Proceedings of HLT-NAACL 2003
Short Papers , pp. 49-51
Edmonton, May-June 2003

Word Alignment with Cohesion Constraint

Dekang Lin and **Colin Cherry**
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada, T6G 2E8
{lindek,colinc}@cs.ualberta.ca

Abstract

We present a syntax-based constraint for word alignment, known as the cohesion constraint. It requires disjoint English phrases to be mapped to non-overlapping intervals in the French sentence. We evaluate the utility of this constraint in two different algorithms. The results show that it can provide a significant improvement in alignment quality.

1 Introduction

The IBM statistical machine translation (SMT) models have been extremely influential in computational linguistics in the past decade. The (arguably) most striking characteristic of the IBM-style SMT models is their total lack of linguistic knowledge. The IBM models demonstrated how much one can do with pure statistical techniques, which have inspired a whole new generation of NLP research and systems.

More recently, there have been many proposals to introduce syntactic knowledge into SMT models (Wu, 1997; Alshawi et al., 2000; Yamada and Knight, 2001; Lopez et al., 2002). A common theme among these approaches is the assumption that the syntactic structures of a pair of source-target sentences are isomorphic (or nearly isomorphic). This assumption seems too strong. Human translators often use non-literal translations, which result in differences in syntactic structures. According to a study in (Dorr et al., 2002), such translational divergences are quite common, involving 11-31% of the sentences.

We introduce a constraint that uses the dependency tree of the English sentence to maintain phrasal cohesion in the French sentence. In other words, if two phrases are disjoint in the English sentence, the alignment must not map them to overlapping intervals in the French sentence. For example, in Figure 1, the cohesion constraint will rule out the possibility of aligning *to* with *à*. The phrases *the reboot* and *the host to discover all the devices* are disjoint, but the partial alignment in Figure 1 maps them to overlapping intervals. This constraint is weaker than isomorphism. However, we will show that it can produce a significant increase in alignment quality.

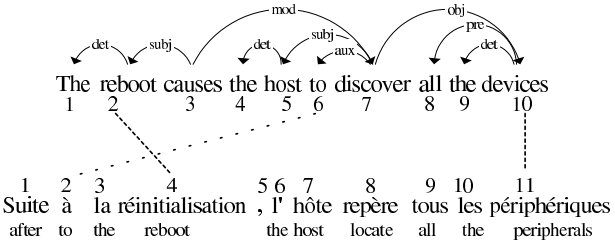

Figure 1: A cohesion constraint violation

2 Cohesion Constraint

Given an English sentence $E = e_1 e_2 \dots e_l$ and a French sentence $F = f_1 f_2 \dots f_m$, an **alignment** is a set of links between the words in E and F. An alignment can be represented as a binary relation A in $[1, l] \times [1, m]$. A pair (i, j) is in A if e_i and f_j are a translation (or part of a translation) of each other. We call such pairs **links**. In Figure 2, the links in the alignment are represented by dashed lines.

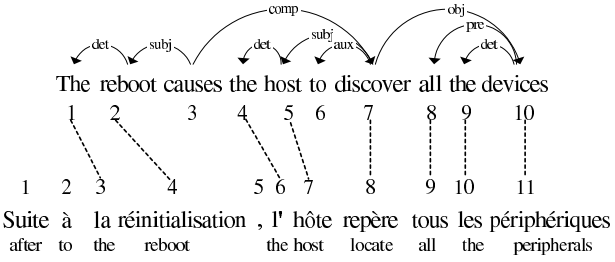

Figure 2: An example pair of aligned sentence

The **cohesion constraint** (Fox, 2002) uses the dependency tree T_E (Mel'čuk, 1987) of the English sentence

to restrict possible link combinations. Let $T_E(e_i)$ be the subtree of T_E rooted at e_i. The **phrase span** of e_i, $span_P(e_i, T_E, A)$, is the image of the English phrase headed by e_i in F given a (partial) alignment A. More precisely, $span_P(e_i, T_E, A) = [k_1, k_2]$, where

$$k_1 = \min\{j|(u,j) \in A, e_u \in T_E(e_i)\}$$
$$k_2 = \max\{j|(u,j) \in A, e_u \in T_E(e_i)\}$$

The **head span** is the image of e_i itself. We define $span_H(e_i, T_E, A) = [k_1, k_2]$, where

$$k_1 = \min\{j|(i,j) \in A\}$$
$$k_2 = \max\{j|(i,j) \in A\}$$

In Figure 2, the phrase span of the node *discover* is [6, 11] and the head span is [8, 8]; the phrase span of the node *reboot* is [3, 4] and the head span is [4, 4]. The word *cause* has a phrase span of [3,11] and its head span is the empty set $\emptyset$.

