Adjunct Alignment in Translation Data with an Application to Phrase-Based Statistical Machine Translation

Sophie Arnoult
Institute of Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, the Netherlands
s.arnoult@gmail.com

Khalil Sima'an
Institute of Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, the Netherlands
k.simaan@uva.nl

Abstract

Enriching statistical models with linguistic knowledge has been a major concern in Machine Translation (MT). In monolingual data, adjuncts are optional constituents contributing secondarily to the meaning of a sentence. One can therefore hypothesize that this secondary status is preserved in translation, and thus that adjuncts may align consistently with their adjunct translations, suggesting they form optional phrase pairs in parallel corpora. In this paper we verify this hypothesis on French-English translation data, and explore the utility of compiling adjunct-poor data for augmenting the training data of a phrase-based machine translation model.

1 Introduction

Phrase-Based Statistical Machine Translation (PB-SMT) (Koehn et al., 2003) exploits symmetrized word alignments (Brown et al., 1993) to form phrase pairs that capture the translation probabilities of idiomatic expressions. However, data sparsity is a major issue for phrase-based systems. It affects longer phrase pairs in particular, which are overestimated by the unsmoothed heuristic counts. Smoothing has been proposed to improve probability estimations in the phrase table (Kuhn et al., 2006; Foster et al., 2006), and minimal phrase pairs to alleviate data sparsity: see the tuples of Schwenk (2007) and the minimal translation units of Quirk and Menezes (2006). In both cases, these new units of translation are utilized in an n-gram translation model that allows to capture contextual dependencies, and their estimates are smoothed.

Morphology has also been proposed to reduce data sparsity, either by integrating morphological information into the translation model or as a preprocessing step. For instance, Nießen and Ney (2004) propose hierarchical lexicon models in a German-English system, with feature functions to integrate different levels of morphosyntactic abstraction, from full word forms to lemmas, whereas El Kholy and Habash (2010) present different morphological tokenization models for Arabic.

In this work we propose to use adjuncts to *augment and smooth training data* for machine translation. The term 'adjunct' is used here to refer to both clausal adjuncts and phrase modifiers, regardless of the nature of the modified category, e.g., verbal or nominal. As optional constituents that further qualify a complete clause or phrase, adjuncts can be removed from or added to a monolingual sentence without affecting its grammaticality. This idea is embodied most visibly in Tree-Adjoining Grammar (Joshi, 1983) where recursion in syntax is factored out into auxiliary lexical elements that modify initial sentences by adjunction operation. Here, we hypothesize that adjuncts, by their secondary semantic status, are likely to be preserved in translation. To our knowledge, this constitutes the first study of adjunct alignment in parallel data, though this idea is related to the Direct Correspondence Assumption (DCA) of Hwa et al. (2002). The DCA postulates that syntactic relations, e.g., between heads and arguments or between heads and modifiers, are preserved in translation. Hwa et al. (2002) project English unlabeled dependency parses on Chinese data and report 30.1% precision and 39.1% recall for the Chinese dependencies. Another related work is that of Dorr et al. (2002), which measured structural di-

vergence, in terms of predicate-argument-modifier structure, by automatically detecting regular expressions in a Spanish corpus. The detected expressions were then verified manually, and are seen as giving a lower bound on structural divergence. The authors found that 11% of sentences contained divergent structures, and 35% with relaxed regular expressions.

We measured adjunct alignment on a French-English parallel treebank, showing that English adjuncts tend to be consistent with word alignments for machine translation and to be aligned to French adjunct-like constituents. Section 2 presents criteria to identify English adjuncts in phrase-structure parses and provides alignment measures for these adjuncts into French.

If adjuncts can be paired by word alignments, they can be deleted from or inserted in translation data, thus unfolding latent translation data. The resulting data can then be used to smooth the original distribution. In this work we start by investigating the effect of adjunct deletion, which is far simpler than adjunct insertion. On the linguistic side, adjunct insertion is complicated because modifiers are subjected by their heads to lexical and syntactical constraints, e.g., verbs take adverbial modifiers and nouns take adjectival modifiers. And on the computational side, adjunct deletion can be done in the phrase-based framework, whereas adjunct insertion requires a synchronous grammar with insertion/adjunction as operation, e.g., Synchronous Tree-Adjoining Grammar, (Abeillé et al., 1990; Shieber, 2007). In section 3 we show how adjunct-pair deletion from parallel data allows us to generate more training data to smooth a PBSMT baseline. Section 4 then provides experimental results for the smoothed model. We conclude in Section 5.

