Modelling the Adjunct/Argument Distinction 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

We present the first application of the adjunct/argument distinction to Hierarchical Phrase-Based SMT. We use rule labelling to characterize synchronous recursion with adjuncts and arguments. Our labels are bilingual obtained from dependency annotations and extended to cover non-syntactic phrases. The label set we derive in this manner is extremely small, as it contains only thirty-six labels, and yet we find it useful to cluster these labels even further. We present a clustering method that uses label similarity based on left-hand-side/right-hand-side joint trained-model estimates. The results of initial experiments show that our model performs similarly to Hiero on in-domain French-English data.

1 Introduction

Labelling Hierarchical Phrase-Based models (Hiero) (Chiang, 2005) allows to disambiguate Hiero, while benefitting from its broad coverage. Using syntactic labels for labelling as Zollmann and Venugopal (2006) do with Syntax-Augmented Machine Translation (SAMT) or, e.g., Li et al. (2012) in an inspired approach, yields however unwieldy models with large non-terminal vocabularies. We propose to approach the labelling problem from the other end, using the adjunct-argument distinction to minimally label Hiero.

We interpret adjuncts in the general sense of modifiers, and not only of adjuncts in semantic frames. Generally speaking, the adjunct-argument distinction accounts for a difference in selectional preferences: arguments are selected by their heads, while adjuncts select their heads. This distinction is modelled in Tree-Adjoining Grammar (Joshi et al., 1975; Joshi and Schabes, 1997), through substitution and adjunction. Shieber and Schabes (1990) and Shieber (2007) have proposed Synchronous Tree-Adjoining Grammar (STAG) by for SMT, and the adjunct/argument distinction has been applied to Syntax-Based models notably by DeNeefe and Knight (2009) and Liu et al. (2011).

We do not attempt here to model adjunction in Hiero, rather we reduce the adjunct-argument distinction to one of type. The semantic aspect of this distinction–adjuncts modify the meaning of a phrase, while arguments complete it–makes it appealing for Machine Translation, as one may expect that it can be preserved across a bitext. To circumvene mismatches, we label both sides of the data to derive bilingual labels. The label set that we derive is minimal as we start from two labels for adjuncts and arguments on both sides of the data, and derive only four new labels for non-syntactic phrases; after combining source and target labels into bilingual labels, the label set contains thirty-six labels only.

We conduct experiments on French-English data, and show that while direct application of this small adjunct/argument label set leads to sub-optimal results, promising results can be obtained by clustering bilingual labels. While further tests are required, our model is currently limited by Hiero's phrase-length limit; to fully apply adjunct/argument labelling, one needs to extend this model, with reordering rules for instance, or by exchanging the phrase-length constraint for a constraint on recursion.

Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015), pages 2–11,
Praha, Czech Republic, 3–4 September 2015.

2 Labelling Adjuncts and Arguments for Hiero

Our labelling procedure follows that of SAMT (Zollmann and Venugopal, 2006) to some extent. We start from sentence pairs that have been parsed on both sides of the data into dependencies, and we map dependency labels to either adjuncts or arguments, as is Figure 1[1].

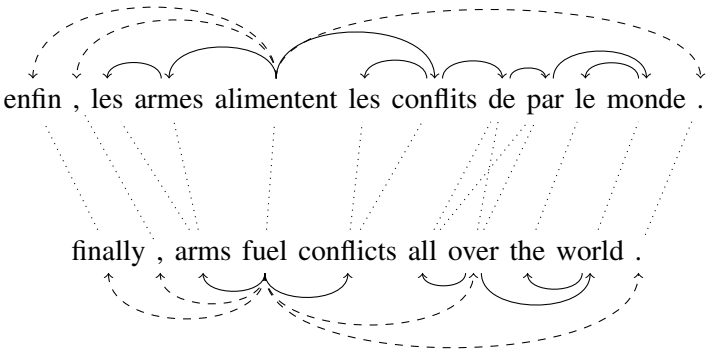

Figure 1: Example sentence pair. Adjunct dependencies are indicated with dashes.