With these definitions of phrase and head spans, we define two notions of overlap, originally introduced in (Fox, 2002) as **crossings**. Given a head node e_h and its modifier e_m, a **head-modifier overlap** occurs when:

$$span_H(e_h, T_E, A) \cap span_P(e_m, T_E, A) \neq \emptyset$$

Given two nodes e_{m_1} and e_{m_2} which both modify the same head node, a **modifier-modifier overlap** occurs when:

$$span_P(e_{m_1}, T_E, A) \cap span_P(e_{m_2}, T_E, A) \neq \emptyset$$

Following (Fox, 2002), we say an alignment is cohesive with respect to T_E if it does not introduce any head-modifier or modifier-modifier overlaps. For example, the alignment A in Figure 1 is not cohesive because there is an overlap between $span_P(reboot, T_E, A)$=$[4, 4]$ and $span_P(discover, T_E, A)$=$[2, 11]$.

If an alignment A' violates the cohesion constraint, any alignment A that is a superset of A' will also violate the cohesion constraint. This is because any pair of nodes that have overlapping spans in A' will still have overlapping spans in A.

Cohesion Checking Algorithm:
We now present an algorithm that checks whether an individual link (e_i, f_j) causes a cohesion constraint violation when it is added to a partial alignment. Let $e_{p_0}, e_{p_1}, e_{p_2}, \ldots$ be a sequence of nodes in T_E such that e_{p_0}=e_i and e_{p_k}=*parentOf*$(e_{p_{k-1}})$ $(k = 1, 2, \ldots)$

1. For all $k \geq 0$, update the $span_P$ and the $span_H$ of e_{p_k} to include j.
2. For each e_{p_k} $(k > 0)$, check for a modifier-modifier overlap between the updated the phrase span of $e_{p_{k-1}}$ and the the phrase span of each of the other children of e_{p_k}.
3. For each e_{p_k} $(k > 0)$, check for a head-modifier overlap between the updated phrase span of $e_{p_{k-1}}$ and the head span of e_{p_k}.
4. If an overlap is found, return true (the constraint is violated). Otherwise, return false.

3 Evaluation

To determine the utility of the cohesion constraint, we incorporated it into two alignment algorithms. The algorithms take as input an English-French sentence pair and the dependency tree of the English sentence. Both algorithms build an alignment by adding one link at a time. We implement two versions of each algorithm: one with the cohesion constraint and one without. We will describe the versions without cohesion constraint below. For the versions with cohesion constraint, it is understood that each new link must also pass the test described in Section 2.

The first algorithm is similar to Competitive Linking (Melamed, 1997). We use a sentence-aligned corpus to compute the ϕ^2 correlation metric (Gale and Church, 1991) between all English-French word pairs. For a given sentence pair, we begin with an empty alignment. We then add links in the order of their ϕ^2 scores so that each word participates in at most one link. We will refer to this as the ϕ^2 method.

The second algorithm uses a best-first search (with fixed beam width and agenda size) to find an alignment that maximizes $P(A|E, F)$. A state in this search space is a partial alignment. A transition is defined as the addition of a single link to the current state. The algorithm computes $P(A|E, F)$ based on statistics obtained from a word-aligned corpus. We construct the initial corpus with a system that is similar to the ϕ^2 method. The algorithm then re-aligns the corpus and trains again for three iterations. We will refer to this as the $P(A|E, F)$ method. The details of this algorithm are described in (Cherry and Lin, 2003).

