kinds of non-projective phenomena (Yuelong, 2012) with reports of as many as one in four sentences in tree banks having non-projective dependency structures (Nivre, 2007). Even for a language with relatively rigid word order such as English non-projectivity is still common, due to Wh-fronting, topicalisation, scrambling and extraposition. Table 1 shows the frequency of non-projective dependency structures in WMT parallel data sets for German and English when parsed with a state-of-the-art non-projective dependency parser (Bohnet, 2010).

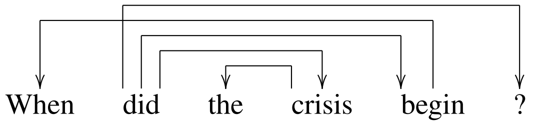

Figure 3: Non-projective Dependency Structure

4 Constrained Model

We define the dependency constraint as follows: *to create a hierarchical rule by replacing a word or phrase with a non-terminal, all the words of that phrase must belong to a single complete dependency tree and its head must remain lexicalised in the rule*. In this way, the hierarchical rules of the SCFG position missing dependents relative to the position of lexicalised heads. Before extracting SCFG rules for the dependency-constrained models, we transform non-projective structures into projective ones, in order to allow the substitution of non-projective dependency trees by a single non-terminal. Although the transformation simplifies the dependency structure, it will introduce some dis-fluency to the training data, and we therefore include experiments to examine such effects.

Figure 4 shows a German-English translation constrained by means of the German dependency structure and Figure 5 shows the full set of dependency-constrained hierarchical SCFG rules, where dependents are specified with reference to lexicalised heads.

5 Implementation with Moses

For rule extraction we use Moses (Williams and Koehn, 2012) implementation of GHKM (Galley et al., 2004; Galley et al., 2006), which although is conventionally used to extract syntax-augmented SCFGs from phrase-structure parses (Zollmann and Venugopal, 2006), we apply the same rule extraction tool to dependency parses. Rule extraction is implemented in such a way as not to be restricted to any particular set of node labels. The conventional input format is for example:

```
<tree label="NP">
  <tree label="DET"> the </tree>
  <tree label="NN"> cat </tree>
</tree>
```

The dependency-constrained ruleset can be extracted with this implementation by arranging dependency structures into tree structures as follows:[1]

```
<tree label="X">
  <tree label="X">
    <tree label="X"> the </tree>
    <tree label="X"> black </tree>
    cat
  </tree>
  ate
  <tree label="X">
    <tree label="X"> the </tree>
    rat
  </tree>
</tree>
```

Since XML format requires nesting of substructures, only projective dependency structures can be input to the tool in the way we use it, as non-projectivity breaks nesting.

6 Non-Projectivity Transform

We therefore transform non-projective dependency structures into projective ones by relocating the dislocated dependent to a position closer to its head so that it no longer violates projectivity. We do this in such a way as not to break any of the existing dependency relations between pairs of words. Figure 6 shows an example non-projective structure (a) before and (b) after the transformation, where the transformation results in the constituent comprised of words *when* and *begin* forming a continuous string, making possible the substitution of this constituent with a non-terminal. The fact that one side of the training data from which hierarchical rules are extracted, however, is no longer guaranteed to be fluent, raises the question as to what effect this disfluency might have when the constraint is applied on the target side. We therefore include in our evaluation for both language directions (and for the case where the constraints are applied to the source) the effects of word reorder cause by the transformation. The algorithm for converting non-projective structures is an inorder traversal of the dependency structure as follows, where words are indexed according to their position in the original string prior to the transformation:

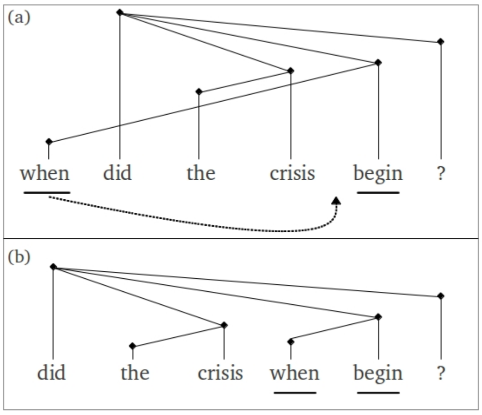

Figure 6: Non-Projectivity Transformation

Algorithm 6.1: DEP_IN_ORD($root$)

```
for each $d \in D$ and $d.index < root.index$
  do $dep\_in\_ord(d)$
PRINT($root$)
for each $d \in D$ and $d.index > root.index$
  do $dep\_in\_ord(d)$
```

[1] Note that is is possible to replace X with dependency labels.

