Soft Dependency Constraints for Reordering in Hierarchical Phrase-Based Translation

Yang Gao, Philipp Koehn and Alexandra Birch
School of Informatics
University of Edinburgh
Edinburgh, UK, EH8 9AB
yanggao1119@gmail.com, pkoehn@inf.ed.ac.uk, a.birch@ed.ac.uk

Abstract

Long-distance reordering remains one of the biggest challenges facing machine translation. We derive soft constraints from the source dependency parsing to directly address the reordering problem for the hierarchical phrase-based model. Our approach significantly improves Chinese–English machine translation on a large-scale task by 0.84 BLEU points on average. Moreover, when we switch the tuning function from BLEU to the LRscore which promotes reordering, we observe total improvements of 1.21 BLEU, 1.30 LRscore and 3.36 TER over the baseline. On average our approach improves reordering precision and recall by 6.9 and 0.3 absolute points, respectively, and is found to be especially effective for long-distance reodering.

1 Introduction

Reordering, especially movement over longer distances, continues to be a hard problem in statistical machine translation. It motivates much of the recent work on tree-based translation models, such as the hierarchical phrase-based model (Chiang, 2007) which extends the phrase-based model (Koehn et al., 2003) by allowing the so-called *hierarchical phrases* containing subphrases.

The hierarchical phrase-based model captures the recursiveness of language without relying on syntactic annotation, and promises better reordering than the phrase-based model. However, Birch et al. (2009) find that although the hierarchical phrase-based model outperforms the phrase-based model in terms of medium-range reordering, it does equally poorly in long-distance reordering due to constraints to guarantee efficiency.

Syntax-based models that use phrase structure constituent labels as non-terminals in their transfer rules, exemplified by that of Galley et al. (2004), produce smarter and syntactically motivated reordering. However, when working with off-the-shelf tools for parsing and alignment, this approach may impose harsh limits on rule extraction and requires serious efforts of optimization (Wang et al., 2010).

An alternative approach is to augment the general hierarchical phrase-based model with soft syntactic constraints. Here, we derive three word-based, complementary constraints from the source dependency parsing, including:

- A dependency orientation feature, trained with maximum entropy on the word-aligned parallel data, which directly models the head-dependent orientation for source words;
- An integer-valued cohesion penalty that complements the dependency orientation feature, and fires when a word is not translated with its head. It measures derivation well-formedness and is used to indirectly help reordering;
- An auxiliary unaligned penalty feature that mitigates search error given the other two features.

We achieve significant improvements in terms of the overall translation quality and reordering behavior. To our knowledge we are the first to use the source dependency parsing to target the reordering problem for hierarchical phrase-based MT.

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 857–868,
Edinburgh, Scotland, UK, July 27–31, 2011.

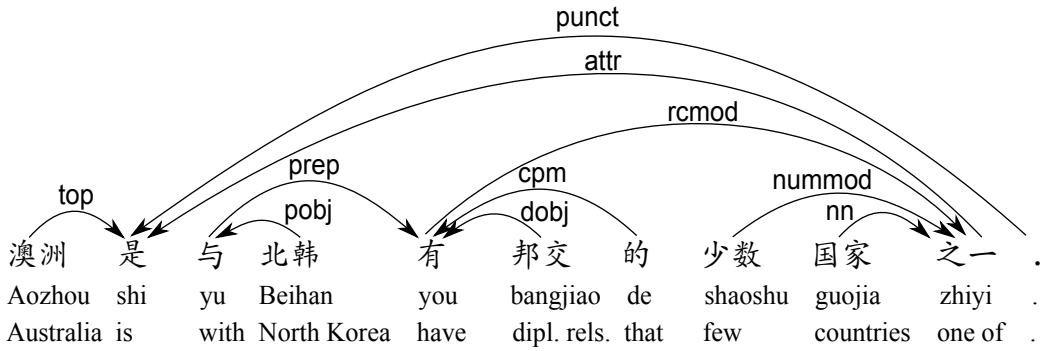


Figure 1: Example dependency parsing generated by the Stanford Parser. The Chinese source sentence and its English translation come from (Chiang, 2007).

2 Three Soft Dependency Constraints

Our features are based on the source dependency parsing, as shown in Figure 1. The basic unit of dependency parsing is a triple consisting of the dependent word, the head word and the dependency relation that connects them. For example, in Figure 1, an arrow labelled prep goes from the word *yu* (English *with*) to the word *you* (English *have*), showing that *yu* is a prepositional modifier of *you*.

We use the Stanford Parser[1] to generate dependency parsing, which automatically extracts dependency relations from phrase structure parsing (de Marneffe et al., 2006).

2.1 Dependency Orientation

Based on the assumption that constituents generally move as a whole (Quirk et al., 2005), we decompose the sentence reordering probability into the reordering probability for each aligned source word with respect to its head, excluding the root word at the top of the dependency hierarchy which does not have a head word. Similarly, Hayashi et al. (2010) also take a word-based reordering approach for HPBMT, but they model *all possible* pairwise orientation from the source side as a general linear ordering problem (Tromble and Eisner, 2009).

