Factoring Adjunction in Hierarchical Phrase-Based SMT

Sophie Arnoult
ILLC
University of Amsterdam
s.i.arnoult@uva.nl

Khalil Sima'an
ILLC
University of Amsterdam
k.simaan@uva.nl

Abstract

While much work has been done to inform Hierarchical Phrase-Based SMT (Chiang, 2005) models linguistically, the adjunct/argument distinction has generally not been exploited for these models. But as Shieber (2007) points out, capturing this distinction allows to abstract over 'intervening' adjuncts, and is thus relevant for (machine) translation in general. We contribute an adjunction-driven approach to hierarchical phrase-based modelling that uses source-side adjuncts to relax extraction constraints–allowing to capturing long-distance dependencies–, and to guide translation through labelling. The labelling scheme can be reduced to two adjunct/non-adjunct labels, and improves translation over Hiero by up to 0.6 BLEU points for English-Chinese.

1 Introduction

Hiero (Chiang, 2005) extends phrase-based Statistical Machine Translation models (Koehn et al., 2003) by allowing phrase pairs to rewrite through a Synchronous CFG mechanism. Rewriting is unconstrained, and the model thus learns local dependencies and reorderings in a very general manner. This lack of restrictions allows the grammar to achieve good coverage, but begs the question of how to guide Hiero with linguistic information. Since SAMT (Zollmann and Venugopal, 2006), a branch of work has focused on labelling Hiero, with different types of labels: phrase-structure labels (Zollmann and Venugopal, 2006), dependency head labels (Li et al., 2012), CCG labels (Almaghout et al., 2011), (non-syntactic) hierarchical alignment labels (Maillette de Buy Wenniger and Sima'an, 2013), etc. Most of these models use large nonterminal vocabularies, as syntactic labels or POS tags are combined into phrase labels.

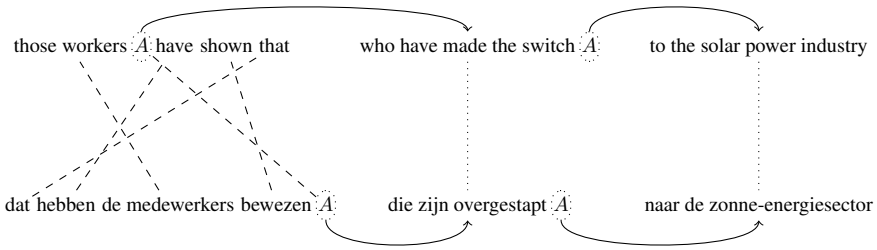
The general character of Hiero is balanced by constraints on the extraction and the form of rules, and another branch of work has focused on rebalancing such constraints. For instance, Li et al. (2013) constrain rewriting to constituents or sequences of constituents, allowing them to relax phrase length at extraction. Perhaps the most obvious limitation of Hiero is its limited capacity to capture sentence-level reordering, as it can only monotonically concatenate larger fragments. This has motivated work on reordering, e.g., (Mylonakis and Sima'an, 2011; Huck et al., 2012).

We propose to extend the scope of rule extraction in Hiero around adjuncts. As adjunction introduces long-distance dependencies, allowing the extraction of larger phrases that contain adjuncts should lead to phrases that still capture useful dependencies. This is akin to the linguistic motivation for Tree-Adjoining Grammar (Joshi et al., 1975), where factoring recursion allows to keep dependencies local (Joshi and Schabes, 1997). Our model relaxes length constraints for phrase extraction by discounting the length of adjuncts contained in a phrase. This allows to learn phrases that Hiero may not learn, such as in the example of Figure 1. Ignoring intervening adjuncts at phrase extraction reduces the apparent length of the source sentence, allowing for its extraction under the standard Hiero phrase-length constraint. As the adjuncts in this example introduce a complex phrase permutation, our model is able to extract rules from this phrase, that cannot otherwise be rendered with Hiero and monotonic glue rules. Besides, we inform the model by labelling adjuncts and non-adjuncts separately. While adjuncts form only a fraction of the

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the 2nd Deep Machine Translation Workshop (DMTW 2016), pages 11–20,
Lisbon, Portugal, 21 October 2016.