2 Adjunct alignment between French and English

As an illustration for adjunct alignment, consider the sentence pair in Figure 1. The English sentence contains three adjuncts that are translated as adjuncts in the French sentence. The example shows that the paired adjuncts can be of a different syntactical nature, as well as the phrases they appear in. Here, *"governing existing vehicles"* is a verb phrase while *"pour les véhicules existants"* is a prepositional phrase; and *"there must be rules"* is only globally equivalent to *"il faut trouver des règles"*. In other words, adjunct pairing can occur relatively independently of the syntactical realization of the involved adjuncts and of the degree of translation equivalence of the phrases they modify.

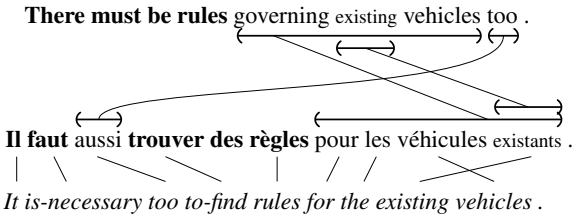

Figure 1: Example sentence pair with adjunct pairs

Conversely, adjuncts are not always preserved in translation. For instance, Example 1 presents a case of head swapping taken from (Dorr et al., 2002).

(1) *Yo entro el cuarto corriendo*
I enter the room running
I run into the room

There the manner of motion, i.e., 'running', is expressed by the verbal head in the English sentence and by a modifier in the Spanish sentence while the direction, i.e., 'into', is expressed by the head in Spanish and by a modifier in English. So, while (Dorr et al., 2002) investigated structural divergence in general and not only on modifiers, we can expect that adjuncts are not always translated as such in the target language.

Another limitation on adjunct alignment is not linguistic but technical. In fact, we depend on word alignments to align adjuncts into the target language. Consequently, to take the example of Figure 1, one can only know that the phrase *"pour les véhicules existants"* is paired with *"governing existing vehicles"* if the word alignments are able to align the semi-equivalent *'governing'* and *'pour'* properly. An unfavorable alignment in this case might align the English phrase with, e.g., *"les véhicules existants"*.

Finally, the method we follow to identify adjunct pairs in the data consists in first identifying English adjuncts, before aligning them to their French counterpart using word alignments. We identify adjuncts using a phrase-structure parser, which allows to quickly parse very large translation corpora, but does not directly annotate modifiers. Instead one can apply categorial and distributional criteria to identify constituents that are likely to be adjuncts. We present our identifica-

tion criteria in section 2.1, and adjunct alignment experiments and results in section 2.2.

2.1 Identifying adjuncts and adjunct pairs

The identification criteria for the English adjuncts are set by manually analysing the fifty first parses of the English Europarl corpus, parsed with the Charniak parser. Constituent categories that function as modifiers in most cases and given some distributional constraints are subsequently regarded as adjuncts. The identification criteria are summed up in Table 1. The tags are those of the Penn Treebank[1], except for NNx, which stands for NN(P)(S). The English adjuncts thus identified are paired by the GIZA++ word alignments to their French counterpart. The phrase pairs that are consistent with the word alignments are then assumed to be pairs of adjuncts.

category	parent	additional restriction
ADJP	NP	
JJ	NP	
NNx	NP	NN/NNS right sibling
VP	NP	
S	NP	
PP	$\neq$PP	
SBAR	$\neq$VP	
RB	$\neq$ADVP	
ADVP		
PRN		
NP		adposed: left and right comma

Table 1: English-adjuncts identification criteria

2.2 Adjunct alignment between English and French

To assess how well English adjuncts are aligned to French adjuncts, we analyzed adjunct alignment in a parallel treebank. The French treebank was obtained from the automatically annotated Europarl-section of the 'Arboratoire' treebank[2], and contains 30421 sentences and parses that roughly correspond to the beginning of the Europarl corpus. The English treebank was obtained from the English Europarl Corpus with the Charniak parser. After aligning both treebanks with the French and English corpora and with the GIZA++ word alignments trained on the whole corpus and merged with `'grow-diag-final'`, one obtains 13620 aligned parses, sentences and word alignments.