Dependency labels vary per parser, but we broadly map modifier and punctuation labels to adjuncts, and remaining labels to arguments. Table 1 presents the mapping from the dependency converters of Candito et al. (2010) for French and of Johansson and Nugues (2007) for English.

Table 1: Adjunct-mapping criteria for English and French

	head-governor relation	other constraints on head h, governor g, etc.
English	ADV, APPO, PRN	
	AMOD	relation(g, governor(g)) ! = ADV
	PMOD	h precedes g
	NMOD	POS(h) $\notin$ {*CC, DT, EX, POS, MD, PRP, PRP$, RP, SYM, WDT, WP, WP$, WRB*}
	P	h has no dependents
French	mod, mod_rel	
	ponct	h has no dependents

2.1 Phrase-Labelling scheme

Next, we define phrase labels to allow for recursion over non-syntactic phrases. Our phrase labelling procedure is summarized as Algorithm 1. This scheme follows SAMT and beyond that Combinatorial Categorial Grammar (CCG) (Steedman, 2000), but it is coarser on the one hand, and it is adapted to syntactic differences between adjunct and arguments, on the other.

It is coarser in that the added phrasal labels, while corresponding to incomplete constituents or constituent sequences, are in fact kept to a minimum, that do not reflect the combination logic of CCG: we distinguish incomplete adjuncts, incomplete arguments, sequences of arguments, and default all remaining phrases to a single type.

To reflect specific adjunct/argument behaviour, we let constituents that miss adjuncts keep their type, thus reflecting the fact that adjuncts do not alter the syntactic type of the phrases they modify; for the same reason, we label sequences of arguments and adjuncts as a sequence of arguments; finally, we label

[1]Word alignments are tentative, but the adjunct/argument labels are factual; the labelling of *"de par le monde"* is the result of a parsing error.

sequences of adjuncts as a single adjunct, reflecting the absence of linguistic restriction on the number of adjuncts for a given phrase (although we do not actually test whether adjuncts have the same governor).

Algorithm 1: Labelling procedure for extracted phrases.

input : A phrase ϕ and a dependency tree with adjunct (A) and argument (C) labels.
output: A phrase label for ϕ
if ϕ *matches a dependent* D **then** Label(ϕ) $\leftarrow$ Label(D)
else if ϕ *matches a sequence of dependents* D_i **then**
  if *all* D_i *are adjuncts* **then** Label(ϕ) $\leftarrow$ A **else** Label(ϕ) $\leftarrow$ $\mathtt{C_S}$
else if ϕ *matches a dependent* D *less some left and/or right sub-dependents* SD_i **then**
  if *all* SD_i *are adjuncts* **then** Label(ϕ) $\leftarrow$ Label(D)
  else
    if D *is an adjunct* **then** Label(ϕ) $\leftarrow$ $\mathtt{A_I}$ **else** Label(ϕ) $\leftarrow$ $\mathtt{C_I}$
else Label(ϕ) $\leftarrow$ P

2.2 Bilingual labelling

The adjunct/argument label set presented thus far can be equally applied on the source and target sides of the data. To account for parsing differences and linguistic divergence (Dorr, 1994; Hwa et al., 2002; Arnoult and Sima'an, 2014), we combine source and target labels into composite, bilingual labels. The resulting label set consists of 36 labels.

Table 2 shows some phrase pairs for the example of Figure 1.