Figure 1: Example sentence; adjuncts cause long-distance dependencies (10 tokens separate 'workers' from 'have' in the English sentence) and complex reorderings (adjunction introduces a 2-4-1-3 permutation).



phrase pairs that can be extracted by Hiero, we find that this labelling is useful, allowing to gain up to 0.6 BLEU points on English-Chinese combined with a basic feature set.

Factoring adjunction also allows to learn more general rules. DeNeefe and Knight (2009) show this improves translation for syntax-based models. We propose to extend the Hiero grammar by excising adjuncts from extraction phrases. This is similar in spirit to the approach of (Arnoult and Sima'an, 2012) for phrase-based models, but with the added capacity to extract SCFG rules from modified phrases. In our example, this allows to extract rules from the (adjunct-free) phrase "those workers have shown that"/"dat hebben de medewerkers bewezen".

The rest of this article is organized as follows: section 2 deals with adjunction, and how we identify it; section 3 presents our extensions to Hiero; section 4 presents experiments on three language pairs, with English as source language, and Chinese, Dutch and French as target languages; we discuss the results of these experiments in section 5 and conclude in section 6.

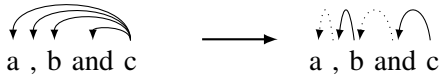
2 Identifying adjuncts

Adjunction in Tree-Adjoining Grammar allows to explain a number of linguistic phenomena like raising or wh movement (Kroch and Joshi, 1985), but we focus here on modification, which we identify with dependency labels (we use the Turbo parser[1]).

We identify modifier labels and punctuation with adjuncts: `AMOD`, `NMOD`, `VMOD` and `P`. We exclude cases where the dependent appears obligatory, based on the head's POS tag: `DT`, `EX`, `IN`, `POS`, `MD`, `PRP`, `PRP$`, `RP`, `SYM`, `TO`, `WDT`, `WP`, `WP$`, `WRB`, `.`; we also exclude possessors in genitive constructions, by excluding dependents preceding a dependent with a `POS` part of speech.

We include dependents of enumerations and conjunctions in the list of modifiers. This follows from a dependency analysis that treats one of the conjuncts as the head, and conjunctions and other conjuncts as its dependents. In the case of the Turbo parser, the last conjunct or element in an enumeration is regarded as the head. We modified this representation to a nested one, where conjunctions are treated as heads of the dependent conjunct, as shown in Figure 2.

Figure 2: Modifying the representation of enumerations and conjunctions. Dotted lines represent non adjuncts.



[1] `http://www.cs.cmu.edu/~ark/TurboParser/`

3 Model

We present an adjunction-driven extension to Hiero with two distinguishing features: we use adjuncts to guide the extraction from larger phrases than normally allowed by Hiero; and we apply labelling to let the model distinguish adjuncts from arguments and other phrases. We present the extraction constraints in section 3.1, the labelling method in section 3.2, and the features we use for the model in section 3.3. Additionally, section 3.4 presents another extension, inspired from (Arnoult and Sima'an, 2012), that leverages adjunct optionality to extract additional rules.

3.1 Adjunction-driven extraction constraints

Hiero limits phrase spans for rule extraction through a *max-phrase-length* constraint (of typically 10 tokens). This limit is needed to restrict the number of extractable phrases, that may grow exponentially with sentence length. Further, reorderings may manifest themselves differently locally than at the sentence level, so that the task of learning sentence-level rules may be better handled separately. However, as adjunction introduces long-distance dependencies, factoring it out should allow to extract more relevant phrase pairs.

We use adjunction as a guide to extending rule extraction for larger phrases. Like Hiero, we allow extraction and unconstrained rewriting of all phrases under *max-phrase-length*. For larger phrases, we subject extraction and rewriting to three constraints: *max-effective-length*, *non-adjunct-crossing*, and *max-target-symbols*. Besides, we apply specific constraints to adjuncts and adjunct sequences.

max-effective-length

We define the effective length as the non-adjunct token count of phrase. Let a phrase ϕ, that contains $\alpha_0..\alpha_n$ adjuncts (disregarding adjuncts embedded in other adjuncts), its effective length $\lambda(\phi)$ is:

$$\lambda(\phi) = \text{len}(\phi) - \Sigma_{i=0}^{n}\text{len}(\alpha_i)$$

In practice, we set *max-effective-length* to the same value as *max-phrase-length*.