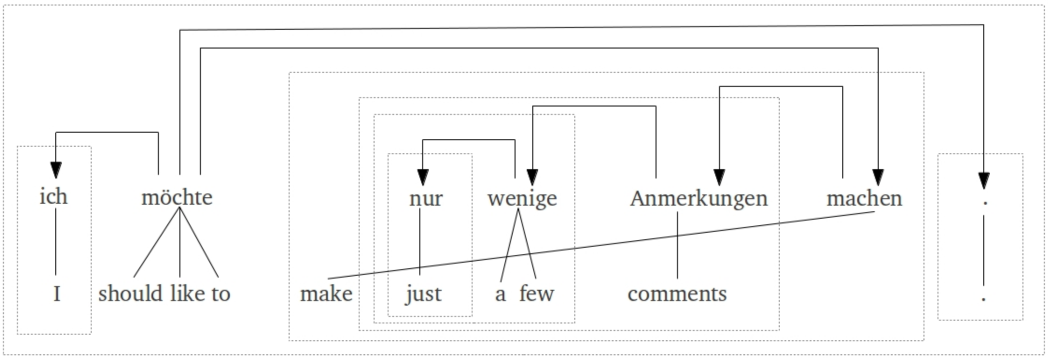

Figure 4: German English translation with German dependency structure, words surrounded by a dashed box form a complete dependency tree.

Rules spanning source words 0-6: ich möchte nur wenige anmerkungen machen .

X_0 möchte nur wenige anmerkungen machen .	X_0 should like to make just a few comments .
ich möchte X_0 .	i should like to X_0 .
ich möchte X_0 machen .	i should like to make X_0 .
ich möchte X_0 anmerkungen machen .	i should like to make X_0 comments .
ich möchte X_0 wenige anmerkungen machen .	i should like to make X_0 a few comments .
ich möchte nur wenige anmerkungen machen X_0	i should like to make just a few comments X_0
non-proj X_0 möchte X_1 .	X_0 should like to X_1 .
X_0 möchte X_1 machen .	X_0 should like to make X_1 .
X_0 möchte X_1 anmerkungen machen .	X_0 should like to make X_1 comments .
X_0 möchte X_1 wenige anmerkungen machen .	X_0 should like to make X_1 a few comments .
X_0 möchte nur wenige anmerkungen machen X_1	X_0 should like to make just a few comments X_1
ich möchte X_0 X_1	i should like to X_0 X_1
ich möchte X_0 machen X_1	i should like to make X_0 X_1
ich möchte X_0 anmerkungen machen X_1	i should like to make X_0 comments X_1
ich möchte X_0 wenige anmerkungen machen X_1	i should like to make X_0 a few comments X_1
X_0 möchte X_1 X_2	X_0 should like to X_1 X_2
X_0 möchte X_1 machen X_2	X_0 should like to make X_1 X_2
X_0 möchte X_1 anmerkungen machen X_2	X_0 should like to make X_1 comments X_2
X_0 möchte X_1 wenige anmerkungen machen X_2	X_0 should like to make X_1 a few comments X_2

Rules spanning source words 2-5: nur wenige anmerkungen machen

X_0 machen	make X_0
X_0 anmerkungen machen	make X_0 comments
X_0 wenige anmerkungen machen	X_0 a few comments

Rules spanning source words 2-4: nur wenige anmerkungen

X_0 anmerkungen	X_0 comments
X_0 wenige anmerkungen	X_0 a few comments

Rules spanning source words 2-3: nur wenige

X_0 wenige	X_0 a few

Figure 5: Complete set of dependency-constrained hierarchical SCFG rules for Figure 4