[1] `ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz`

[2] `http://corp.hum.sdu.dk/arboratoire.html`

For each English adjunct category, we aligned English adjuncts to their French counterparts, and measured the relative frequency of the following cases: (1) adjuncts pairs that are *not* consistent with the word alignments (nc/A) [3] ; (2) the French counterpart could not be located in the parse ($f_?$); (3) the French counterpart is not consistent with the parse (nc/P); (4) adjuncts aligned to the empty string ($f_\varnothing$); (5) the French counterpart is consistent with the French parse, i.e., it corresponds to one or more *complete* constituents (c/P). Measurements are reported in Table 2, along with the average number of English adjuncts per sentence (r), and upper bounds (UB) for adjunct alignment into French.

	r	nc/A (%)	$f_?$ (%)	nc/P (%)	$f_\varnothing$ (%)	c/P (%)	UB (%)
JJ	0.98	18.4	0.4	3.0	3.5	74.7	78.2
RB	0.31	35.1	0.8	3.1	5.1	55.9	61.0
ADVP	0.63	24.0	0.9	4.9	6.4	63.9	70.3
SBAR	0.41	26.2	0.9	6.5	0.5	66.0	66.5
S	0.03	27.3	0.9	6.4	0.6	64.8	65.4
VP	0.09	30.4	0.8	6.5	0.9	61.5	62.4
PP	2.05	23.7	0.8	9.0	1.4	65.0	66.4
NNx	0.28	22.9	0.5	9.3	1.9	65.5	67.4
ADJP	0.10	26.9	0.7	9.1	1.0	62.3	63.3
PRN	0.04	19.3	4.3	10.6	5.9	59.9	65.8
NP	0.03	11.5	1.7	17.2	0.7	68.8	69.5

Table 2: English-French adjunct alignment

Depending on the category, 11.5% to 35.1% of the English adjuncts lead to a phrase pair that is not consistent with the word alignments. Low alignment-consistency for the RB category is due in part to discontinuous alignments as, e.g., *'not'/'ne ... pas'*. A second informative measure for adjunct alignment is the proportion of aligned French phrases that are not consistent with the French parse, i.e., that fall across constituent boundaries. The worse results are obtained for the parenthetical PRN and adposed NP's and are caused by the the lack of punctuation handling of the French parse, which results in wrong attachments. The latter issue also concerns the categories PP, ADJP and NNx . Consequently, figures for these categories can be partially imputed to parsing quality. What remains are English adjuncts aligning to zero, one or more French constituents, with

[3] A phrase pair is consistent with word alignments iff none of the words in one of the two phrases is aligned outside the other phrase, see also (Koehn et al., 2003).

figures varying between 61.0% for RB and 78.2% for JJ. We interpret these figures as an upper bound on adjunct alignment under word alignments, and with the restriction that our identification criteria also include a portion of false adjuncts.

To try and answer how often aligned, parse-consistent French constituents are also adjuncts, we then looked at their categorial tag(s). Depending on the category, English adjuncts have 52 to 1952 different projections on the French side. This has to do with the number of tags used by the Arboratoire treebank, 37, all of which but one appear in the projections, in combination with a flat parse structure. To simplify the analysis, we only looked at the three most frequent projections for each category. Results are displayed in Table 3, showing a fairly high dispersion of the projections: in the worst case of VP, the three first projections cover only 38.6% of all 229 cases. In the best case with RB, 84.1% of adjuncts are covered.

category	three most frequent projections (%)			LB
JJ	ADJ - 69.0	N - 10.1	NUM - 3.0	64.8
RB	ADV - 78.5	ADJ - 3.5	N - 2.1	51.1
ADVP	ADV - 60.4	PP - 5.8	ADJ - 3.0	50.6
SBAR	FCL - 35.1	PP - 6.2	NP - 4.3	30.6
S	PP - 45.7	ICL - 6.3	PRP ICL - 2.2	35.7
VP	ICL - 18.6	FCL - 10.2	V-PCP2 PP- 9.8	30.2
PP	PP - 55.1	PP PP - 7.0	NP - 3.5	44.0
NNx	N - 35.2	ADJ - 26.6	PP - 14.5	28.8
ADJP	ADJP - 29.7	PAR - 9.4	N ADJ - 6.6	29.5
PRN	NUM - 29.9	N - 11.6	NP - 10.5	37.0
NP	NP - 28.6	PROP - 24.6	PROP NP - 11.2	45.0

Table 3: Most frequent French projections

The most frequent projections illustrate that English adjunct constituents tend to be aligned to French constituents of comparable nature. The only noticeable anomaly is that of NNx aligning to N in French. French uses much less nominal qualifiers than English, and a closer look reveals that in most cases, the NNx constituent was translated by a PP modifier in French, but that the word alignments aligned it to the PP's nominal constituent, instead of the entire PP.