This constraint only applies to non-adjunct phrases, as we allow the extraction of all phrases that match an adjunct on the source side, or a group of adjuncts: we group together adjuncts that have the same orientation with regard to their head, and that form contiguous sequences on source and target sides.

no-adjunct-crossing

This constraint prevents the extraction of larger phrases that cross adjuncts, or groups of adjuncts. This forces rewriting to an adjunct group as a whole. When rewriting from an adjunct group, one only forbids adjunct crossings, allowing rewriting to sub-groups.

max-target-symbols

Hiero limits the number of right-hand-side symbols on the rules' source sides. The length of the target side can also be limited: for Hiero, we apply *max-phrase-length* to the target side of extraction phrase pairs; for the adjunction-based models, we limit the number of target right-hand-side symbols to the same value as *max-phrase-length*.

Table 1 shows a possible derivation for the example of Figure 1. Allowing the extraction of rules from larger phrases permits to capture long-range dependencies and reorderings inaccessible to Hiero. While rule r_1 is likely in fact to be learnt by Hiero in a different context, rule r_2 displays a pattern (extraposed modifier in the Dutch sentence but not in the English sentence) that is only likely to occur with a long modifier.

3.2 Labelling

To further guide the model, we apply labelling to distinguish adjuncts from other phrases. We identify source adjuncts and adjunct sequences, and label both sides of rules and rule gaps accordingly: with an `A` label for adjuncts; an `A`x label for adjunct groups of size x; and a default label for other phrases.

Table 1: Example rules for the example in Figure 1

r_1	X $\rightarrow \langle$ X that , dat X $\rangle$
r_2	X $\rightarrow \langle$ those workers $A^{[1]}$ have shown , hebben de medewerkers bewezen $A^{[1]}$ $\rangle$
r_3	A $\rightarrow \langle$ $X^{[1]}$ made the switch $A^{[2]}$, $X^{[1]}$ overgestapt $A^{[2]}\rangle$
r_4	X $\rightarrow \langle$ who have , die zijn $\rangle$
r_5	A $\rightarrow \langle$ to the solar power industry , naar de zonne-energiesector $\rangle$

3.3 Features

The model uses two rule features to distinguish larger phrase pairs from Hiero-extractable phrase pairs: a *long-distance* feature, corresponding to the probability estimate that a rule was extracted from a larger phrase pair (exceeding Hiero's *max-phrase-length*); and an *adjunct-crossing* feature corresponding to the probability that a rule was extracted from a (shorter) phrase pair violating the *non-adjunct-crossing* constraint.

Besides, we tested a version of the model with a simplified labelling for adjunct sequences. These sequences are then labelled with $\mathtt{A}$ instead of $\mathtt{A}x$, while their size x appears in the following feature:

$$f_x = e^{1-x} \tag{1}$$

For other rules (adjuncts and other phrase pairs), f_x is taken to be 1.

3.4 Factoring out adjuncts

TAG factors adjunction by extracting auxiliary trees and initial trees separately (Joshi and Schabes, 1997). This leads to a more compact grammar (Chiang, 2000) that is able to generate unseen adjunction patterns. Synchronous Tree Adjunction Grammar (STAG) (Shieber and Schabes, 1990) applies TAG to translation, and DeNeefe and Knight (2009) propose a probabilistic implementation for string-to-tree translation. Their model identifies target-side adjuncts and takes their projection on the source side as a basis for auxilliary-tree extraction.

In the case of Hiero, one cannot directly implement STAG, as CFG rules do not have the (tree) structure that is necessary for modelling adjunction. One can still however extract generalized versions of rules, by factoring out adjuncts contained in extraction phrases. This follows (Arnoult and Sima'an, 2012), who apply this idea to a phrase-based model. The hierarchical nature of Hiero further allows to apply substitution in these generalized phrases.