Table 2: Adjunct/argument-based phrasal labels

label	French phrase	English phrase	label	French phrase	English phrase
AA	enfin ,	finally ,	C_IC_I	alimentent	fuel
CC	les armes	arms	C_IC_I	monde	world
CA	de par le monde	all over the world	C_SC_S	le monde .	the world .
C_IA_I	de par	all over	PP	monde .	world .
C_IP	les conflits de par	conflicts all over			

The phrase *"finally ,"* is labelled as an adjunct as it is a sequence of adjuncts with the same governor; *"the world ."* is labelled as an argument sequence as it is a multi-headed sequence containing an argument; *"conflicts all over"* is labelled as a phrase (P) as it is a multi-headed sequence containing an incomplete argument.

3 Model

The model is a SAMT-like, labelled-variant of Hiero (Chiang, 2005). The model is similar to Hiero, but for the fact that the single non-terminal of Hiero is replaced by a set of labels.

Model derivations are scored by a log-linear model over features; our model uses most of the Hiero and SAMT features. Like Hiero, the model features comprehend phrase-translation weights, lexical weights, rule penalty, glue-rule penalty and word penalty; the phrase-translation-weights feature also applies for adjunct/argument-labelled models, and is then computed on unlabelled rule equivalents, i.e., on lexical content only. Like SAMT (Zollmann, 2011), the model uses features for left-hand-side-conditioned rule weights, labelled-rule translation weights, rule-rarity penalty, and flags for lexical-only rules, abstract rules, monotone rules, and abstract-target rules. Unlike SAMT, we do not condition labelled on unlabelled sides: phrase-translation weights are computed on labelled rules on the one hand, and on unlabelled rules on the other hand.

4 Adjunction-label clustering

Hanneman and Lavie (2013) propose a clustering method for SAMT labels to reduce their amount and the resulting computational load. Their method employs source labels next to the usual SAMT target labels: combining source and target labels allows them to compute relative-frequency estimates of source/target labels, which serve to compute distance measures between source labels on one hand, and target labels on the other. The distance measure between two source labels s_1 and s_2 is defined as the marginal difference between $P(t|s_1)$ and $P(t|s_2)$ estimates; the distance between target labels is defined similarly. Clustering proceeds by searching the source or target labels that minimize either one of the source-label and the target-label distances, and collapsing the resulting label pair. Clustering stops after a predefined number of iterations, after which only the clustered target labels are used to extract an SAMT grammar. The resulting model proves superior to SAMT on a Chinese-English task, and generally superior to Hiero.

Even though our bilingual label set is very small, the combination of source and target labels is ad-hoc, and initial experiments show it is misadapted. To correct this, we adapt the method of Hanneman and Lavie (2013) to cluster combined, bilingual labels.

4.1 Label-distance measures

Rather than using a joint distribution of source and target labels to compute label distance, we use a joint distribution of left-hand-side and right-hand-side labels. We define a distance d_{LHS} between left-hand-side label occurrences, and a distance d_{RHS} between right-hand-side label occurrences.

The lhs distance between two non-terminals v_1 and v_2 in the bilingual label set U is computed by marginalizing the difference between non-terminal rewriting probabilities, where probability estimates are obtained by heuristic counting of joint LHS/RHS labels in extracted labelled rules:

$$d_{LHS}(v_1, v_2) = \sum_{v \in U} (P_{RHS|LHS}(v|v_1) - P_{RHS|LHS}(v|v_2)) \tag{1}$$

This distance captures similarities in the rewriting behaviour of non-terminals.

For the rhs distance, we tested two definitions. The first one, d^n_{RHS}, mirrors the lhs distance, by marginalizing the difference between inverse non-terminal rewriting probabilities:

$$d^n_{RHS}(v_1, v_2) = \sum_{v \in U} (P_{LHS|RHS}(v|v_1) - P_{LHS|RHS}(v|v_2)) \tag{2}$$

Under this definition, two non-terminals are similar if they have similar generating distributions.

The second one, d^u_{RHS}, marginalizes the difference between joint lhs/rhs probabilities over left-hand-side non-terminal labels:

$$d^u_{RHS}(v_1, v_2) = \sum_{v \in U} (P_{LHS,RHS}(v, v_1) - P_{LHS,RHS}(v, v_2)) \tag{3}$$

Under this definition, the similarity in right-hand-side label occurrences is not normalized anymore by right-hand-side label probabilities, so this rhs distance is also conditioned on right-hand-side labels having similar frequencies.