With the exclusion of the NNx→N derivation, taking the proportion of parse-consistent French constituents with the three most frequent projections, and adding it to the proportion of null-aligned English adjuncts gives a lower bound (LB) on adjunct alignment. The resulting lower-bound figures displayed in Table 3 could be much refined. On one hand, we assume here, based on a succinct qualitative analysis of the data, that all first three projections, excepted NNx→N, actually concern French adjuncts; on the other hand, considering more projections for each category is bound to increase figures.

3 Smoothing a PBSMT model by factoring out adjuncts

Figure 2 shows a schematic view of the procedure to smooth a phrase-based model by adjunct-pair deletion. We train a baseline using the Moses toolkit (Koehn et al., 2007). Besides, the training data and the word alignments trained on this data are used to generate new training data by adjunct-pair deletion; Section 3.1 explains how this is done. We then execute part of the Moses training to extract and score phrase pairs from the generated data. Finally, the resulting model is interpolated with the baseline as explained in section 3.2.

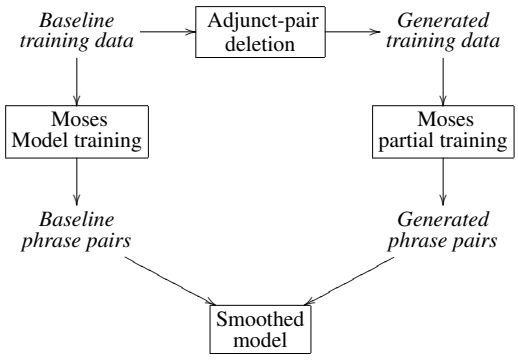

Figure 2: Building a smoothed model

3.1 Training-data generation

We identified 4.9M English adjuncts in the 0.95M parsed sentences of the English Europarl corpus, 3.7M of which lead to consistent phrase pairs. For each sentence pair, we try to generate as many sentence pairs and associated word alignments as there are combinations of adjunct pairs. Data growth is then exponential, and we obtain 95M possible adjunct combinations, though more than half of these contain overlapping adjuncts. To further limit the amount of generated data, combinations are filtered based on the distance between adjuncts. This filtering is combined with measures to control the quality of the generated data. These measures lead to the generation of 9.4M sentence pairs, which can be further brought down by a language-model filter. Next we flush out the details of the filtering methods and explain how we interpolate the model trained on the original data

with the model trained on the thus generated data.

Distance-based filtering

Deleting adjunct-pair combinations allows to obtain more phrase pairs than would be possible by deleting adjunct pairs separately. However, as phrase length is typically limited in phrase-based models, there is no benefit in deleting combinations of distant adjunct pairs. We therefore only considered combinations in which all English adjuncts are separated by less than $l_M - 1$ tokens, where l_M is the maximum phrase length. Note that using only the distance between English adjuncts relies on the assumption that French adjuncts will be distant if their English counterparts are.

Adjunct-gap junction correction

As adjuncts can be marked typographically, typically by surrounding commas, we try to prevent adjunct deletion from resulting in incorrect sequences of punctuation marks. A sequence of at least two of the following tokens is considered incorrect, as is the occurrence of any of these tokens at the start of a sentence:

, . : ; ? ! - .. ...

We try to remove misplaced punctuation marks as follows: if punctuation is aligned to the empty string or if it is aligned to something together with some other token, then it is deleted; if punctuation is aligned to a punctuation mark, and no other token is aligned to that punctuation mark, then punctuation is deleted on both sides. Sentence pairs that contain sequences of incorrigible punctuation marks are discarded.

Conversely, one also uses punctuation to try and increase the number of potentially interesting adjunct pairs: If a given adjunct pair is found not to be consistent with word alignments, one tries to extend it to adjoining punctuation.