We extend Hiero by extracting rules both by standard phrase substitution, and by adjunct factorization. For each phrase pair in the training data, we first extract rules by substitution. For each adjunct contained in the phrase pair, we instantiate a copy of the extraction phrase where the adjunct is *blind*: the adjunct blocks the extraction of overlapping gaps, and its yield is excised from the rule. We then extract rules by phrase substitution from this extraction phrase; Table 2 shows some of the resulting rules for the example of Figure 1. The rules extracted in this manner form a subset of the rules that Hiero would extract from the phrase pair $\langle$ those workers have shown, hebben de medewerkers bewezen $\rangle$, as we forbid gaps from overlapping with blind adjuncts.

Table 2: Some rules added by adjunct factorization

X	$\rightarrow$	$\langle$ those workers have shown , hebben de medewerkers bewezen $\rangle$
X	$\rightarrow$	$\langle$ those $X^{[1]}$ have shown , hebben de $X^{[1]}$ bewezen $\rangle$
X	$\rightarrow$	$\langle$ those workers have $X^{[1]}$, hebben de medewerkers $X^{[1]}$ $\rangle$
X	$\rightarrow$	$\langle$ $X^{[1]}$ have $X^{[2]}$, hebben $X^{[1]}$ $X^{[2]}$ $\rangle$

The combinations of adjuncts that can be excised from a phrase grow exponentially with the number of adjuncts in the phrase. Even if this number remains small in general, adjunct factorization is applied to all phrases, in an extraction space that is already increased by extending extraction-phrase spans. Besides, the number of adjuncts in a phrase may also be high occasionally, especially since we regard enumeration tails as adjuncts. This concern motivates the hierarchical nesting of enumerated elements presented in section 2.

We contain grammar size increase by excising one adjunct at a time in adjunct-group phrases, and one adjunct group (or stand-alone adjunct) at a time in other phrases.

The adjunct factorization we propose for Hiero is incomplete as it does not fully extract adjuncts from phrase pairs. Compared to STAG, our grammar extracts 'derived' rules with generalized adjunction patterns, rather than separating 'auxiliary' from 'initial' rules. Consequently, our grammar increases in size rather than becoming more compact.

4 Experiments

4.1 Data

We performed experiments on three language pairs: English-Chinese, English-Dutch and English-French. For all experiments, word alignments were obtained using GIZA++ with 'grow-diag-final-and' symmetrization (Och and Ney, 2003). The English side of the data were parsed using the Turbo parser, and converted to adjunct parses following the criteria of section 2. We used a 4-gram language model, trained with KenLM (Heafield et al., 2013).

The English-Chinese data were taken from the MultiUN corpus (Eisele and Chen, 2010), limited to sentences of up to 40 tokens. We first extracted an in-domain development and test set by randomly drawing 4000 sentences without replacement from the corpus (after having removed English-side duplicates), and splitting the resulting set in two. Word alignments were trained on the rest of the corpus (ca. 5.6M sentence pairs). The language model was trained on the Xinhua section of the Chinese Gigaword corpus (LDC2003T09).

The English-Dutch data were taken from the Europarl corpus (v7). We extracted a development and test set of 2000 sentence pairs each following the same method as for the English-Chinese data. The language model was trained on the target side of the training corpus.

The English-French data were taken from the Europarl corpus (v7), limited to sentences of up to 40 tokens. We used the Europarl 2006 development and test sets, and trained the language model on the target side of the corpus.

For English-Chinese, we used training sets of two different sizes. Table 3 summarizes the sizes and average sentence length of the different data sets.

Table 3: Data-set sizes

		train		dev	test
fr	sentences	500k		2k	2k
	avg. tokens	20.6		29.0	29.7
nl	sentences	500k		2k	2k
	avg. tokens	27.4		27.6	27.1
zh	sentences	500k	2M	2k	2k
	avg. tokens	22.5	22.5	22.7	22.6

4.2 Tuning and Decoding

All models use an extended set of dense features (not counting adjunction features), following Maillette de Buy Wenniger and Sima'an (2013). Feature weights are tuned with MIRA (Cherry and Foster, 2012), for 20 iterations.

Decoding is performed with Joshua (Li et al., 2009), with a relaxation of the decoding span to 100 tokens. This allows hierarchical rules to span an entire sentence in the case of the extended models.