To be more specific, we have a maximum entropy orientation classifier that predicts the probability of a source word being translated in a monotone or reversed manner with respect to its head. For example, given the alignment in Figure 2(a), with the alignment points (i_{dep}, j_{dep}) for the source dependent word and (i_{head}, j_{head}) for the source head word, we define two orientation classes as:

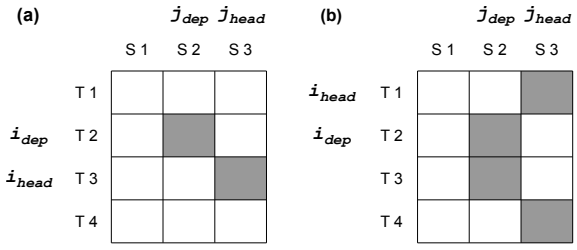


Figure 2: Word alignments to illustrate orientation classification. In (a), monotone (M); in (b), reversed (R).

$$c = \begin{cases} R & \text{if } (j_{dep} - j_{head})(i_{dep} - i_{head}) < 0 \\ M & \text{otherwise} \end{cases} \quad (1)$$

When a source head or dependent word is aligned to multiple target words, as shown in Figure 2(b), we always take the first target word for orientation classification.

The orientation classifier is trained on the large word-aligned parallel corpus. Various features can potentially be used, based on the source and target context as well as syntactic and semantic analysis. The orientation probability is evaluated in the following log-linear equation, where f is the source context, d is the source dependency parsing, e^* is the target context produced so far, a^* is the alignment produced so far and c is the orientation class:

[1] http://nlp.stanford.edu/software/lex-parser.shtml

Word	p(M)	p(R)
Aozhou	0.81	0.19
shi	NA	NA
yu	0.45	0.55
Beihan	0.88	0.12
you	0.12	0.88
bangjiao	0.83	0.17
de	0.58	0.42
shaoshu	0.30	0.70
guojia	0.19	0.81
zhiyi	0.85	0.15
.	1.00	0.00

Table 1: The dependency orientation probabilities for words of the Figure 1 sentence, in both monotone and reversed cases.

$$p(c|f,d,e^*,a^*) = \frac{exp(\sum_{n=1}^{N}\lambda_n h_n(f,d,e^*,a^*,c))}{\sum_{c'\in\{M,R\}} exp(\sum_{n=1}^{N}\lambda_n h_n(f,d,e^*,a^*,c'))} \quad (2)$$

Currently, we only use two kinds of features: (1) the concatenation of the source dependent word with the dependency relation and (2) the concatenation of the source head word with the dependency relation. So for the word *yu* (English *with*) in Figure 1, we extract these features for orientation classification: prep_DEP_*yu* and prep_HEAD_*you*.

We define the dependency orientation feature score for a translation hypothesis as the sum of the log orientation probabilities for each source word. This score is used as one feature in the log-linear formulation of the hierarchical phrase-based model.

Table 1 shows the dependency orientation probabilities for all words in the Figure 1 sentence. Most interestingly, the orientation probabilities for *you* (English *have*) strongly support global reordering of *one of the few countries* with the relative clause *that have diplomatic relations with North Korea*. We find that it is a general trend for long-distance reordering to gain stronger support, since it is often correlated with prominent reordering patterns (such as relative clause and preposition) as well as lexical evidences (such as *"... zhiyi"* (English *"one of ..."*)) for which the reversed orientation takes up the majority of the training cases.

Consider the following rules (both terminals and nonterminals are coindexed):

$$X \rightarrow (\underline{yu_1\ Beihan_2}\ you_3\ bangjiao_4,\ have_3\ dipl._4\ rels._4\ \underline{with_1\ North_2\ Korea_2}) \quad (3)$$

$$X \rightarrow (\underline{yu_1\ Beihan_2}\ you_3\ bangjiao_4,\ \underline{with_1\ North_2\ Korea_2}\ have_3\ dipl._4\ rels._4) \quad (4)$$

According to Table 1, the hypothesis that applies Rule 3 receives a probability of 0.55 for *yu* getting reversed with its head *you*, as well as 0.88 and 0.83 for translating *Beihan* and *bangjiao* in a monotone manner with respect to their heads. Rule 4 is associated with probabilities 0.45, 0.88 and 0.83 for monotone translation of *yu*, *Beihan* and *bangjiao*. Thus our dependency orientation feature is able to trace the difference in ordering the PP *with North Korea* (as underlined) and the VP *have dipl. rels.* down to the orientation of the preposition *yu* (English *with*) with respect to its head *you* (English *have*), and promote Rule 3 which has the right word order.

The word *you* (English *have*) cannot be scored in Rules 3 or 4, since its head word *zhiyi* (English *one of*) is not covered. In this case, we say that the word *you* is unresolved. We carry an unresolved word along in the derivation process until we reach a terminator hypothesis which translates the head word. Then the resulting dependency orientation score is added to the terminator hypothesis. This means that the dependency orientation feature is "stateless", i.e., hypotheses that cover the same source span with the same orientation information will receive the same feature score, regardless of the derivation history. Therefore, Derivation 5 in the following will have the same dependency orientation score as Derivation (Rule) 3, and Derivation 6 will score the same as Derivation (Rule) 4.

