Competitive Grouping in Integrated Phrase Segmentation and Alignment Model

Ying Zhang **Stephan Vogel**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{joy+,vogel+}@cs.cmu.edu

Abstract

This article describes the competitive grouping algorithm at the core of our Integrated Segmentation and Alignment (ISA) model. ISA extracts phrase pairs from a bilingual corpus without requiring the pre-calculated word alignment as many other phrase alignment models do. Experiments conducted within the WPT-05 shared task on statistical machine translation demonstrate the simplicity and effectiveness of this approach.

1 Introduction

In recent years, various phrase translation approaches (Marcu and Wong, 2002; Och et al., 1999; Koehn et al., 2003) have been shown to outperform word-to-word translation models (Brown et al., 1993). Many of these phrase alignment strategies rely on the pre-calculated word alignment and use different heuristics to extract the phrase pairs from the Viterbi word alignment path. The Integrated Segmentation and Alignment (ISA) model (Zhang et al., 2003) does not require such word alignment. ISA segments the sentence into phrases and finds their alignment simultaneously. ISA is simple and fast. Translation experiments have shown comparable performance to other phrase alignment strategies which require complicated statistical model training. In this paper, we describe the key idea behind this model and connect it with the competitive linking algorithm (Melamed, 1997) which was developed for word-to-word alignment.

2 Translation Likelihood as a Statistical Test

Given a bilingual corpus of language pair *F* (Foreign, source language) and *E* (English, target language), if we know the word alignment for each sentence pair we can calculate the co-occurrence frequency for each source/target word pair type $C(f, e)$ and the marginal frequency $C(f) = \sum_e C(f, e)$ and $C(e) = \sum_f C(f, e)$. We can apply various statistical tests (Manning and Schütze, 1999) to measure how likely is the association between f and e, in other words how likely they are mutual translations. In the following sections, we will use χ^2 statistics to measure the the mutual translation likelihood (Church and Hanks, 1990).

3 The Core of the Integrated Phrase Segmentation and Alignment

The competitive linking algorithm (CLA) (Melamed, 1997) is a greedy word alignment algorithm. It was designed to overcome the problem of indirect associations using a simple heuristic: whenever several word tokens f_i in one half of the bilingual corpus co-occur with a particular word token e in the other half of the corpus, the word that is most likely to be e's translation is the one for which the likelihood $L(f, e)$ of translational equivalence is highest. The simplicity of this algorithm depends on a one-to-one alignment assumption. Each word translates to at most one other word. Thus when one pair $\{f, e\}$ is "linked", neither f nor e can be aligned with any other words. This assumption renders CLA unusable in phrase level alignment.

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 159–162,
Ann Arbor, June 2005.

We propose an extension, the competitive grouping, as the core component in the ISA model.

3.1 Competitive Grouping Algorithm (CGA)

The key modification to the competitive linking algorithm is to make it less greedy. When a word pair is found to be the winner of the competition, we allow it to invite its neighbors to join the "winner's club" and group them together as an aligned phrase pair. The one-to-one assumption is thus discarded in CGA. In addition, we introduce the *locality* assumption for phrase alignment. *Locality* states that a source phrase of adjacent words can only be aligned to a target phrase composed of adjacent words. This is not true of most language pairs in cases such as the relative clause, passive tense, and prepositional clause, etc.; however this assumption renders the problem tractable. Here is a description of CGA:

For a sentence pair $\{\mathbf{f}, \mathbf{e}\}$, represent the word pair statistics for each word pair $\{f, e\}$ in a two dimensional matrix $L_{I\times J}$, where $L(i,j) = \chi^2(f_i, e_j)$ in our implementation. [1]

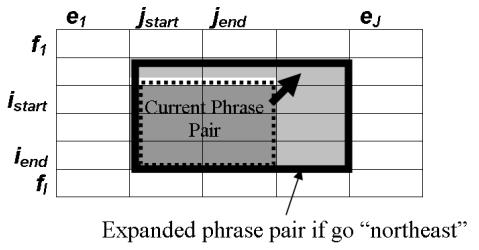

Figure 1: Expanding the current phrase pair

Denote an aligned phrase pair $\{\tilde{f}, \tilde{e}\}$ as a tuple $[i_{start}, i_{end}, j_{start}, j_{end}]$ where $\tilde{f}$ is $f_{i_{start}}, f_{i_{start+1}}, \ldots, f_{i_{end}}$ and similarly for $\tilde{e}$.

