Improving Bitext Word Alignments via Syntax-based Reordering of English

Elliott Franco Drábek and David Yarowsky
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218, USA
{edrabek,yarowsky}@cs.jhu.edu

Abstract

We present an improved method for automated word alignment of parallel texts which takes advantage of knowledge of syntactic divergences, while avoiding the need for syntactic analysis of the less resource rich language, and retaining the robustness of syntactically agnostic approaches such as the IBM word alignment models. We achieve this by using simple, easily-elicited knowledge to produce syntax-based heuristics which transform the target language (e.g. English) into a form more closely resembling the source language, and then by using standard alignment methods to align the transformed bitext. We present experimental results under variable resource conditions. The method improves word alignment performance for language pairs such as English-Korean and English-Hindi, which exhibit longer-distance syntactic divergences.

1 Introduction

Word-level alignment is a key infrastructural technology for multilingual processing. It is crucial for the development of translation models and translation lexica (Tufiş, 2002; Melamed, 1998), as well as for translingual projection (Yarowsky et al., 2001; Lopez et al., 2002). It has increasingly attracted attention as a task worthy of study in its own right (Mihalcea and Pedersen, 2003; Och and Ney, 2000).

Syntax-light alignment models such as the five IBM models (Brown et al., 1993) and their relatives have proved to be very successful and robust at producing word-level alignments, especially for closely related languages with similar word order and mostly local reorderings, which can be captured via simple models of relative word distortion. However, these models have been less successful at modeling syntactic distortions with longer distance movement. In contrast, more syntactically informed approaches have been constrained by the often weak syntactic correspondences typical of real-world parallel texts, and by the difficulty of finding or inducing syntactic parsers for any but a few of the world's most studied languages.

Our approach uses simple, easily-elicited knowledge of divergences to produce heuristic syntax-based transformations from English to a form (English′) more closely resembling the source language, and then using standard alignment methods to align the transformed version to the target language. This approach retains the robustness of syntactically agnostic models, while taking advantage of syntactic knowledge. Because the approach relies only on syntactic analysis of English, it can avoid the difficulty of developing a full parser for a new low-resource language.

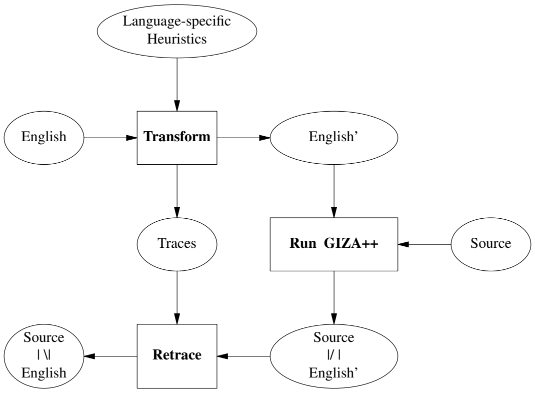

Figure 1: System Architecture

Our method is rapid and low cost. It requires only coarse-grained knowledge of basic word order, knowledge which can be rapidly found in even the briefest grammatical sketches. Because basic word order changes very slowly with time, word order of related languages tends to be very similar. For example, even if we only know that a language is of the Northern-Indian/Sanskrit family, we can easily guess with high confidence that it is systematically head-final. Because our method can be restricted to only bi-text pre-processing and post-processing, it can be used as a wrapper around any existing word-alignment tool, without modification, to provide improved performance by minimizing alignment distortion.

2 Prior Work

The 2003 HLT-NAACL Workshop on Building and Using Parallel Texts (Mihalcea and Pedersen, 2003) reflected the increasing importance of the word-alignment task, and established standard performance measures and some benchmark tasks.

There is prior work studying systematic cross-

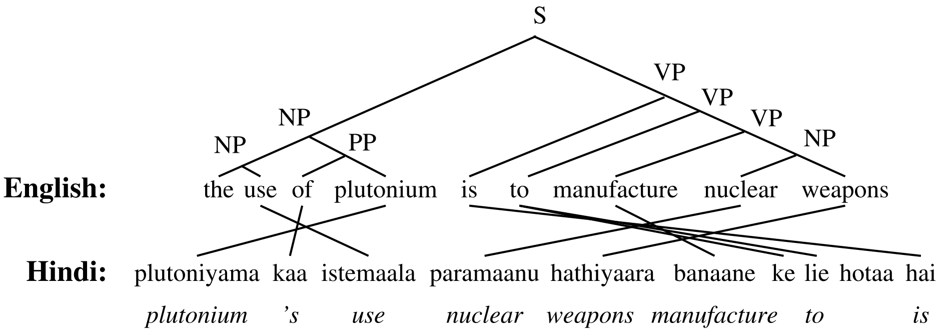
S
VP
VP
VP
NP
NP
NP
PP
English:
the use of plutonium is to manufacture nuclear weapons
Hindi:
plutoniyama kaa istemaala paramaanu hathiyaara banaane ke lie hotaa hai
plutonium 's use nuclear weapons manufacture to is

Figure 2: Original Hindi-English sentence pair with gold-standard word-alignments.