5.1 $X \rightarrow (yu_1, with_1)$

5.2 $X \rightarrow (Beihan_1, North_1\ Korea_1)$

5.3 $X \rightarrow (X_1\ X_2, X_1\ X_2)$

5.4 $X \rightarrow (X_1\ you_2\ bangjiao_3,\ have_2\ dipl._3\ rels._3\ X_1)$ (5)

6.1 $X \rightarrow$ (*Beihan*$_1$ *you*$_2$, *North*$_1$ *Korea*$_1$ *has*$_2$)

6.2 $X \rightarrow$ (X_1 *bangjiao*$_2$, X_1 *dipl.*$_2$ *rels.*$_2$)

6.3 $X \rightarrow$ (*yu*$_1$ X_2, *with*$_1$ X_2)

(6)

2.2 Cohesion Penalty

When the dependency orientation for a word is temporarily unavailable ("unresolved"), a cohesion penalty fires. Cohesion penalty counts the total occurrences of unresolved words for a translation hypothesis, which involve newly encountered unresolved words as well as old unresolved words carried on from the derivation history. Therefore, the cohesion penalty is "stateful", i.e., an unresolved word is repeatedly penalized until it gets resolved. Under this definition, the most cohesive derivation translates the entire sentence with one rule, where every word is locally resolved. The least cohesive derivation translates each word individually and glues word translations together. Consulting Figure 1, the cohesion penalty in Derivation 5 is 4, since the word *yu* (English *with*) is unresolved twice (in 5.1 and 5.3), and both *Beihan* (English *North Korea*) and *you* (English *have*) are unresolved once (in 5.2 and 5.4, respectively); the cohesion penalty in Derivation 6 is 5: 2 from *Beihan* (English *North Korea*) (in 6.1 and 6.2) and 3 from *you* (English *have*). As a result, Derivation 5 gets promoted, which echoes with human intuition since Derivation 5 translates syntactic constituents. To sum up, our cohesion penalty provides an integer-valued measure of derivation well-formedness in the hierarchical phrase-based MT. Same as dependency orientation, the cohesion penalty is not applicable to the root word of the sentence.

We propose the cohesion penalty in order to further improve reordering, especially in long-distance cases, since a well-formed derivation at an earlier stage makes it more likely to explore hierarchical rules that perform more reliable reordering. In this respect, the cohesion penalty can be seen as an aid to the glue rule penalty and as an alternative to constituency-based constraints.

Specifically, the glue rule penalty (Chiang, 2007) promotes hierarchical rules. Hierarchical rules whose lexical evidence helps resolve words locally will also be favored by our cohesion penalty feature. However, ignorant of the syntactic structure, the glue rule penalty may penalize a reasonably cohesive derivation such as Derivation 5 and at the same time promote a less cohesive hierarchical translation, such as Derivation 6.

Compared with constituency constraints based on the phrase structure, our cohesion penalty derived from the binary dependency parsing has two different characteristics.

First, our cohesion penalty is by nature more tolerant to some meaningful noncontituent translations. For example, constituency constraints in (Chiang, 2005; Marton and Resnik, 2008; Chiang et al., 2009) would penalize Rule 7 below which is useful for German–English translation (Koehn et al., 2003), and Rule 8 which can be applied to the Figure 1 sentence. Fuzzy constituency constraints can solve this problem with a combination of product categories and slash categories (Chiang, 2010). Yet our cohesion penalty by nature admits these translations as cohesive (with no extra cost from *es* and *Aozhou* since both are locally resolved). Admittedly, our current implementation of the cohesion penalty is blind to some other meaningful nonconstituent collocations, such as neighbouring siblings of a common uncovered head (regulated as the "floating structure" in (Shen et al., 2008)). A concrete example is Rule 9 which is useful for the Figure 1 sentence. To address this problem, another feature can be defined in the same manner to capture how each head word is translated with its children.

$$X \rightarrow (es_1\ gibt_2,\ there_1\ is_2) \quad (7)$$

$$X \rightarrow (Aozhou_1\ shi_2,\ Australia_1\ is_2) \quad (8)$$

$$X \rightarrow (shaoshu_1\ guojia_2,\ few_1\ countries_2) \quad (9)$$

Second, our cohesion penalty can be by nature more discriminative. Compared with the constituency constraints, the cohesion penalty is integer-valued, and can be made sensitive to the depth of each word in the dependency hierarchy (see Section 2.4). Inspired by (Marton and Resnik, 2008; Chiang et al., 2009), the cohesion penalty could also be made sensitive to the dependency relation of each word. However, this drastically increases the number of features and requires a tuning algorithm which scales better to high-dimensional model spaces, such as MIRA (Watanabe et al., 2007; Chiang et al., 2008).