A second measure aiming at improving the quality of the generated data consists in ensuring that if an English adjunct is deleted just after the indefinite article *'a'/'an'*, the form of the article is modified to account for the first letter of the new following word: *'a'* is changed to *'an'* if it is now followed by a vowel, and likewise for the form *'an'*.

Language-model filter

A final filtering measure consists in comparing the language-model probability P_{LM} of each generated French sentence f and English sentence e with that of the French sentence f_0 and the English sentence e_0 they are generated from. Sentences are corrected for length, and an additional threshold k is used to control the amount of generated data. Accordingly, only those sentence pairs that satisfy the following equations are actually generated:

$$P_{LM}(e)^{1/|e|} \geq k \cdot P_{LM}(e_0)^{1/|e_0|} \quad (1)$$

$$P_{LM}(f)^{1/|f|} \geq k \cdot P_{LM}(f_0)^{1/|f_0|} \quad (2)$$

3.2 Model smoothing

The baseline's translation model is smoothed by linear interpolation with the model trained on the generated data, following Equation 3.

$$\phi_I(\bar{s}|\bar{t}) = \lambda\phi_B(\bar{s}|\bar{t}) + (1-\lambda)\phi_A(\bar{s}|\bar{t}) \quad (3)$$

where $\phi_B(\bar{s}|\bar{t})$ and $\phi_A(\bar{s}|\bar{t})$ are the translation probability distributions in the baseline and the new model, respectively. The probability distributions are normalized to ensure model consistency [4].

We used either a constant interpolation parameter λ or one inspired from the Good-Turing estimate. In this case, the probability mass allocated to the probability distributions $\phi_A(\bullet, \bar{t})$ increases with the relative frequency of single-occurrence phrase pairs with a constituent $\bar{t}$. The interpolation parameter $\lambda(\bar{t})$ is defined by:

$$1 - \lambda(\bar{t}) = \frac{n1}{n1 + N} \quad (4)$$

where $n1$ is the count of single-occurrence phrase pairs, and N the total count of phrase pairs with a constituent target phrase $\bar{t}$. As most target constituent phrases in the baseline are associated with singleton phrase pairs, adding n1 to the denominator of Equation 4 ensures that $1 - \lambda(\bar{t})$ never reaches 1. To prevent the opposite, $1 - \lambda(\bar{t})$ is set to 10^{-4} by default.

The phrase-pair tables contain both translation probability estimates conditioned on the target phrases, and inverse translation probability estimates conditioned on the source phrases. Interpolation is performed for both distributions.

Probabilities in the reordering model are estimated individually for each phrase pair, consequently one can directly enrich the reordering table with the new model's table without smoothing. The enriched reordering model consists therefore of the baseline model and of the new model's reordering probabilities for the phrase pairs in $A-B$.

[4] The normalization factor is λ, 1 or $1 - \lambda$, depending on whether the conditioned target phrase is known to the baseline model only, to both models, or to the new model only.

4 Experiments

The basic set-up for the experiments uses the 2007 Workshop on Machine Translation (WMT07) baseline's training data. The generated training data is obtained with a language-model filter threshold $k = 0.7$, yielding 4M sentence pairs. Models are built to decode from French to English. The tuning parameters of the baseline are re-used for the smoothed models.

We used four test sets: the in-domain WMT07 test set `devtest`, the out-of-domain WMT07 test set `nc-test`, an adjunct-poor test set `adjpoor` and a second out-of-domain test set `hansards` derived from the Hansards corpus.

The `adjpoor` test set is derived from `devtest` by adjunct-pair deletion, following the same procedure as for the training data: The new test set contains the sentence pairs that are generated by removing combinations of adjunct pairs in `devtest`, without replication of the original sentence pairs. The language-model threshold is set to 1.0 in order to enhance the quality of the generated sentence pairs while limiting their number. The resulting test set consists of 8586 sentence pairs. While not all sentence pairs are equally grammatical, the test set allows to compare the performance of the generated models and of the baseline on adjunct-poor data.

The `hansards` test set consists of the 2000 first non-comment sentence pairs of the Hansards' House Debates Test Set, where non-comment sentence pairs are defined as ones for which the English sentence ends with a period. The selected sentence pairs are tokenized and lowercased as for the WMT07 test sets.