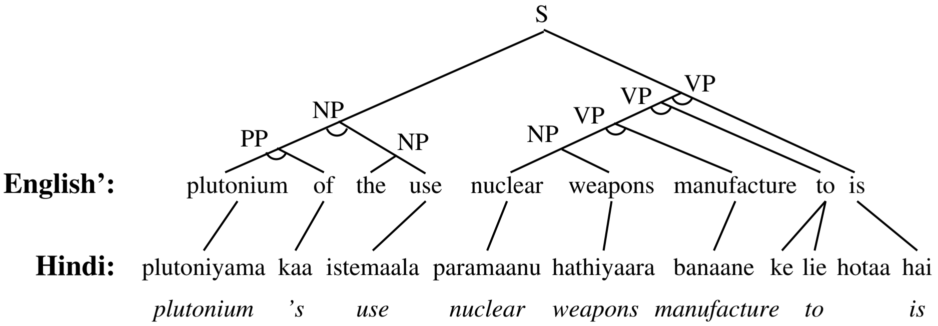
S
NP
PP
NP
VP
VP
VP
NP
English':
plutonium of the use nuclear weapons manufacture to is
Hindi:
plutoniyama kaa istemaala paramaanu hathiyaara banaane ke lie hotaa hai
plutonium 's use nuclear weapons manufacture to is

Figure 3: Transformed Hindi-English′ sentence pair with gold-standard word-alignments. Rotated nodes are marked with an arc.

linguistic structural divergences, such as the DUSTer system (Dorr et al., 2002). While the focus on major classes of structural variation such as manner-of-motion verb-phrase transformations have facilitated both transfer and generation in machine translation, these divergences have not been integrated into a system that produces automatic word alignments and have tended to focus on more local phrasal variation rather than more comprehensive sentential syntactic reordering.

Complementary prior work (e.g. Wu, 1995) has also addressed syntactic transduction for bilingual parsing, translation, and word-alignment. Much of this work depends on high-quality parsing of both target and source sentences, which may be unavailable for many "lower density" languages of interest. Tree-to-string models, such as (Yamada and Knight, 2001) remove this dependency, and such models are well suited for situations with large, cleanly translated training corpora. By contrast, our method retains the robustness of the underlying aligner towards loose translations, and can if necessary use knowledge of syntactic divergences even in the absence of any training corpora whatsoever, using only a translation lexicon.

3 System

Figure 1 shows the system architecture. We start by running the Collins parser (Collins, 1999) on the English side of both training and testing data, and apply our source-language-specific heuristics to the resulting trees. This yields English′ text, along with traces recording correspondences between English′ words and the English originals. We use GIZA++ (Och and Ney, 2000) to align the English′ with the source language text, yielding alignments in terms of the English′. Finally, we use the traces to map these alignments to the original English words.

Language	**VP**	**AP**	**NP**
English	VO	AO	AN, NR
Hindi	OV	OA	AN, RN
Korean	OV	OA	AN, RN
Chinese	VO	AOA	AN, RN
Romanian	VO	AO	NA, NR

Table 1: Basic word order for three major phrase types – VP: verb phrases with **V**erb and **O**bject, AP: appositional (prepositional or postpositional) phrases with **A**pposition and **O**bject, and NP: noun phrases with **N**oun and **A**djective or **R**elative clause. Chinese has both prepositions and postpositions.

Figure 2 shows an illustrative Hindi-English sentence pair, with true word alignments, and parse-tree over the English sentence. Although it is only a short sentence, the large number of crossing alignments clearly show the high-degree of reordering, and especially long-distance motion, caused by the syntactic divergences between Hindi and English.

Figure 3 shows the same sentence pair after English has been transformed into English′ by our system. Tree nodes whose children have been reordered