We trained our alignment programs with the same 50K pairs of sentences as (Och and Ney, 2000) and tested it on the same 500 manually aligned sentences. Both the training and testing sentences are from the Hansard corpus. We parsed the training and testing corpora with Minipar.[1] We adopted the evaluation methodology in (Och and Ney, 2000), which defines three metrics: precision, recall and alignment error rate (AER).

Table 1 shows the results of our experiments. The first four rows correspond to the methods described above. As a reference point, we also provide the results reported in (Och and Ney, 2000). They implemented IBM Model 4 by bootstrapping from an HMM model. The rows *F*→*E*

[1] available at http://www.cs.ualberta.ca/~lindek/minipar.htm

Table 1: Evaluation Results

Method		Prec	Rec	AER
ϕ^2	w/o cohesion	82.7	84.6	16.5
	w/ cohesion	89.2	82.7	13.8
$P(A\|E,F)$	w/o cohesion	87.3	85.3	13.6
	w/ cohesion	95.7	86.4	8.7
Och&Ney	F→E	80.5	91.2	15.6
	E→F	80.0	90.8	16.0
	Refined	85.9	92.3	11.7

and $E \rightarrow F$ are the results obtained by this model when treating French as the source and English as the target or vice versa. The row *Refined* shows results obtained by taking the intersection of $E \rightarrow F$ and $F \rightarrow E$ and then refining this intersection to increase recall.

From Table 1, we can see that the addition of the cohesion constraint leads to significant improvements in performance with both algorithms. The relative reduction in error rate is 16% with the ϕ^2 method and 36% with the $P(A|E,F)$ method. The improvement comes primarily from increased precision. With the $P(A|E,F)$ method, this increase in precision does not come at the expense of recall.

4 Related Work

There has been a growing trend in the SMT community to attempt to leverage syntactic data in word alignment. Methods such as (Wu, 1997), (Alshawi et al., 2000) and (Lopez et al., 2002) employ a synchronous parsing procedure to constrain a statistical alignment. The work done in (Yamada and Knight, 2001) measures statistics on operations that transform a parse tree from one language into another.

The syntactic knowledge that is leveraged in these methods is tightly coupled with the alignment method itself. We have presented a modular constraint that can be plugged into different alignment algorithms. This has allowed us to test the contribution of the constraint directly.

(Fox, 2002) studied the extent to which the cohesion constraint holds in a parallel corpus and the reasons for the violations, but did not apply the constraint to an alignment algorithm.

5 Conclusion

We have presented a syntax-based constraint for word alignment, known as the cohesion constraint. It requires disjoint English phrases to be mapped to non-overlapping intervals in the French sentence. Our experiments have shown that the use of this constraint can provide a relative reduction in alignment error rate of 36%.

Acknowledgments

We wish to thank Franz Och for providing us with manually aligned evaluation data. This project is funded by and jointly undertaken with Sun Microsystems, Inc. We wish to thank Finola Brady, Bob Kuhns and Michael McHugh for their help.

References

Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 2000. Learning dependency translation models as collections of finite state head transducers. *Computational Linguistics*, 26(1):45–60.

Colin Cherry and Dekang Lin. 2003. A probability model to improve word alignment. Submitted.

Bonnie J. Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. Duster: A method for unraveling cross-language divergences for statistical word-level alignment. In Stephen D. Richardson, editor, *Proceedings of AMTA-02*, pages 31–43, Tiburon, CA, October. Springer.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP-02*, pages 304–311.

W.A. Gale and K.W. Church. 1991. Identifying word correspondences in parallel texts. In *Proceedings of the 4th Speech and Natural Language Workshop*, pages 152–157. DARPA, Morgan Kaufmann.

Adam Lopez, Michael Nossal, Rebecca Hwa, and Philip Resnik. 2002. Word-level alignment for multilingual resource acquisition. In *Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data.*

I. Dan Melamed. 1997. A word-to-word model of translational equivalence. In *Proceedings of the ACL-97*, pages 490–497. Association for Computational Linguistics.

Igor A. Mel'čuk. 1987. *Dependency syntax: theory and practice*. State University of New York Press, Albany.

Franz J. Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong, China, October.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):374–403.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Meeting of the Association for Computational Linguistics*, pages 523–530.