1. Find i^* and j^* such that $L(i^*, j^*)$ is the highest. Create a *seed* phrase pair $[i^*, i^*, j^*, j^*]$ which is simply the word pair $\{f_{i^*}, e_{j^*}\}$ itself.

2. Expand the current phrase pair $[i_{start}, i_{end}, j_{start}, j_{end}]$ to the neighboring territory to include adjacent source and target words in the phrase alignment group. There are 8 ways to group new words into the phrase pair. For example, one can expand to the north by including an additional source word $f_{i_{start}-1}$ to be aligned with all the target words in the current group; or one can expand to the northeast by including $f_{i_{start}-1}$ and $e_{j_{end}+1}$ (Figure 1).

   Two criteria have to be satisfied for each expansion:

   (a) If a new source word $f_{i'}$ is to be grouped, $\max_{j_{start}\le j\le j_{end}} L(i', j)$ should be no smaller than $\max_{1\le j\le J} L(i', j)$. Since CGA is a greedy algorithm as described below, this is to guarantee that $f_{i'}$ will not "regret" the decision of joining the phrase pair because it does not have other "better" target words to be aligned with. Similar constraint is applied if a new target word $e_{j'}$ is to be grouped.

   (b) The highest value in the newly-expanded area needs to be "similar" to the seed value $L(i^*, j^*)$.

   Expand the current phrase pair to the largest extend possible as long as both criteria are satisfied.

3. The locality assumption means that the aligned phrase cannot be aligned again. Therefore, all the source and target words in the phrase pair are marked as "invalid" and will be skipped in the following steps.

4. If there is another valid pair $\{f_i, e_j\}$, then repeat from Step 1.

Figure 2 and Figure 3 show a simple example of applying CGA on the sentence pair {*je déclare reprise la session/i declare resumed the session*}.

	i	declare	resumed	the	session
je	40316.90	0.79	0.01	19.39	0.04
déclare	0.40	760.79	40.85	0.33	86.78
reprise	0.01	24.66	312.73	0.31	402.86
la	10.50	0.01	0.17	667.49	1.60
session	0.00	40.42	5.13	0.80	3795.00

Figure 2: Seed pair {*je / i*}, no expansion allowed

[1] χ^2 statistics were found to be more discriminative in our experiments than other symmetric word association measures, such as the averaged mutual information, ϕ^2 statistics and Dice-coefficient.

	i	declare	resumed	the	session
je					
déclare		760.79	40.85	0.33	86.78
reprise		24.66	312.73	0.31	402.86
la		0.01	0.17	667.49	1.60
session		40.42	5.13	0.80	3795.00

Figure 3: Seed pair {*session/session*}, expanded to {*la session/the session*}

3.2 Exploring all possible groupings

The similarity criterion 2-(b) described previously is used to control the granularity of phrase pairs. In cases where the pairs $\{f_1 f_2, e_1 e_2\}$, $\{f_1, e_1\}$ and $\{f_2, e_2\}$ are all valid translations pairs, similarity is used to control whether we want to align $\{f_1 f_2, e_1 e_2\}$ as one phrase pair or two shorter ones.

The granularity of the phrase pairs is hard to optimize especially when the test data is unknown. On the one hand, we prefer long phrases since interaction among the words in the phrase, for example word sense, morphology and local reordering could be encapsulated. On the other hand, long phrase pairs are less likely to occur in the test data than the shorter ones and may lead to low coverage. To have both long and short phrases in the alignment, we apply a range of similarity thresholds for each of the expansion operations. By applying a low similarity threshold, the expanded phrase pairs tend to be large, while a higher similarity threshold results in shorter phrase pairs. As described above, CGA is a greedy algorithm and the expansion of the seed pair restricts the possible alignments for the rest of the sentence. Figure 4 shows an example as we explore all the possible grouping choices in a depth-first search. In the end, all unique phrase pairs along the path traveled are output as phrase translation candidates for the current sentence pair.

3.3 Phrase translation probabilities

Each aligned phrase pair $\{\tilde{f}, \tilde{e}\}$ is assigned a likelihood score $L(\tilde{f}, \tilde{e})$, defined as:

$$\frac{\sum_i \max_j \log L(f_i, e_j) + \sum_j \max_i \log L(f_i, e_j)}{|\tilde{f}| + |\tilde{e}|}$$

where i ranges over all words in $\tilde{f}$ and similarly j in $\tilde{e}$.