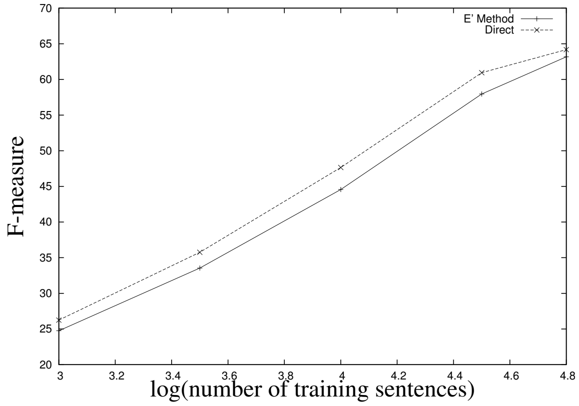

Figure 4: Hindi alignment performance

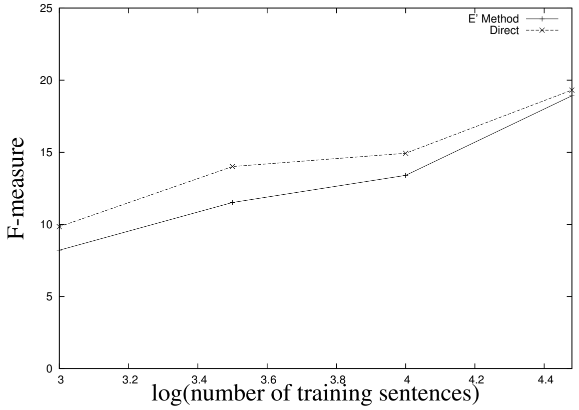

Figure 5: Korean alignment performance

are marked by a subtended arc. Crossings have been eliminated, and the alignment is now monotonic.

Table 1 shows the basic word order of three major phrase types for each of the languages we treated. In each case, our heuristics transform the English trees to achieve these same word orders. For the Chinese case, we apply several more language-specific transformations. Because Chinese has both prepositions and postpositions, we retain the original preposition and add an additional bracketing postposition. We also move verb modifiers other than noun phrases to the left of the head verb.

4 Experiments

For each language we treated, we assembled sentence-aligned, tokenized training and test corpora, with hand-annotated gold-standard word alignments for the latter[1]. We did not apply any sort of morphological analysis beyond basic word tokenization. We measured system performance with `wa_eval_align.pl`, provided by Rada Mihalcea and Ted Pedersen.

Each training set provides the aligner with information about lexical affinities and reordering patterns. For Hindi, Korean and Chinese, we also tested our system under the more difficult situation of having only a bilingual word list but no bitext available. This is a plausible low-resource language scenario

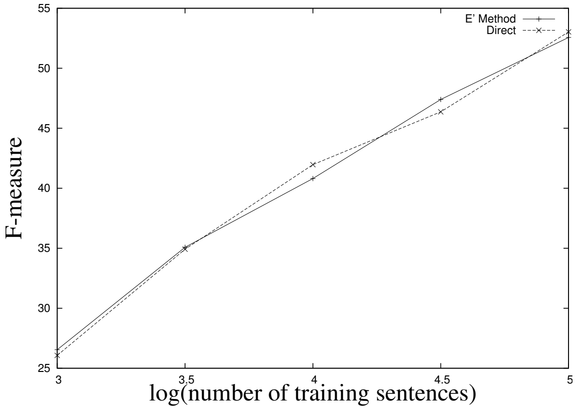

Figure 6: Chinese alignment performance

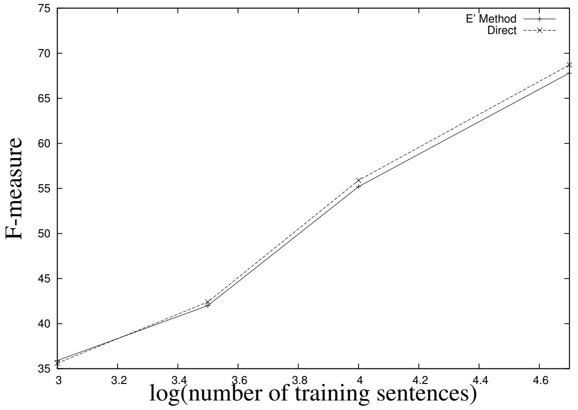

Figure 7: Romanian alignment performance

# Train	**Direct**			**English′**		
Sents	**P**	**R**	**F**	**P**	**R**	**F**
	Hindi					
Dict only	16.4	13.8	15.0	18.5	15.6	**17.0**
1000	26.8	23.0	24.8	28.4	24.4	**26.2**
3162	35.7	31.6	33.5	38.4	33.5	**35.8**
10000	46.6	42.7	44.6	50.4	45.2	**47.6**
31622	60.1	56.0	58.0	63.6	58.5	**61.0**
63095	64.7	61.7	63.2	66.3	62.2	**64.2**
	Korean					
Dict only	26.6	12.3	16.9	27.5	12.9	**17.6**
1000	9.4	7.3	8.2	11.3	8.7	**9.8**
3162	13.2	10.2	11.5	16.0	12.4	**14.0**
10000	15.2	12.0	13.4	17.0	13.3	**14.9**
30199	21.5	16.9	18.9	21.9	17.2	**19.3**
	Chinese					
Dict only	44.4	30.4	36.1	44.5	30.5	**36.2**
1000	33.0	22.2	**26.5**	30.8	22.6	26.1
3162	44.6	28.9	**35.1**	41.7	30.0	34.9
10000	51.1	34.0	40.8	50.7	35.8	**42.0**
31622	60.4	39.0	**47.4**	55.7	39.7	46.4
100000	66.0	43.7	52.6	63.7	45.4	**53.0**
	Romanian					
1000	49.6	27.7	35.6	50.1	28.0	**35.9**
3162	57.9	33.4	**42.4**	57.6	33.0	42.0
10000	72.6	45.5	**55.9**	71.3	45.0	55.2
48441	84.7	57.8	**68.7**	83.5	57.1	67.8

Table 2: Performance in Precision, Recall, and F-measure (per cent) of all systems.