7 Experiments

WMT training data sets were used for both parallel (1.49 million German/English sentence pairs) and monolingual training (11.51 million English & 4.74 million German sentences). Mate non-projective dependency parser (Bohnet, 2010) was used for parsing both the German and English parallel data with standard pre-trained models, the same parser was used for projective parsing with non-projectivity turned off.[2] Parallel training data lines containing multiple sentences were merged into a single pseudo-dependency structure by adding an artificial root and head-dependent relation between the head of the initial sentence and any subsequent sentences. Non-projective dependencies were converted into projective structures using Algorithm 6.1.

Giza++ (Och et al., 1999) was employed for automatic word alignment, and Moses GHKM rule extraction (Williams and Koehn, 2012) was used for hierarchical rule extraction for the dependency-constrained models. Default settings were used for rule extraction for all models with the exception on non-fractional counting being used, as well as Good-turing discounting. Both the dependency-constrained and standard models use the same set of initial rules. For decoding, since only a single non-terminal, X, is present for all models, Moses hierarchical decoder (Koehn et al., 2007) was used with default settings with the exception of rule span limit being removed for all models. SRILM (Stolke, 2002) was used for 5-gram language modeling and Kneser-Ney smoothing (Kneser and Ney, 1995) for both German-to-English and English-to-German translation. MERT (Och, 2003) was carried out on WMT newstest2009 development set optimizing for BLEU, and final results are reported for held-out test sets, newstest2010 and newstest2011, with BLEU (Papineni et al., 2001) and LR-score (Birch and Osborne, 2010) for evaluation.

7.1 Results

Table 2 shows automatic evaluation results for both the dependency-constrained and standard hierarchical models for both language directions. Compared to the standard hierarchical model (orig), the best performing dependency-constrained models, sl_npr (de-en) and tl_npr (en-de), show significant decreases in mean BLEU score, -0.44 for German to English and -0.13 for English to German. However, there is a trade-off, as the dependency-constrained models achieve vast reductions in model size, approx. 93% for German to English and 89% for English to German in numbers of SCFG hierarchical rules. This results in decreased decoding times, with the best performing dependency-constrained models achieving a decrease of 26% for German to English and 34% for English to German in mean decoding times.

The decrease in BLEU scores is not likely to be attributed to less accurate long-distance reordering for German to English translation, as the Kendall Tau LR-scores for this language direction show an increase over the standard hierarchical models of +0.25 mean LR. Although this is not the case for English to German, as mean LR scores show a slight decrease (-0.11 LR).

The number of hierarchical rules (not including glue rules) employed during decoding provides a useful indication of to what degree each model actually uses hierarchical rules to construct translations, i.e. not simply concatenating phrases with glue rules. For English to German translation, while the number of hierarchical rules present in the SCFG is vastly reduced, the number of hierarchical rules used during decoding actually increases, with double the number of hierarchical rules used to translate test segments compared to the standard hierarchical model, from an average of only 0.58 hierarchical rules per segment for the standard model to 1.19 per segment. This indicates that the set of hierarchical rules is refined by the dependency constraint.

When the more linguistically valid non-projective dependency structure, as opposed to the projective dependency structure, is used to constrain rule extraction significant increases in BLEU scores are achieved for all configurations. The most significant gains in this respect occur when constraints are applied on the source side, +0.58 mean BLEU for German to English and +0.50 mean BLEU for English to German.

In general, when constraints are applied to the more free word order language, German, regardless of whether or not translation is *into* or *out of* German, marginally higher BLEU scores result, with an increase of +0.03 mean BLEU for German to English translation and similarly an increase of

[2] OpenNLP (Feinerer, 2012) sentence splitter is recommended with the parser we and was used for preprocessing.