4.3 Adjunction-based-model results

Results for English-Chinese, with a small training set

Table 4 presents test results on the smaller English-Chinese training set (500k sentence pairs). These tests compare the adjunction-based models, with and without labelling or features, to a Hiero baseline. We also tested the effect of relaxing the decoding span on Hiero. We use the following identifiers for the models: `H-100` is a Hiero model with a relaxed decoding span, `adj` uses adjunction-based constraints, but no labels or adjunction features; `adj-F` also uses the *long-distance* and *adjunct-crossing* features; `adj-L` uses labels (including adjunct-sequence labels); `adj-FL` uses both features and labels; `adj-L2F` replaces the adjunct-sequence labels by their corresponding feature.

Table 4: Experimental results for English-Chinese; training set size=500k[a]

	Hiero	H-100	adj	adj-F	adj-L	adj-FL	adj-L2F
BLEU	21.8	21.7	21.5*	22.0	22.1*	22.3*	22.3*
BEER	11.2	11.2	11.1	11.3	11.4*	11.4*	11.4*
TER	63.8	64.4*	64.1*	64.1*	64.3*	63.9	64.3*
LENGTH	99.8	100.1*	98.1*	99.5*	100.0*	99.7	99.8
LR-KB1[b]	0.265	0.262	0.258	0.260	0.261	0.263	0.261

[a] We mark significance levels of $p = 0.05$; each model was tuned and decoded three times.
[b] The LR-KB1 scores were computed giving equal weight to BLEU-1 and Kendall's tau ($\alpha = 0.5$)

While extending extraction spans with the `adj` model decreases performance, both labelling and the base adjunction features allow to guide decoding in the adjunct-driven models, and outperform Hiero, both in terms of BLEU and BEER (Stanojević and Sima'an, 2014). The highest improvements are obtained for the models employing both labelling and features, with little or no difference between the full-label model `adj-FL` and the label-to-feature model `adj-L2F`. The lack of improvement in LR score (Birch and Osborne, 2011) suggest that the adjunct-driven models improve lexical selection rather than reordering.

Effect of training set size

Table 5 presents results for the larger English-Chinese data set (2M training sentence pairs). With a larger data set, relaxing the decoding span for Hiero (`H-100`) is beneficial for English-Chinese–locally learned rules are useful when applied to larger spans. As before, extending extraction spans alone decreases performance, but labels and features allow to guide the model and improve performance; the `adj-L2F` model outperforms Hiero by 0.6 BLEU point.

Table 5: Experimental results for English-Chinese; training set size=2M

	Hiero	H-100	adj	adj-L2F
BLEU	23.2	23.5*	23.0	23.8*
BEER	12.5	12.6	12.5	12.8
TER	61.9	62.0	61.4*	62.1
LENGTH	98.9	98.7	96.9*	99.1
LR-KB1	0.272	0.268	0.268	0.270

[a] Results are based on two tuning runs

Tests on other language pairs

Table 6 presents results for English-French and English-Dutch for training sets of 500k sentence pairs. We find that the adjunction-driven model performs similarly to Hiero for both these language pairs.

Table 6: Experimental results for English-Dutch and English-French; training size=500k

	en-nl				en-fr			
	Hiero	H-100	adj	adj-L2F	Hiero	H-100	adj	adj-L2F
BLEU	27.5	27.5	27.5	27.4	32.9	32.8	33.0	32.7
BEER	16.4	16.3	16.4	16.3	23.6	23.6	23.6	23.5
TER	59.5	59.6	59.5	59.6	53.9	54.3	53.9	54.1
LENGTH	99.9	100.3	99.8	99.7	99.1	99.3	99.3	99.3
LR-KB1	0.307	0.307	0.306	0.305	0.390	0.388	0.390	0.389

[a] Results are based on a single tuning round

Inspection of output translations shows several cases of improved lexical selection for French. For instance, the `adj-L2F` model is able to capture the dependency between 'enthusiasm' and 'wane' in the first example in Table 7, and to translate both words appropriately. One also can find examples of improved reordering, as in the second example in the table. While both Hiero and the `adj-L2F` model wrongly reorder the translations of 'geopolitical' and 'geographical', making them appear as dependents of 'outpost' and 'population' respectively, the `adj-L2F` model is able, unlike Hiero, to preserve the dependency between 'outpost' and 'of europe'.