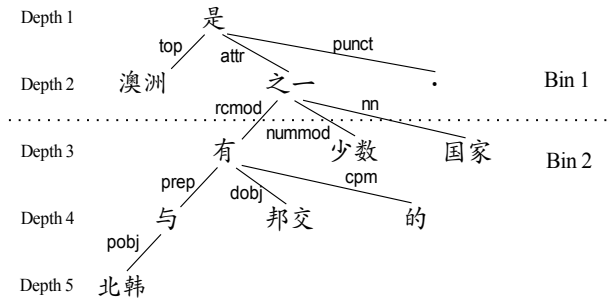


Figure 3: Using 2 bins for the dependency parse tree of the Figure 1 sentence.

2.3 Unaligned Penalty

The dependency orientation and cohesion penalty cannot be applied to unaligned source words. This may lead to search error, such as dropping (i.e., unaligning) key content words that are important for lexical translation and reordering. The problem is mitigated by an unaligned penalty applicable to all words in the dependency hierarchy.

2.4 Grouping Words into Bins

Having defined dependency orientation, cohesion penalty and unaligned penalty, we section the source dependency tree *uniformly by depth*, group words at different depths into bins and only add the feature scores of a word into its respective bin. In this way one feature is split into several sub-features and each can be trained discriminatively by MERT.

There are two motivations for binning. The primary motivation is to distinguish long-distance reordering which is still problematic for the hiero-style model, since local reorderings generally operate at low levels of the tree while high tree levels tend to take more care of long-distance reordering. Parsing accuracy is another concern, yet its impact on feature performance is intricate and our MaxEnt-trained dependency orientation feature also buffers against odd parsing. Using bins, we simply let the tuning process decide how much to trust feature scores coming from different levels of parsing.

We experiment with 1, 2 and 3 bins. An example of binning for the Figure 1 sentence can be found in Figure 3. With 2 bins (hereafter "bin-2"), words at Depth 1 and 2 are grouped into Bin 1, and words at Depth 3, 4, 5 are grouped into Bin 2. As a simple approach, binning does not take into account how the tree levels spread out.

3 Experiments

3.1 General Settings

We used a parallel training corpus with 2.1 million Chinese–English sentence pairs, aligned by GIZA++. The Chinese side was parsed by the Stanford Parser. Then we extracted 33.8 million examples from the parsed Chinese side to discriminatively train 1.1 million features (using the MegaM software[2]) for dependency orientation classification.

We trained three 5-gram language models with modified Kneser-Ney smoothing (Kneser and Ney, 1995): one on the English half of the parallel corpus, one on the Xinhua part of the Gigaword corpus, one on the AFP part, and interpolated them for best fit to the tuning set (Schwenk and Koehn, 2008).

We used NIST MT06 evaluation data (1664 lines) as our tuning set, and tested on NIST MT02 (878 lines), MT05 (1082 lines) and MT08 (1357 lines).

Our baseline system was the Moses implementation of the hierarchical phrase-based model with standard settings (Hoang et al., 2009). When only 1 bin was used, 3 additional features were added to the baseline, one each from the soft dependency constraints. When we used 2 or 3 bins, the additional feature counts doubled or tripled. We preserved terminal alignment alongside nonterminal alignment during the rule extraction and output word alignments together with translated strings. Since the features we currently define are based entirely on the source side, we used preprocessing to speed up decoding of our feature-augmented model. All experiments were tuned with MERT (Och, 2003).

3.2 Using BLEU as the Tuning Metric

As a standard practice, we first used BLEU (Papineni et al., 2002) as the objective function for tuning. Table 2 shows the results of the baseline model as well as our complete feature-augmented model with different bin numbers. With the "bin-2" setting, we get substantial improvement of up to 1.03 BLEU points (on MT02 data), and 0.84 BLEU points on average. Using more than one bin (i.e., differentiating tree depths) is generally beneficial, although the

[2]http://www.umiacs.umd.edu/∼hal/megam/index.html

Setting	BLEU / LRscore / TER			
	MT02	MT05	MT08	Average
baseline	34.01 / 41.85 / 68.93	32.23 / 40.50 / 68.15	28.09 / 37.17 / 66.82	31.44 / 39.84 / 67.97
bin-2	**35.04 / 43.07 / 65.58**	**33.18 / 41.62 / 65.59**	**28.63 / 38.12 / 65.36**	**32.28 / 40.94 / 65.51**
baseline-lr	34.23 / 42.06 / 68.08	32.28 / 40.61 / 67.61	27.99 / 37.27 / 66.98	31.50 / 39.98 / 67.56
bin-2-lr	**35.42 / 43.25 / 64.82**	**33.44 / 41.80 / 64.88**	**29.10 / 38.38 / 64.14**	**32.65 / 41.14 / 64.61**

Table 4: Results for the baseline model and the complete feature-augmented model with 2 bins ("bin-2"), using BLEU and LRscore ("-lr") as the tuning function. The BLEU scores of "bin-2" and "bin-2-lr" are significantly better than baseline ($p < 0.05$), computed by paired bootstrap resampling (Koehn, 2004).