We derive a single label distance measure by adding the lhs and rhs distances. Depending on the variant of rhs distance (*normalized* or *unnormalized*), we obtain either d_n or d_u:

$$d_n(v_1, v_2) = d_{LHS}(v_1, v_2) + d^n_{RHS}(v_1, v_2) \tag{4}$$

$$d_u(v_1, v_2) = d_{LHS}(v_1, v_2) + d^u_{RHS}(v_1, v_2) \tag{5}$$

4.2 Clustering

Clustering proceeds by searching at each step for the label pair that minimizes label distance. The two closest labels are clustered into a single label, and probability estimates are updated for the next round. Clustering stops when a predefined label-set size has been reached. The clustered bilingual labels can then be used to extract a new grammar.

5 Experiments

5.1 Experimental set-up

5.1.1 Data

We conduct experiments of French-English data from the Europarl corpus (v7) with in-domain test data from the WMT07 Europarl development and test sets (devtest2006 and test2006).

We use the Berkeley aligner[2] for training word alignments, with 5 rounds of IBM1 and HMM training; the training data consist of the French-English Europarl training set, containing 1.97M sentence pairs with a maximum length of 40 tokens. The data are tokenized with a script adapted from the Moses tokenizer and lowercased. The language model is a 4-gram model with interpolated Kneser-Ney smoothing, and is trained with KenLM[3] on the English side of the training set with a sentence-length limit of 80 tokens; the set contains 52.5M tokens.

The training data consist of 200k sentence pairs of length limited to 40 tokens, taken from the training set used for the language model and the word alignments; the data contain 4.18M English tokens;

5.1.2 Annotations

We parse both sides of the training data with the Berkeley Parser[4]–the data are then true-cased–, and then convert parses to dependency parses: with the Pennconverter of Johansson and Nugues (2007) for English, and the Functional Role Labeller of Candito et al. (2010) for French.

5.1.3 Model Training and decoding

We train models using an in-house grammar extractor, and a decoder based on Joshua[5]. Training and decoding constraints and defaults are the same as for Hiero, but we disallow consecutive non-terminals on both sides, and not only on the source side.

Model parameters are tuned with Mira, allowing up to 20 iterations. Following (Clark et al., 2011) we average results over three rounds of tuning/decoding.

5.2 First results

Table 3 reports tests on adjunct/argument label sets, where we use source-language labels only (`AA-Src`), target-language labels (`AA-Trg`), or combined, bilingual labels (`AA-Bi`).

Table 3: Performance of monolingual and bilingual labelling schemes with regard to Hiero; significant differences are marked with one $\triangledown$ for $p = 0.05$ and two for $p = 0.01$

	BLEU		METEOR		TER	
	dev	test	dev	test	dev	test
Hiero	32.1	31.8	34.9	34.8	52.9	53.3
AA-Src	31.9$^{\triangledown\triangledown}$	31.3$^{\triangledown\triangledown}$	34.8$^{\triangledown}$	34.7$^{\triangledown\triangledown}$	53.0	53.5$^{\triangledown\triangledown}$
AA-Trg	32.0$^{\triangledown}$	31.6$^{\triangledown\triangledown}$	34.9	34.7$^{\triangledown}$	52.9	53.5$^{\triangledown\triangledown}$
AA-Bi	31.9$^{\triangledown}$	31.5$^{\triangledown\triangledown}$	34.8	34.7$^{\triangledown\triangledown}$	53.0	53.5$^{\triangledown}$