Table 7: Example translations

Improved lexical selection	
src	the problem is that , if you set a date , there is a danger that the enthusiasm for reform in these countries will wane .
Hiero	le problème est que , si vous wane fixer une date , il y a un risque que l' enthousiasme de réforme dans ces pays .
adj-L2F	le problème est que , si vous fixer une date , il y a un risque que l' enthousiasme de réforme dans ces pays diminue .
Limited reordering improvement	
src	because of its geopolitical position as the last outpost of europe , at the crossroads with the middle east and north africa , the importance of malta goes far beyond its geographical size and its small population.
Hiero	en raison de sa position en tant que dernier retranchement géopolitique , au carrefour avec le moyen-orient et l' afrique du nord , l' importance de malte va bien au-delà de sa taille et sa petite population géographique de l' europe .
adj-L2F	en raison de sa position en tant que dernier retranchement géopolitique de l' europe , à la croisée des chemins avec le moyen-orient et l' afrique du nord , l' importance de malte va bien au-delà de sa taille et sa petite population géographique .

Adjunct factorization model

Table 8 presents preliminary results for the adjunct factorization model of section 3.2 (`adj-Opt`). While this model does not use adjunction or features, the gap in performance with regard to Hiero appears bigger than for the `adj` model.

5 Discussion

We have presented an adjunction-based hierarchical phrase-based model, that extends Hiero in two ways: by letting adjuncts guide the extraction of larger phrase pairs, and by marking adjuncts through labelling.

Our model outperforms Hiero for English-Chinese on training sets of moderate size (500k and 2M sentence pairs), by 0.5 and 0.6 BLEU points respectively. The improvement is brought by the combination of extraction features with a minimal adjunct/non-adjunct source labelling scheme. This is a very positive result, that shows that the adjunct/argument distinction can be useful for machine translation, even

Table 8: Experimental results for the adjunct-optionality model, for English-Chinese

	training=500k				training=2M			
	BLEU	BEER	TER	LEN	BLEU	BEER	TER	LEN
Hiero	21.7	11.1	64.6	100.0	23.4	12.6	62.0	98.8
adj-Opt	21.0**	10.9	64.1*	97.9**	22.8**	12.3	61.8	97.5**

[a] Results are based on a single tuning round

though our means to identify adjuncts are coarse (in the example of Figure 1 for instance, "to the solar power industry" is argueably an argument of "made the switch", and not an adjunct). Beside we assumed here that source adjuncts project into target adjuncts. This is an optimistic assumption (Hwa et al., 2002; Arnoult and Sima'an, 2014), and we are bound to extract many erroneous rules. The adjunct-driven model is however able to guide the model sufficiently well to ward off these rules for English-Chinese. Refining the labels and features is likely to further enhance the model.

While our feature set may be improved, we face the difficulty that the current features are informative of the extraction of a rule, and we accordingly store their values along with the rules in the grammar. This increases the size occupied by the grammar in memory, making it harder to extract grammars for larger training sets.

We found that the adjunct-driven model provides no improvement over Hiero for English-Dutch and English-French. We believe that the improvement for English-Chinese is related to the extent of reordering in this language pair: while system scores suggest a mostly lexical improvement, reordering in English-Chinese may favor the application of hierarchical rules (rather than glue rules), and benefit more from the linguistic constraint brought by the adjunct-driven model.

Additionally, we have presented another extension, that leverages adjunct optionality to extract rules by excising adjuncts and their projections, following what Arnoult and Sima'an (2012) had done for a phrase-based model. The resulting model underperforms Hiero for English-Chinese, and while an adapted feature and label set may improve results, selecting which adjuncts to excise appears necessary.

6 Conclusion

We have presented an adjunct-driven extension to Hiero: the model uses source-side adjuncts to extract larger phrases and to label rules. The model is able to improve over Hiero for English-Chinese with minimal labelling and a few features. This improvement appears to be mostly lexical: the model captures long-distance dependencies better, but not long-distance reorderings. We found no improvement for English-Dutch and English-French. The lesser extent of reordering in these language pairs may limit the application of rules involving adjuncts; further constraining the model may then be beneficial for these language pairs too.