Source Language	**# Test Sents**	**Mean Length**	**Correlation**	
			Direct	**E′**
Hindi	46	16.3	54.1	60.1
Korean	100	20.2	10.2	31.6
Chinese	88	26.5	60.2	63.7
Romanian	248	22.7	81.1	80.6

Table 3: Test set characteristics, including number of sentence pairs, mean length of English sentences, and correlation r^2 between English and source-language normalized word positions in gold-standard data, for direct and English′ situations.

and a test of the ability of the system to take sole responsibility for knowledge of reordering.

Table 3 describes the test sets and shows the correlation in gold standard aligned word pairs between the position of the English word in the English sentence and the position of the source-language word in the source-language sentence (normalizing the positions to fall between 0 and 1). The baseline (direct) correlations give quantitative evidence of differing degrees of syntactic divergence with English, and the English′ correlations demonstrate that our heuristics do have the effect of better fitting source language word order.

5 Results

Figures 4, 5, 6 and 7 show learning curves for systems trained on parallel sentences with and without the English′ transforms. Table 2 provides further detail, and also shows the performance of systems trained without any bitext, but only with access to a bilingual translation lexicon. Our system achieves consistent, substantial performance improvement under all situations for English-Hindi and English-Korean language pairs, which exhibit longer distance SOV→SVO syntactic divergence. For English-Romanian and English-Chinese, neither significant improvement nor degradation is seen, but these are language pairs with quite similar sentential word order to English, and hence have less opportunity to benefit from our syntactic transformations.

6 Conclusions

We have developed a system to improve the performance of bitext word alignment between English and a source language by first reordering parsed English into an order more closely resembling that of the source language, based only on knowledge of the coarse basic word order of the source language, such as can be obtained from any cross-linguistic survey of languages, and requiring *no* parsing of the source language. We applied the system to the task of aligning English with Hindi, Korean, Chinese and Romanian. Performance improvement is greatest for Hindi and Korean, which exhibit longer-distance constituent reordering with respect to English. These properties suggest the proposed English′ word alignment method can be an effective approach for word alignment to languages with both greater cross-linguistic word-order divergence and an absence of available parsers.

[1] Hindi training: news text from the LDC for the 2003 DARPA TIDES Surprise Language exercise; Hindi testing: news text from Rebecca Hwa, then at the University of Maryland; Hindi dictionary: The Hindi-English Dictionary, v. 2.0 from IIIT (Hyderabad) LTRC; Korean training: Unbound Bible; Korean testing: half from Penn Korean Treebank and half from Universal declaration of Human Rights, aligned by Woosung Kim at the Johns Hopkins University; Korean dictionary: EngDic v. 4; Chinese training: news text from FBIS; Chinese testing: Penn Chinese Treebank news text aligned by Rebecca Hwa, then at the University of Maryland; Chinese dictionary: from the LDC; Romanian training and testing: (Mihalcea and Pedersen, 2003).

References

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

B. J. Dorr, L. Pearl, R. Hwa, and N. Habash. 2002. DUSTer: A method for unraveling cross-language divergences for statistical word-level alignment. In *Proceedings of AMTA-02*, pages 31–43.

A. Lopez, M. Nosal, R. Hwa, and P. Resnik. 2002. Word-level alignment for multilingual resource acquisition. In *Proceedings of the LREC-02 Workshop on Linguistic Knowledge Acquisition and Representation*.

I. D. Melamed. 1998. Empirical methods for MT lexicon development. *Lecture Notes in Computer Science*, 1529:18–9999.

R. Mihalcea and T. Pedersen. 2003. An evaluation exercise for word alignment. In Rada Mihalcea and Ted Pedersen, editors, *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts*, pages 1–10.

F. J. Och and H. Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of COLING-00*, pages 1086–1090.

D. I. Tufiş. 2002. A cheap and fast way to build useful translation lexicons. In *Proceedings of COLING-02*, pages 1030–1036.

D. Wu. 1995. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proceedings of IJCAI-95*, pages 1328–1335.

K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL-01*, pages 523–530.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT-01*, pages 161–168.