All models underperform Hiero, on the test set more than on the development set, and on BLEU more than Meteor or TER. The `AA-Src` model performs worse: source-labelling models are most known to guide reordering, which is relatively absent in French-English. The `AA-Bi` model appears to give poorer results than the `AA-Trg` model, and that while it disposes of more information; argueably, even if the source-language labels are not directly useful, they might serve to refine target labels. We attribute the relatively poor results of the `AA-Bi` model to the undirected combination of source and target labels, and

[2] https://code.google.com/p/berkeleyaligner/

[3] http://kheafield.com/code/kenlm/

[4] https://github.com/slavpetrov/berkeleyparser

[5] http://joshua-decoder.org/

we use label-rewriting statistics on the development set grammar of the AA−Bi model to cluster labels as described in section 4.

5.3 Label clustering

We apply both definitions of the rhs distance of Equations 2 and 3 to extract two label sets. In both cases, we limit the final, clustered label-set size to six labels. Table 4 presents the label set obtained with the rhs-normalized distance d_n (equation 4), and Table 5 the label set obtained with the rhs-unnormalized distance d_u (Equation 5).

Table 4: Clusters obtained with normalized (conditional) RHS distance d_n and relative frequency of LHS occurrence

	clustered bilingual labels	$P(LHS)$
1	CA, CA$_I$, CC, CC$_I$, CC$_S$, CP, C$_I$A, C$_I$A$_I$, C$_I$C, C$_I$C$_I$, C$_I$C$_S$, C$_I$P	0.381
2	AA, AA$_I$, AC, AC$_I$, AC$_S$, AP, C$_S$A, C$_S$A$_I$, C$_S$C, C$_S$C$_I$, C$_S$C$_S$, C$_S$P, PA, PA$_I$, PC, PC$_I$, PC$_S$	0.255
3	PP	0.328
4	A$_I$C$_S$	0.024
5	A$_I$C, A$_I$C$_I$	0.012
6	A$_I$A, A$_I$A$_I$, A$_I$P	0.001

Table 5: Clusters obtained with unnormalized (joint) RHS distance d_u and relative frequency of LHS occurrence

	clustered bilingual labels	$P(LHS)$
1	CC, CC$_I$, C$_I$C, C$_I$C$_I$	0.288
2	CA, CA$_I$, CC$_S$, CP, C$_I$A, C$_I$A$_I$, C$_I$C$_S$, C$_I$P, C$_S$C, C$_S$C$_I$, C$_S$C$_S$, C$_S$P, PC, PC$_I$, PC$_S$, PP	0.595
3	AC, A$_I$C, A$_I$C$_I$	0.016
4	AA, AA$_I$, A$_I$A , A$_I$A$_I$	0.050
5	AC$_I$, AC$_S$, AP, A$_I$C$_S$, A$_I$P	0.018
6	C$_S$A, C$_S$A$_I$, PA, PA$_I$	0.032

Clusters obtained with d_n (Table 4) show a dominance of the source-label component: labels with an A_I source component form half of all clusters (clusters 4,5 and 6), and other labels–PP excepted–are clustered by their source component only.

In contrast, clusters obtained with d_u (Table 5) show some symmetry between source and target components, and they group together adjuncts and incomplete adjuncts, arguments and incomplete arguments, and multi-headed dependent sequences and phrases: cluster 1 corresponds to argument/argument translations; cluster 2 to argument/adjunct pairs and phrasal (multi-headed) or semi-phrasal equivalences; cluster 3 to adjunct/argument pairs; cluster 4 to adjunct/adjunct pairs; cluster 5 to adjunct/phrase pairs; and cluster 6 to phrase/adjunct pairs.

These clusters also lead to better translation results, as Table 6 shows.

The model trained with the label set of Table 4, AA−Cn performs worse than the original labelled model AA−Bi. The second label set leads to a better AA−Cu model, which performs significantly better than AA−Bi on the test set. Compared to Hiero, AA−Cu is still less performant on the development set–at least in terms of BLEU scores–, but achieves comparable results on the test set.