We have also presented a second extension, that factors adjunction to derive rules with simpler adjunction patterns. This extension leads to a decrease in performance compared to Hiero: while an adapted feature and label set may help this model, constraints on which adjuncts to excise are likely to be necessary as well.

Acknowledgments

This research is part of the project "Statistical Translation of Novel Constructions", which is supported by NWO VC EW grant 612.001.122 from the Netherlands Organisation for Scientific Research (NWO).

References

Hala Almaghout, Jie Jiang, and Andy Way. 2011. CCG contextual labels in hierarchical phrase-based SMT. In *Proceedings of the 15th conference of the European Association for Machine Translation*, pages 281–288.

Sophie Arnoult and Khalil Sima'an. 2012. Adjunct Alignment in Translation Data with an Application to Phrase-Based Statistical Machine Translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 287–294.

Sophie Arnoult and Khalil Sima'an. 2014. How Synchronous are Adjuncts in Translation Data? In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 157–165, Doha, Qatar.

Alexandra Birch and Miles Osborne. 2011. Reordering Metrics for MT. In *Proceedings of the Association for Computational Linguistics*, Portland, Oregon, USA.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada.

David Chiang. 2000. Statistical Parsing with an Automatically-Extracted Tree Adjoining Grammar. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 456–463.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270.

Steve DeNeefe and Kevin Knight. 2009. Synchronous Tree Adjoining Machine Translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 727–736.

Andreas Eisele and Yu Chen. 2010. MultiUN: A Multilingual Corpus from United Nation Documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Matthias Huck, Stephan Peitz, Markus Freitag, and Hermann Ney. 2012. Discriminative Reordering Extensions for Hierarchical Phrase-Based Machine Translation. In *16th Annual Conference of the European Association for Machine Translation*, pages 313–320, Trento, Italy.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating Translational Correspondence Using Annotation Projection. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 392–399.

Aravind K. Joshi and Yves Schabes. 1997. Tree-Adjoining Grammars. In G. Rosenberg and A. Salomaa, editors, *Handbook of Formal Languages*. Springer-Verlag, New York, NY.

Aravind K. Joshi, Leon S. Levy, and Masako Takahashi. 1975. Tree Adjunct Grammars. *Journal of Computer and System Sciences*, 10(1):136–163.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT/NAACL*, pages 127–133.

Anthony Kroch and Aravind Joshi. 1985. The Linguistic Relevance of Tree Adjoining Grammars. Technical Report MC CIS 85 18, Department of Computer and Information Science, University of Pennsylvania.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece.

Junhui Li, Zhaopeng Tu, Guodong Zhou, and Josef van Genabith. 2012. Using Syntactic Head Information in Hierarchical Phrase-Based Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 232–242.

Junhui Li, Philip Resnik, and Hal Daumé III. 2013. Modeling Syntactic and Semantic Structures in Hierarchical Phrase-based Translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–549, Atlanta, Georgia.

Gideon Maillette de Buy Wenniger and Khalil Sima'an. 2013. Hierarchical Alignment Decomposition Labels for Hiero Grammar Rules. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 19–28, Atlanta, Georgia.

Markos Mylonakis and Khalil Sima'an. 2011. Learning Hierarchical Translation Structure with Linguistic Annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 642–652.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

Stuart Shieber and Yves Schabes. 1990. Synchronous Tree-Adjoining Grammars. In *Handbook of Formal Languages*, pages 69–123. Springer.

Stuart M. Shieber. 2007. Probabilistic Synchronous Tree-Adjoining Grammars for Machine Translation: The Argument from Bilingual Dictionaries. In *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, Rochester, New York.

Miloš Stanojević and Khalil Sima'an. 2014. Evaluating Word Order Recursively over Permutation-Forests. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 138–147, Doha, Qatar.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax Augmented Machine Translation via Chart Parsing. In *Proceedings of NAACL 2006 - Workshop on statistical machine translation*, pages 138–141.