Setting	BLEU			
	MT02	MT05	MT08	Average
baseline	34.01	32.23	28.09	31.44
bin-1	34.20	32.13	28.41	31.58(+.14)
bin-2	**35.04**	**33.18**	**28.63**	**32.28(+.84)**
bin-3	34.35	32.79	28.37	31.84(+.40)

Table 2: Results of the baseline model as well as our complete feature-augmented model with 1, 2 and 3 bins. BLEU is the tuning function.

Setting	BLEU			
	MT02	MT05	MT08	Average
baseline	34.01	32.23	28.09	31.44
dep	34.26	32.58	28.07	31.64(+.20)
dep+coP	34.47	32.81	28.61	31.96(+.52)
dep+coP+unP	35.04	33.18	28.63	32.28(+.84)

Table 3: Contributions of the three soft dependency constraints, with the "bin-2" setting

problem of overfitting sets in when we use 3 bins (with slightly higher tuning BLEU, not shown here).

We also studied the effect of adding features incrementally onto the baseline with the "bin-2" setting, as shown in Table 3. On average, all three features seem to have similar contributions.

3.3 Using LRscore as the Tuning Metric

Since our features are proposed to address the reordering problem and BLEU is not sensitive enough to reordering (especially in long-distance cases), we have also tried tuning with a metric that highlights reordering, i.e., the LRscore (Birch and Osborne, 2010). LRscore is a linear interpolation of a lexical metric and a reordering metric. We interpolated BLEU (as the lexical metric) with the Kendall's tau permutation distance (as the reordering metric). The Kendall's tau permutation distance measures the relative word order difference between the translation output and the reference(s) and is particularly sensitive to long-distance reordering. Testing results in terms of BLEU, LRscore and TER (Snover et al., 2006) are shown in Table 4. Tuned with the LRscore, our feature-augmented model achieves further average improvements (compare "bin-2" and "bin-2-lr") of 0.20 LRscore *as well as* 0.37 BLEU and 0.90 TER. Note that while the BLEU increase can largely be seen as a projection of the LRscore increase back into its lexical component, the consistent TER drop confirms that our improvement is not metric-specific[3]. Altogether the final improvement is 1.21 BLEU, 1.30 LRscore and 3.36 TER on average over the baseline.

However, an important question is how our features affect short, medium and long-distance reorderings. In the next section, we conduct quantitative analysis on reordering precision and recall, as well as qualitative analysis on translation examples.

4 Analysis

4.1 Precision and Recall of Reordering

The key to obtaining precision and recall for reordering is to investigate whether reorderings in the references are reproduced in the translations. We calculate precision as the number of reproduced reorderings divided by the total number of reorderings in the translation, and recall as the number of reproduced reorderings divided by the number of reorder-

[3]One of our reviewers points out that according to the inductive learning theory, it is counter-intuitive to improve on BLEU and TER if we optimize by the LRscore. Yet we do observe some other papers reporting increased TER or other metric scores when BLEU is used for tuning (Carpuat and Wu, 2007; Shen et al., 2008), suggesting that MT evaluation might be too complicated to be characterized just with inductive learning. Similar results based on extensive experiments can also be found in (Birch and Osborne, 2011).

Setting	MT02	MT05	MT08	Average
baseline	37.0	35.3	35.6	36.0
bin-2	42.7	40.8	38.7	40.7 (+4.7)
baseline-lr	37.3	35.0	34.2	35.5 (-0.5)
bin-2-lr	44.1	42.0	42.5	42.9 (+6.9)

Table 5: Overall precision for the test sets.

Setting	MT02	MT05	MT08	Average
baseline	37.5	36.2	33.2	35.6
bin-2	36.8	35.9	31.8	34.8 (-0.8)
baseline-lr	37.0	35.6	32.2	34.9 (-0.7)
bin-2-lr	37.7	36.7	33.2	35.9 (+0.3)

Table 6: Overall recall for the test sets.

ings in the reference. Then we average the precision and recall over all four reference translations.

Details of measuring reproduced reordering can be found in Birch et al. (2008). An important difference in this work is in handling many-to-one and one-to-many alignments, as we only retain the first word alignment for any source or target word which has multiple alignments. This is consistent with our treatment in dependency orientation classification, and results in more reorderings being extracted.

From Table 5 we can see that our features improve precision by an average of 4.7 absolute points when BLEU is used for tuning ("bin-2"). Switching from BLEU to the LRscore ("bin-2-lr"), we gain 2.2 points more and have a total improvement of 6.9 absolute points on average. This is a novel and important finding as we directly show that the quality of reordering has been improved.

From Table 6, we observe a small but consistent increase in recall with the "bin-2-lr" setting, averaging 0.3 absolute points. However, the drop of recall with the "bin-2" setting (by an average 0.8 points from the baseline) is unexpected. It seems that when applying our features alone, we are trading a small drop in recall for a large gain in precision.