			SCFG hier. rules (millions)	newstest 2010		newstest 2011		mean	mean hier. rules decoder	mean segment decode time (seconds)
				BLEU	LR-K	BLEU	LR-K	BLEU		
de-en	hpb	orig	35.25	22.30	**71.86**	20.47	**70.55**	21.39	2.51	6.76
		tl_re	34.77	22.31	71.43	**20.49**	70.27	**21.40**	2.63	6.39
		tl_are	34.77	**22.41**	71.16	20.36	69.89	21.39	2.68	6.14
		sl_are	33.87	22.40	70.78	20.27	69.78	21.34	2.71	6.02
		sl_re	33.87	22.06	71.38	20.15	70.25	21.11	2.41	6.17
	dc	sl_npr	2.49	21.57	71.87	20.09	71.04	20.95	1.15	4.99
		tl_npr	1.45	21.88	72.20	19.95	71.36	20.92	2.85	4.62
		tl_pr	1.12	21.43	71.82	19.75	70.90	20.59	1.40	3.62
		sl_pr	0.34	21.05	72.20	19.69	71.36	20.37	1.10	1.98
en-de	hpb	orig	36.30	16.14	**70.24**	**15.05**	**69.91**	**15.60**	0.58	7.25
		tl_re	35.20	16.13	69.81	14.94	69.45	15.54	1.03	5.16
		tl_are	35.20	**16.15**	69.06	14.57	68.66	15.36	1.89	4.82
		sl_are	35.68	15.72	69.25	14.44	69.06	15.08	1.88	5.23
		sl_re	35.68	15.72	70.21	14.38	69.84	15.05	1.16	5.16
	dc	tl_npr	4.00	16.03	70.12	14.91	69.81	15.47	1.19	4.79
		sl_npr	1.09	15.94	70.07	14.85	69.69	15.40	1.78	3.46
		tl_pr	0.92	15.88	70.46	14.78	69.90	15.33	1.23	4.05
		sl_pr	0.88	15.58	70.18	14.22	69.80	14.90	1.19	2.90

Table 2: Effects of dependency constraints and dependency-based reordering on translation quality for German-to-English (de-en) and English to German (en-de), hpb=hierarchical phrase-based, orig=no reordering, *re=dependency-based word reordering where only hierarchical rules are extracted from reordered training data, *are=dependency-based word reordering where all SCFG rules extracted from reordered training data, dc=dependency-constrained, *pr=projective parse used for dependency constraint, *npr=non-projective parse used for dependency constraint, sl*=constraints or reordering for source language, tl*=constraints or reordering for target language, numbers of hierarchical rules reported do not include glue rules.

+0.07 mean BLEU for English to German, with the increase being statistically significant for German to English for the newstest2010 test set, but not statistically significant for newstest2011 test set or English to German (Koehn, 2004).

Overall the best performing dependency-constrained models are those that retain the highest numbers of hierarchical rules in the SCFG. This indicates that although the dependency-constrained models produce a refined ruleset, they nevertheless discard some SCFG rules that would be useful to translate the unseen test data. One possible reason is that although the non-projective dependency structures are significantly better, these high-quality linguistic structures may still not be optimal for translation. Another possibility is that a the GHKM rule extraction constraints combined with the dependency constraint is causing a small set of very useful rules to be discarded.

7.2 Dependency-based Reordering

We examine the effects of the non-projective transformation in isolation of any dependency-constraints by training a standard hierarchical model on the reordered corpus with no dependency constraints applied. We do this in two set-ups. First, we extract hierarchical rules from the reordered training corpus and initial rules from the original unaltered corpus (*_re in Table 2), as this is the set-up for the dependency-constrained models. Simply for interest sake, we repeat this experiment but extract all rules (hierarchical and initial rules) from the reordered corpus (*_are in Table 2).

Surprisingly, when non-projective reordering is carried out on the target side no significant decrease in BLEU scores occurs for both language directions. In fact, a minor increase in mean BLEU (+0.01) is observed for German to English translation, but this small increase is not statistically significant. For the English to German direction, a minor decrease of -0.06 mean BLEU occurs (not statistically significant).

Similarly for German to English, when reordering is applied to the source side, only a minor decrease (-0.05) results. Non-projective reordering causes the most significant reduction in performance for English to German when the English source is reordered, with a decrease of -0.52 mean BLEU.