4.1 Results

Table 4 reports the BLEU scores obtained by the baseline and the smoothed models when varying the amount of generated data with the language-model filter, and using two interpolation parameters, $\lambda = 0.999$ or the Good-Turing inspired λ_{GT}.

The smoothed models perform only slightly better than the baseline on the in-domain test set `devtest`, but significantly [5] on the `adjpoor` test set. We found that giving more weight to the generated data, using $\lambda = 0.99$ and $\lambda = 0.9$, decreased model performance. The benefit of generating more training data is seen best on `adjpoor` with λ_{GT}. In the rest of this document, results are reported for a model smoothed with λ_{GT}.

	devtest	adjpoor	nc-test	hansards
baseline	32.47	33.18	24.41	22.24
TD$_G = 1M$:				
$\lambda = 0.999$	32.47	33.31	24.42	22.18
λ_{GT}	32.50	33.30	24.44	22.09
TD$_G = 4M$:				
$\lambda = 0.999$	32.52	33.35	24.38	22.16
λ_{GT}	32.51	33.49	24.42	22.12

Table 4: BLEU scores in basic set-up

Table 5 reports the BLEU scores obtained by the baseline and a smoothed model with λ_{GT} when trained on the first 10000 sentence pairs of the normal training set. With a small training set, the smoothed models still perform but slightly better than the baseline. However, they now fail to outperform the baseline on the `adjpoor` test set.

	devtest	adjpoor	nc-test	hansards
baseline	25.94	26.24	15.77	16.56
smoothed	25.97	26.16	15.67	16.66

Table 5: BLEU scores with a small training set

To assess whether the lack of improvement could be related to the asymmetry of the adjunct-deletion process, we used the generated data to smooth an English-to-French model, but this provided inconclusive again.

We found that the smoothed model could perform significantly better than the baseline, both on `devtest` and `adjpoor`, if one uses the tuning parameters of the smoothed model instead of those of the baseline. Results are given in Table 6.

	devtest	adjpoor	nc-test	hansards
baseline	32.31	33.56	24.55	22.27
smoothed	32.50	33.79	24.46	22.27

Table 6: Effect of retuning

The language model used by the decoder is trained on the training data of the baseline, and as such it may penalize new phrase pairs. As a last experiment, we interpolated a language model

[5] Significance was measured at $p = 0.05$ through approximate randomization, using FastMtEval: `http://www.computing.dcu.ie/~nstroppa/softs/fast_mt_eval.tgz`.

trained on the baseline data with one trained on the generated data, with an interpolation parameter value of 0.999. We did not retune the model, but re-used instead the tuning parameters of the baseline with a language model trained on it. Results are reported in Table 7. These results are nearly

	devtest	adjpoor	nc-test	hansards
baseline	32.47	33.18	24.42	22.23
smoothed	32.52	33.50	24.40	22.10

Table 7: Effect of an interpolated language model

identical to those of Table 4, indicating that there is no benefit in using an interpolated language model.

4.2 Results Analysis

To understand why the smoothed model shows only a minor improvement over the baseline, we looked at the repartition of new phrase pairs at different stages of decoding, and we measured the proportion of test sentences affected by smoothing.

Model contents While the deletion of adjunct pairs allows to generate many new phrase pairs, only few of them are selected by the decoder. Table 8 gives the size of the smoothed model and the repartition of its phrase pairs at three stages: in the training table, in the test-set filtered table, and in the phrase pairs used by the decoder. Phrase pairs are partioned in the following categories: phrase pairs contained in the baseline's training data only; phrase pairs contained both in the baseline's training data and the generated data; generated phrase pairs providing new translation options for source phrases that are known to the baseline; generated phrase pairs containing a source phrase unknown to the baseline.

	table size	base. only (%)	shared (%)	trans. options (%)	new input (%)
training	67.1M	10.4	52.0	7.2	30.4
filtered	4.84M	10.0	72.9	17.0	0.2
decoding	26.7k	1.4	98.4	0.0	0.2

Table 8: Model contents

When tables are filtered for decoding, the proportion of phrase pairs providing new input phrases shrinks, showing that the smoothed model brings proportionally little input phrase pairs that match the test data. Nearly all phrase pairs used for decoding are shared by the baseline and the generated table, while none of the generated phrase pairs with new translation options are used. It may be interesting to note that regardless of their origin, all the phrase pairs used at decoding have a target constituent that is used both by baseline and generated phrase pairs. Consequently, even when a generated phrase pair with a new input is used, it provides the system with an existing translation option.