Table 6: Performance of clustered labelling schemes with regard to Hiero and the original bilingual-label model; significant differences with Hiero are marked with one $\triangledown$ for $p = 0.05$ and two for $p = 0.01$; significant differences with the original label-set model are marked with one $\blacktriangle$ for $p = 0.05$ and two for $p = 0.01$;

	BLEU		METEOR		TER	
	dev	test	dev	test	dev	test
Hiero	32.1	31.8	34.9	34.8	52.9	53.3
AA-Bi	31.9	31.5	34.8	34.7	53.0	53.5
AA-Cn	$31.8^{\triangledown\triangledown/\blacktriangledown}$	$31.4^{\triangledown\triangledown}$	34.8	$34.7^{\triangledown}$	$53.1^{\triangledown\triangledown}$	$53.6^{\triangledown\triangledown}$
AA-Cu	$31.9^{\triangledown\triangledown}$	$31.8^{\blacktriangle\blacktriangle}$	34.9	$34.8^{\blacktriangle\blacktriangle}$	53.0	$53.3^{\blacktriangle\blacktriangle}$

6 Discussion

Our first results show that direct matching of source and target labels leads to sub-optimal performance. Our solution uses rule estimates to cluster bilingual labels. This is orthogonal to the approach of Chiang (2010), who applies rule-matching features on both sides of the data, without explicitely matching source and target labels. While using bilingual labels is appealing as these labels are directly interpretable in terms of syntactic correspondence, clustering only allows to merge labels. A more refined method would allow to both split and merge labels, with the original adjunct/argument labels as a starting point for characterizing synchronous recursion linguistically.

As far as the current clustering procedure is concerned, we have shown that a distance based on rewriting patterns of left-hand-side non-terminals and occurrence patterns of right-hand-side non-terminals weighed over left-hand-side contexts leads to meaningful clusters and decent results. The clustered labels pair up labels of type adjunct, argument or (multi-headed) constituent sequence with their incomplete counterparts. This is not surprising: as, e.g., adjunct phrases rewrite largely to the same phrases as the incomplete adjunct phrases, their corresponding labels are close according to the left-hand-side distance; similarly, as both types of phrases are largely extracted from the same phrases, their corresponding labels are close according to the *joint* right-hand-side distance[6]. The results we obtain with these clustered labels suggest that the distinction between full and incomplete constituents is not essential for phrase labelling, which agrees with the labelling method of Li et al. (2012), where phrases are labelled with the highest, undominated head(s).

The translation results we present here are quite limited, both in extent and scores. One can first question whether labelling could increase performance for French-English; we intend to extend the application of the model to other language pairs in future work. Secondly, as we kept the Hiero constraints on phrase length and reordering–using labelled but otherwise standard glue rules–the effect of labelling can only be local. Possible extentions for our model would consist in extending the reordering capacities of Hiero with adjunct/argument reordering rules, or to use adjuncts to restrict recursion, thereby making way to lift Hiero's standard phrase-length constraint.

7 Related Work

Most work on adjunction in SMT takes place in a syntax-based framework, which forms a natural ground for STAG. DeNeefe and Knight (2009) and Liu et al. (2011) for instance have proposed tree-to-string models that differentiate between adjunction and substitution. The only application of adjunction to string-to-string models we know of is that of Arnoult and Sima'an (2012), who exploit the optional character of adjuncts to extract more rules for a Phrase-Based model.

While the first applications of syntax for SMT (Wu, 1997; Poutsma, 2000; Yamada and Knight, 2001)

[6]One can also note that, as the right-hand-side distance is not normalized, it takes lower values than the left-hand-side distance; we have not attempted to weigh them differently, but more experiments in this direction might be worth the while.

used constituency trees, recent work has come to use a larger array of linguistic formalisms: besides applications of TAG (DeNeefe and Knight, 2009; Liu et al., 2011), Xie et al. (2011) apply dependency syntax for tree-to-string modelling and Li et al. (2012) for labelling Hiero; Hassan et al. (2007) apply CCG supertags to phrase-based SMT and Almaghout et al. (2011) to Hiero; Xiong et al. (2012) apply predicate-argument structures in a hierarchical phrase-based model and Li et al.(2013) for in Hiero.