Conclusions

This paper examines non-projectivity and language application for dependency-constrained hierarchical models using Moses open-source toolkit. Experiments show that when applied to English to German translation, vastly reduced model size and subsequently decreased decoding times result with only a minor decrease in BLEU. In addition, higher numbers of (non-glue) hierarchical rules are used to translate test segments. For German to English translation, similar decreases in model size and decoding times occur, but at the expense of a more significant decrease in BLEU.

In general, results for the dependency-constrained models show that applying constraints on the source or target side does not have a major impact on BLEU scores. Rather the use of high quality linguistic structures is more important, as significant improvements are made for all configurations when the non-projective dependency structure is used to constrain rule extraction.

Acknowledgments

Many thanks to the Moses developers, especially Hieu Hoang and Alexandra Birch. Thanks also to Tsuyoshi Okita and anonymous reviewers. This research was supported by the Australian Research Council, as well as Science Foundation Ireland as part of the Centre for Next Generation Localisation.

References

Alexandra Birch and 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*, pages 327–332, Uppsala, Sweden, July. Association for Computational Linguistics.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Ingo Feinerer. 2012. tm: Text mining package. *R package version 0.5-7.1.*

Michel Galley and Christopher D Manning. 2009. Quadratic-time dependency parsing for machine translation. 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*, pages 773–781. Association for Computational Linguistics.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *HLT-NAACL*, pages 273–280.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and 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*, pages 961–968. Association for Computational Linguistics.

Yang Gao, Philipp Koehn, and Alexandra Birch. 2011. Soft dependency constraints for reordering in hierarchical phrase-based translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 857–868. Association for Computational Linguistics.

Fei Huang and Bing Xiang. 2010. Feature-rich discriminative phrase rescoring for smt. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 492–500. Association for Computational Linguistics.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Accoustics, Speech and Signal Processing*, volume 1.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL).*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo* *and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Sandra Kübler, Ryna McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Seung-Wook Lee, Dongdong Zhang, Mu Li, Ming Zhou, and Hae-Chang Rim. 2012. Translation model size reduction for hierarchical phrase-based statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 291–295, Jeju Island, Korea, July. Association for Computational Linguistics.

Junhui Li, Zhaopeng Tu, Guodong Zhou, and Josef van Genabith. 2012. Using syntactic head information in hierarchical phrase-based translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 232–242, Montréal, Canada, June. Association for Computational Linguistics.

Joakim Nivre. 2007. Incremental non-projective dependency parsing. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 396–403, Rochester, NY.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)*, pages 20–28.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report, September 17.

Jan-Thorsten Peter, Matthias Huck, Hermann Ney, and Daniel Stein. 2011. Soft string-to-dependency hierarchical machine translation. In Marcello Federico, Mei-Yuh Hwang, Margit Rödder, and Sebastian Stüker, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 246–253.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2010. String-to-dependency statistical machine translation. *Computational Linguistics*, 36(4):649–671.

Andreas Stolke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 577–585.

Nadi Tomeh, Nicola Cancedda, and Marc Dymetman. 2009. Complexity-based phrase-table filtering for statistical machine translation. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. International Association for Machine Translation.

Philip Williams and Philipp Koehn. 2012. GHKM rule extraction and scope-3 parsing in moses. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 434–440, Montreal, Canada, June. Association for Computational Linguistics.

Jun Xie, Haitao Mi, and Qun Liu. 2011. A novel dependency-to-string model for statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 216–226, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Mei Yang and Jing Zheng. 2009. Toward smaller, faster, and better hierarchical phrase-based smt. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 237–240, Suntec, Singapore, August. Association for Computational Linguistics.

Wang Yuelong. 2012. *Edge-crossing Non-projective Phenomena in Chinese Language*. Ph.D. thesis, National University of Singapore.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City, June. Association for Computational Linguistics.

Andreas Zollmann, Ashish Venugopal, Franz Josef Och, and 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 (Coling 2008)*, pages 1145–1152, Manchester, UK, August. Coling 2008 Organizing Committee.