Given the collected phrase pairs and their likelihood, we estimate the phrase translation probability

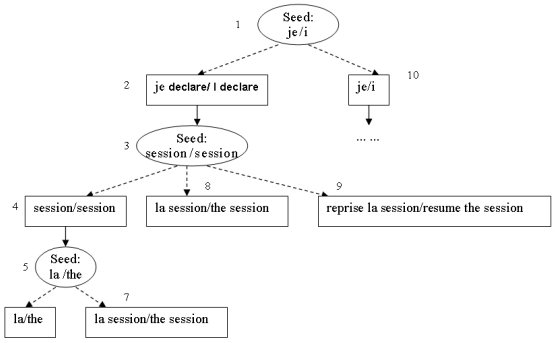

Figure 4: Depth-first itinerary of all possible grouping choices.

by their weighted frequency:

$$P(\tilde{f}|\tilde{e}) = \frac{count(\tilde{f}, \tilde{e}) \cdot L(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} count(\tilde{f}, \tilde{e}) \cdot L(\tilde{f}, \tilde{e})}$$

No smoothing is applied to the probabilities.

4 Learning co-occurrence information

In most cases, word alignment information is not given and is treated as a hidden parameter in the training process. We initialize a word pair co-occurrence frequency by assuming uniform alignment for each sentence pair, i.e. for sentence pair $(\mathbf{f}, \mathbf{e})$ where $\mathbf{f}$ has I words and $\mathbf{e}$ has J words, each word pair $\{f, e\}$ is considered to be aligned with frequency $\frac{1}{I \times J}$. These co-occurrence frequencies will be accumulated over the whole corpus to calculate the initial $L(f, e)$. Then we iterate the ISA model:

1. Apply the competitive grouping algorithm to each sentence pair to find all possible phrase pairs.
2. For each identified phrase pair $\{\tilde{f}, \tilde{e}\}$, increase the co-occurrence counts for all word pairs inside $\{\tilde{f}, \tilde{e}\}$ with weight $\frac{1}{|\tilde{f}| \cdot |\tilde{e}|}$.
3. Calculate $L(f, e)$ again and goto Step 1 for several iterations.

5 Experiments

We participated the shared task in the WPT05 workshop[2] and applied ISA to all four language pairs

[2] http://www.statmt.org/wpt05/mt-shared-task/

(French-English, Finnish-English, German-English and Spanish-English). Table 1 shows the n-gram coverage of the dev-test set. French and Spanish data are better covered by the training data compared to the German and Finnish sets. Since our phrase alignment is constrained by the locality assumption and we can only extract phrase pairs of adjacent words, lower n-gram coverage will result in lower translation scores. We used the training data

Dev-test	DE	ES	FI	FR
N=1	99.2	99.6	98.2	99.8
N=2	88.2	93.3	73.0	94.7
N=3	59.4	71.7	38.2	76.0
N=4	30.0	42.9	17.0	50.6
N=5	13.0	21.7	6.8	29.8
N=16	(8)	(65)	(1)	(101)
N=19	(1)	(23)		(34)
N=23		(1)		(1)

Table 1: Percentage of dev-test n-grams covered by the training data. Numbers in parenthesis are the actual counts of n-gram tokens in the dev-test data.

and the language model as provided and manually tuned the parameters of the Pharaoh decoder[3] to optimize BLEU scores. Table 2 shows the translation results on the dev-test and the test set of WPT05. The BLEU scores appear comparable to those of other state-of-the-art phrase alignment systems, in spite of the simplicity of the ISA model and ease of training.

	DE	ES	FI	FR
Dev-test	18.63	26.20	12.88	26.20
Test	18.93	26.14	12.66	26.71

Table 2: BLEU scores of ISA in WPT05

6 Conclusion

In this paper, we introduced the competitive grouping algorithm which is at the core of the ISA phrase alignment model. As an extension to the competitive linking algorithm which is used for word-to-word alignment, CGA overcomes the assumption of one-to-one mapping and makes it possible to align phrase pairs. Despite its simplicity, the ISA model has achieved competitive translation results. We plan to release ISA toolkit[4] to the community in the near future.

[3] http://www.isi.edu/licensed-sw/pharaoh/

[4] http://projectile.is.cs.cmu.edu/research/public/isa/index.htm

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edomonton, Canada, May 27-June 1.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Philadephia, PA, July 6-7.

I. Dan Melamed. 1997. A word-to-word model of translational equivalence. In *Proceedings of the 8-th conference on EACL*, pages 490–497, Morristown, NJ, USA. Association for Computational Linguistics.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2003. Integrated phrase segmentation and alignment algorithm for statistical machine translation. In *Proceedings of NLP-KE'03*, Beijing, China, October.