Labelling hierarchical models introduces new constraints while providing the opportunity to relax innate Hiero constraints. New constraints are: a limitation of the grammar to observed substitutions, which can be remedied by relaxing matching constraints using features to learn subsitution preferences (Chiang, 2010); an increase of rule sparsity and computational constraints, which can be remedied by label clustering (Hanneman and Lavie, 2013; Mino et al., 2014). The Hiero constraints that one attempts to relax are the monotonic top-level ordering, first and foremost (Huck et al., 2012; Li et al., 2012; Li et al., 2013). Li et al. (2012) also relax the source non-terminal adjacency constraint, while Li et al. (2013) relax the phrase-length constraint for extraction and decoding.

8 Conclusion

We have presented a bilingual labelling scheme for Hiero that is based on the adjunct/argument distinction. Even though our label set is very small, containing only thirty-six labels, we find that clustering these labels is useful. As it is, our model is able to perform similarly to Hiero on in-domain test data for French-English.

For future work, we plan to refine our labelling method, and to extend our model to circumvene the limited reordering capacity of Hiero: either with reordering rules, for which the adjunct/argument distinction should form a good basis, or through restrictions on recursion, which would allow to lift Hiero's phrase-length constraint.

Acknowledgements

We thank Gideon Maillette de Buy Wenniger for his help in using his grammar-extraction and labelling software and for useful discussions. We also thank the reviewers for their comments. This work is supported by The Netherlands Organisation for Scientific Research (NWO), with VC EW grant 612.001.122.

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*, page 157?165, Doha, Qatar, October. Association for Computational Linguistics.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of The seventh international conference on Language Resources and Evaluation (LREC)*.

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.

David Chiang. 2010. Learning to Translate with Source and Target Syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443–1452, Uppsala, Sweden, July. Association for Computational Linguistics.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

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.

Bonnie J. Dorr. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597–633.

Greg Hanneman and Alon Lavie. 2013. Improving Syntax-Augmented Machine Translation by Coarsening the Label Set. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 288–297, Atlanta, Georgia, June. Association for Computational Linguistics.

Hany Hassan, Khalil Sima'an, and Andy Way. 2007. Supertagged Phrase-Based Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 288–295, Prague, Czech Republic, June. Association for Computational Linguistics.

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, may.

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.

Richard Johansson and Pierre Nugues. 2007. Extended Constituent-to-dependency Conversion for English. In *Proceedings of NODALIDA 2007*, pages 105–112, Tartu, Estonia, May 25-26.

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.

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, June. Association for Computational Linguistics.

Yang Liu, Qun Liu, and Yajuan Lü. 2011. Adjoining Tree-to-string Translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1278–1287.

Hideya Mino, Taro Watanabe, and Eiichiro Sumita. 2014. Syntax-Augmented Machine Translation using Syntax-Label Clustering. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 165–171, Doha, Qatar, October. Association for Computational Linguistics.

Arjen Poutsma. 2000. Data-Oriented Translation. In *COLING*, pages 635–641.

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 Dekai Wu and David Chiang, editors, *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, Rochester, New York, 26 April.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA.

Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3):377–404.

Jun Xie, Haitao Mi, and Qun Liu. 2011. A Novel Dependency-to-String Model for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 216–226, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deyi Xiong, Min Zhang, and Haizhou Li. 2012. Modeling the Translation of Predicate-Argument Structure for SMT. In *ACL (1)*, pages 902–911.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Andreas Zollmann. 2011. *Learning Multiple-Nonterminal Synchronous Grammars for Statistical Machine Translation*. Ph.D. thesis.