In Figure 4 we break down the precision and recall statistics in MT08 by the reordering width on the source side. We find that our features consistently help precision over all word ranges, with more substantial improvement in the medium and long word ranges. When recall is concerned, our model does not help for short ranges of up to Width 4, but improves consistently for longer distance re-

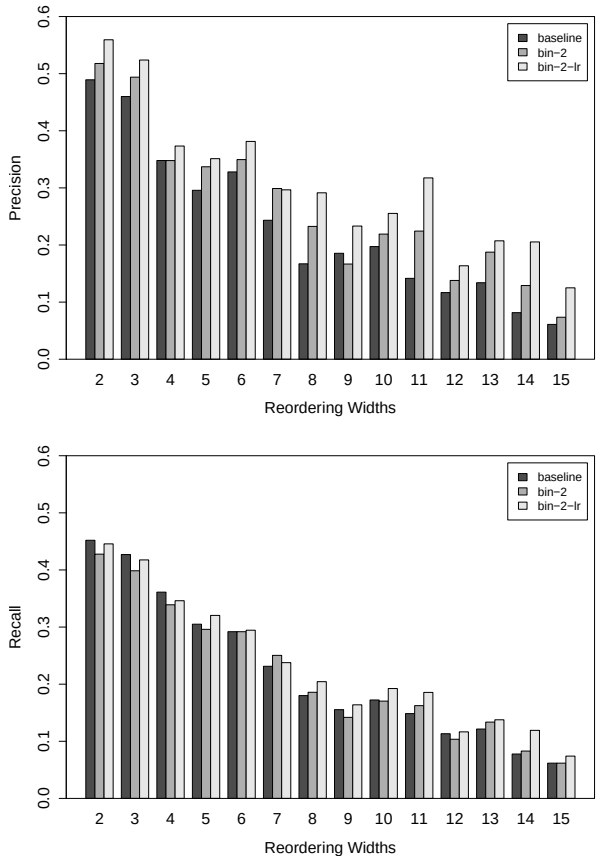
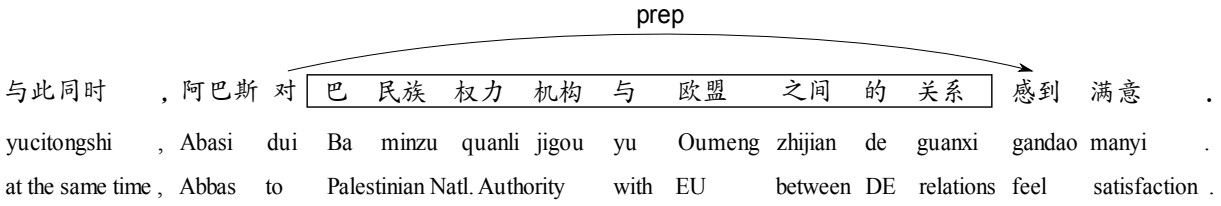


Figure 4: Precision and recall breakdown for the source-side reordering width 2-15 for the NIST MT08 dataset.

orderings. Once again, it seems that the feature-augmented model is able to benefit from tuning with a metric that is more sensitive to reordering, as the performance of "bin-2-lr" is the best in all reordering statistics.

4.2 Translation Examples

We observe a number of outputs with improved word order and more cohesive derivation, as the one in Figure 5. The baseline translation is fragmented and requires more glue rule applications. Specifically, it fails to translate the boxed area as a whole into "the relations between the palestinian national authority (pna) and the european union (eu)". The key dependency orientation that controls the global reordering is between the prepositional modifier *dui* (English *to*) and its head word, the verb *gandao* (English *feel*). The baseline system translates *dui* (English *to*) as "of the" and misorders the sentence. In contrast, the feature-augmented model "bin-2" cap-



baseline: [at the same time , abbas] [of the] [palestinian [national authority (pna)]] [and] [is satisfied with the [relations] [between [the european union (eu)]]] [.]

bin-2: [at the same time , abbas] [expressed satisfaction with [the relations between the [palestinian [national authority (pna) [and the european union (eu)]]]] .]

Figure 5: Example translations from the NIST MT08 set, output by the baseline model and "bin-2" model. The "-lr" version outputs are quite similar and not shown here. Translation outputs are in lower case.

tures the boxed area as a whole and uses Rule 10 to perform the right global reordering.

$$X \rightarrow (\underline{dui_1\ X_2}\ gandao_3\ manyi_4\ .5\ ,\ expressed_3\ satisfaction_4\ \underline{with_1\ X_2}\ .5\ ) \quad (10)$$

5 Related Work

In recent years, there has been a growing body of research on using dependency for statistical machine translation. Some directly encodes dependency in the translation model (Ding and Palmer, 2005; Quirk et al., 2005; Xiong et al., 2007; Shen et al., 2008; Mi and Liu, 2010), while others use dependency as a soft constraint (Cherry, 2008; Bach et al., 2009a,b; Chang et al., 2009). Among them, Shen et al. (2008) report that just filtering the phrase table by the so-called well-formed target dependency structure does not help, yet adding a target dependency language model improves performance significantly. Our intuitive interpretation is that the target dependency language model capitalizes on two characteristics of the dependency structure: it is based on words and it directly connects head and child. Therefore, the target dependency language model makes good use of the dependency representation as well as the target side training data.

We follow the second line of research, and derive three *word-based* soft constraints from the source dependency parsing. Note that although we reuse the word "cohesion" to name one of the constraints, our work is different from (Cherry, 2008; Bach et al., 2009a,b) which have successfully defined *another* cohesion constraint from the source dependency structure, with the aim of improving reordering in phrase-based MT.