Effect on output translation As the contribution of the enriched models in terms of phrase pairs is minimal, it is interesting to see how many output sentences actually differ from the baseline. Table 9 gives the number of sentences with a different translation and the associated BLEU scores for each test-set in the basic set-up. When translation output is identical, one distinguishes sentences with an identical or a different segmentation.

	devtest	adjpoor	nc-test	hansards
$\neq$ translation	645	3722	687	597
BLEU base	29.73	29.71	22.95	20.77
BLEU smoothed	29.81	30.31	22.99	20.44
$\neq$ segmentation	488	2504	440	460
$=$ segmentation	867	2360	880	943
BLEU	34.88	36.38	26.10	23.84

Table 9: Effect of the models on output translation

Table 9 shows that although the enriched model contributes few new phrase pairs, output translation is different for 30% to 43% sentences, indicating that the smoothed probability estimates lead to a different choice of output phrases. This is also reflected by the number of identical translations with a different segmentation (22% to 29%). Note that differences seem very localized, as they tend to concern sequences of two phrases only.

If one only considers different translations, the improvement of the smoothed model over the baseline on `devtest` is slightly higher than overall, but still not significant. It does however indicate that smoothing helps to improve results.

5 Conclusion

We presented projection figures for English adjuncts into French adjunct-like categories, reporting upper-bound values varying between 61.0% to 78.2% depending on the adjunct category, and lower-bound values between 28.8% and 64.8%.

Besides, we presented a novel way of enriching a PBSMT model by factoring out adjuncts. We

found that a model enriched in this manner only leads to a minor improvement over the baseline. Our system could be improved, notably by extending the class of adjuncts to account for other optional constituents that do not have the status of modifiers, e.g., coordinated elements.

However the main hurdle for our system is that one can only remove adjuncts, and not add any. Consequently, our system performs best on adjunct-poor data, but that is not generally the nature of translation data. Therefore we think that it would be interesting to use adjuncts as a label in a basic SCFG as that of Chiang (2005).

Finally, it would be interesting to investigate the effect of adjunct-pair deletion on other language pairs. While we relied on structural similarity between French and English to align adjuncts, the notion of adjunct is not only syntactical but also has semantic, and therefore cross-linguistic value. Future research might tell whether there is more to gain from adjunct-pair deletion on language pairs that are harder to translate.

References

Abeillé, Anne and Yves Schabes and Aravind K. Joshi. 1990. Using Lexicalized Tags for Machine Translation. *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, Finland, 1–6.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2), 263–311.

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

Dorr, Bonnie J., Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment. *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, Tiburon, CA, 31–43.

El Kholy, Ahmed and Nizar Habash. 2010. Orthographic and morphological processing for English-Arabic statistical machine translation. *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montreal, Canada.

Foster, George, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 53–61.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating Translational Correspondence using Annotation Projection. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, 392–399.

Joshi, Aravind K. 1983. Recursion and Dependencies: an Aspect of Tree Adjoining Grammars (TAG) and a Comparison of Some Formal Properties of TAGs, GPSGs, PLGs, and LPGS. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, 7–15.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. *Proceedings of the Human Language Technology conference - North American chapter of the Association for Computational Linguistics annual meeting*, Edmonton, Canada, 127–133.

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

Kuhn, Roland, George Foster, Samuel Larkin, and Nicola Ueffing. 2006. PORTAGE Phrase-Based System for Chinese-to-English Translation. *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 75–80.

Nießen, Sonja and Hermann Ney. 2004 Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics* 30(2), 181–204.

Quirk, Chris and Arul Menezes. 2006. Do we need phrases? Challenging the conventional wisdom in Statistical Machine Translation. *Proceedings of the Human Language Technology conference - North American chapter of the Association for Computational Linguistics annual meeting*, New York, NY, 9–16.

Shieber, Stuart M. 2007. Probabilistic Synchronous Tree-Adjoining Grammars for Machine Translation: The Argument from Bilingual Dictionaries. *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, Rochester, NY, 88–95.

Schwenk, Holger, Marta R. Costa-Jussà, and José A.R. Fonollosa. 2007. Smooth Bilingual N-gram Translation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, 430–438.