To take a glance, Cherry (2008) and Bach et al. (2009b) define cohesion as translating a source dependency subtree contiguously into the target side without interruption (span or subtree overlapping), following Fox (2002). This span-based cohesion constraint has a different criterion from our word-based cohesion penalty and often leads to opposite conclusions. Bach et al. (2009a) also use cohesion to correlate with the lexicalized reordering model (Tillman, 2004; Koehn et al., 2005), whereas we define an orthogonal dependency orientation feature to explicitly model head-dependent reordering.

The fundamental difference, however, is rooted in the translation model. Their span-based cohesion constraint is implemented as an "interruption check" to encourage finishing a subtree before translating something else. This check is very effective for phrase-based decoding which searches over an entire space within the distortion limit in order to advance a hypothesis. In fact, it constrains reordering for the phrase-based model, as Cherry finds that the cohesion constraint is used "primarily to prevent distortion" and to provide "an intelligent estimate as to when source order must be respected" (Cherry, 2008). However, since the hierarchical phrase-based model *already* conducts principled reordering search with rules through the more constrained chart-decoding, ill-formed derivations exhibit themselves more often as nonconstituent translation than interrupted translation as defined in (Cherry, 2008; Bach et al., 2009a,b) (They do have a non-empty intersection, but neither subsumes the other). There-

fore, our cohesion penalty is better suited for the hierarchical phrase-based model.

To discourage nonconstituent translation, Chiang (2005) has proposed a constituency feature to examine whether a source rule span matches the source constituent as defined by phrase structure parsing. Finer-grained constituency constraints significantly improve hierarchical phrase-based MT when applied on the source side (Marton and Resnik, 2008; Chiang et al., 2009), or on the target side in a more tolerant fashion (Zollmann and Venugopal, 2006). Using both source and target syntax, but relaxing on rule extraction and substitution enables HPBMT to produce more well-formed and syntactically richer derivations (Chiang, 2010). Softening constituency matching with latent syntactic distributions proves to be helpful (Huang et al., 2010). Compared to constituency-based approaches, our cohesion penalty based on the dependency structure naturally supports constituent translations as well as some nonconstituent translations, if not all of them (as discussed in Section 2.2).

Our dependency orientation feature is similar to the order model within dependency treelet translation (Quirk et al., 2005). Yet instead of a head-relative position number for each modifier word, we simply predict the head-dependent orientation which is either monotone or reversed. Our coarser-grained approach is more robust from a machine learning perspective, yet still captures prominent and long-distance reordering patterns observed in Chinese–English (Wang et al., 2007), German–English (Collins et al., 2005), Japanese–English (Katz-Brown and Collins, 2008) and translation from English to a group of SOV languages (Xu et al., 2009). Not committed to specific language pairs, we learn orientation classification from the word-aligned parallel data through maximum entropy training as Zens and Ney (2006) and Chang et al. (2009) for phrase-based translation and Xiong et al. (2006) for the BTG model (Wu, 1996). While Chang et al. (2009) also make use of source dependency, their orientation classification concerns two subsequent phrase pairs in the left-to-right phrase-based decoding (as apposed to each dependent word and its head) and is therefore less linguistically-motivated.

6 Conclusion

We have derived three novel features from the source dependency structure for hierarchical phrase-based MT. They work as a whole to capitalize on two characteristics of the dependency representation: it is directly based on words and it directly connects head and child. The effectiveness of our approach has been demonstrated by a final average improvement of 1.21 BLEU, 1.30 LRscore and 3.36 TER. On average we improve reordering precision and recall by 6.9 and 0.3 absolute points, respectively, over the baseline. Moreover, our approach is found to be especially effective for long-distance reodering.

As mentioned in Section 2.2, the cohesion penalty can be extended to also account for how a head word is translated with its children so that we are not biased towards one form of cohesive nonconstituent translation. All our features can be made sensitive to the dependency relations or even words. This fine-grainedness is especially desirable when we want to reward words for being unaligned or unresolved, such as punctuations and function words in certain context. Word alignment quality is crucial for the performance of our features as well as the LRscore which uses word alignment to compute the permutation distance. As an alternative to GIZA++, we would like to experiment with syntactically informed aligners that better handle function words which often exhibit high alignment ambiguity due to low cross-lingual correspondence.

Finally, since our soft dependency constraints promote reordering without increasing model complexity, further gains can be achieved when combining our approach with orthogonal studies to improve the quantity and quality of hierarchical (reordering) rules, such as relaxing hierarchical rule extraction constraints (Setiawan and Resnik, 2010) and selectively lexicalizing rules with function words (Setiawan et al., 2009).

Acknowledgments

We would like to thank Miles Osborne, Adam Lopez, Barry Haddow, Hieu Hoang, Philip Williams and Michael Auli in the Edinburgh SMT group as well as Kevin Knight, David Chiang and Andrew Dai for inspiring discussions. We appreciate Pichuan Chang, Huihsin Tseng, Richard Zens, Matthew Snover and Nguyen Bach for helping us

understand their brilliant work. Many thanks to the anonymous reviewers for their insightful comments and suggestions. This work was supported in part by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme) and in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

References

Bach, N., Gao, Q., and Vogel, S. (2009a). Source-side dependency tree reordering models with subtree movements and constraints. In *Proceedings of the Twelfth Machine Translation Summit (MTSummit-XII)*, Ottawa, Canada. International Association for Machine Translation.

Bach, N., Vogel, S., and Cherry, C. (2009b). Cohesive constraints in a beam search phrase-based decoder. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 1–4, Boulder, Colorado.

Birch, A., Blunsom, P., and Osborne, M. (2009). A quantitative analysis of reordering phenomena. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205, Athens, Greece.

Birch, A. and Osborne, M. (2010). LRscore for evaluating lexical and reordering quality in MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 327–332, Uppsala, Sweden.

Birch, A. and Osborne, M. (2011). Reordering metrics for mt. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies*, pages 1027–1035, Portland, Oregon, USA.

Birch, A., Osborne, M., and Koehn, P. (2008). Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii.

Carpuat, M. and Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic.

Chang, P.-C., Tseng, H., Jurafsky, D., and Manning, C. D. (2009). Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009*, pages 51–59, Boulder, Colorado.

Cherry, C. (2008). Cohesive phrase-based decoding for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 72–80, Columbus, Ohio.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Stroudsburg, PA, USA.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

Chiang, D. (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.

Chiang, D., Knight, K., and Wang, W. (2009). 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 218–226, Boulder, Colorado.

Chiang, D., Marton, Y., and Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, Hawaii.

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*.

Ding, Y. and Palmer, M. (2005). Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 541–548, Ann Arbor, Michigan.

Fox, H. (2002). Phrasal cohesion and statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 304–3111.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. (2004). What's in a translation rule? In *HLT-NAACL 2004: Main Proceedings*, pages 273–280, Boston, Massachusetts, USA.

Hayashi, K., Tsukada, H., Sudoh, K., Duh, K., and Yamamoto, S. (2010). Hierarchical phrase-based machine translation with word-based reordering model. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 439–446, Beijing, China.

Hoang, H., Koehn, P., and Lopez, A. (2009). A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 152–159, Tokyo, Japan.

Huang, Z., Cmejrek, M., and Zhou, B. (2010). Soft syntactic constraints for hierarchical phrase-based translation using latent syntactic distributions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 138–147, Cambridge, MA.

Katz-Brown, J. and Collins, M. (2008). Syntactic reordering in preprocessing for japanese-to-english translation: Mit system description for ntcir-7 patent translation task. In *Proceedings of NTCIR-7 Workshop Meeting*, Tokyo, Japan.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 181–184.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P., Axelrod, A., Birch, A., Callison-burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of IWSLT2005*.

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

Marton, Y. and Resnik, P. (2008). Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of ACL-08: HLT*, pages 1003–1011, Columbus, Ohio.

Mi, H. and Liu, Q. (2010). Constituency to dependency translation with forests. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1433–1442, Uppsala, Sweden.

Och, F. J. (2003). Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279, Ann Arbor, Michigan.

Schwenk, H. and Koehn, P. (2008). Large and diverse language models for statistical machine translation. In *Proceedings of International Joint Conference on Natural Language Processing*.

Setiawan, H., Kan, M. Y., Li, H., and Resnik, P. (2009). Topological ordering of function words in hierarchical phrase-based translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 324–332, Suntec, Singapore.

Setiawan, H. and Resnik, P. (2010). Generalizing hierarchical phrase-based translation using rules with adjacent nonterminals. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 349–352, Los Angeles, California.

Shen, L., Xu, J., and Weischedel, R. (2008). A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT*, pages 577–585, Columbus, Ohio.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Tillman, C. (2004). A unigram orientation model for statistical machine translation. In *HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA.

Tromble, R. and Eisner, J. (2009). Learning linear ordering problems for better translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1007–1016, Singapore.

Wang, C., Collins, M., and Koehn, P. (2007). Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and*

Computational Natural Language Learning (EMNLP-CoNLL), pages 737–745, Prague, Czech Republic.

Wang, W., May, J., Knight, K., and Marcu, D. (2010). Re-structuring, re-labeling, and re-aligning for syntax-based machine translation. *Computational Linguistics*, 36(2).

Watanabe, T., Suzuki, J., Tsukada, H., and Isozaki, H. (2007). Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 764–773, Prague, Czech Republic.

Wu, D. (1996). A polynomial-time algorithm for statistical machine translation. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 152–158, Santa Cruz, California, USA.

Xiong, D., Liu, Q., and Lin, S. (2006). Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528, Sydney, Australia.

Xiong, D., Liu, Q., and Lin, S. (2007). A dependency treelet string correspondence model for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 40–47, Prague, Czech Republic.

Xu, P., Kang, J., Ringgaard, M., and Och, F. (2009). Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253, Boulder, Colorado.

Zens, R. and Ney, H. (2006). Discriminative reordering models for statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 55–63, New York City.

Zollmann, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City.