Symmetric Probabilistic Alignment

Ralf D. Brown **Jae Dong Kim** **Peter J. Jansen** **Jaime G. Carbonell**

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{ralf,jdkim,pjj,jgc}@cs.cmu.edu

Abstract

We recently decided to develop a new alignment algorithm for the purpose of improving our Example-Based Machine Translation (EBMT) system's performance, since subsentential alignment is critical in locating the correct translation for a matched fragment of the input. Unlike most algorithms in the literature, this new Symmetric Probabilistic Alignment (SPA) algorithm treats the source and target languages in a symmetric fashion.

In this short paper, we outline our basic algorithm and some extensions for using context and positional information, and compare its alignment accuracy on the Romanian-English data for the shared task with IBM Model 4 and the reported results from the prior workshop.

1 Symmetric Probabilistic Alignment (SPA)

In subsentential alignment, mappings are produced from words or phrases in the source language sentence and those words or phrases in the target language sentence that best express their meaning.

An alignment algorithm takes as input a bilingual corpus consisting of corresponding sentence pairs and strives to find the best possible alignment in the second for selected n-grams (sequences of n words) in the first language. The alignments are based on a number of factors, including a bilingual dictionary (preferably a probabilistic one), the position of the words, invariants such as numbers and punctuation, and so forth.

For our baseline algorithm, we make the following simplifying assumptions, each of which we intend to relax in future work, and the last of which has already been partially relaxed:

1. A fixed bilingual probabilistic dictionary is available.
2. Fragments (word sequences) are translated independently of surrounding context.
3. Contiguous fragments of source language text are translated into contiguous fragments in the target language text.

Unlike the work of (Marcu and Wong, 2002), our alignment algorithm is not generative and does not use the idea of a bag of concepts from which the phrases in the sentence pair arise. It is, rather, intended to find the corresponding target-language phrase given a specific source-language phrase of interest, as required by our EBMT system after finding a match between the input and the training data (Brown, 2004).

1.1 Baseline Algorithm

Our baseline algorithm is based on maximizing the probability of bi-directional translations of individual words between a selected n-gram in the source language and every possible n-gram in the corresponding paired target language sentence. No positional preference assumptions are made, nor are any length preservation assumptions made. That is, an n-gram may translate to an m-gram, for any values of n or m bounded by the source and target sentence lengths, respectively. Finally a smoothing factor is used to avoid singularities (i.e. avoiding zero-probabilities for unknown words, or words never translated before in a way consistent with the dictionary).

Given a source-language sentence

$$S1 : s_0, s_1, ..., s_i, ..., s_{i+k}, ..., s_n \tag{1}$$

in the bilingual corpus, where $s_i, ..., s_{i+k}$ is a phrase of interest, and the corresponding target language sentence S2 is

$$S2 : t_0, t_1, ..., t_j, ..., t_{j+l}, ..., t_m \tag{2}$$

the values of j and l are to be determined.

Then the segment we try to obtain is the target fragment $\hat{F}_T$ with the highest probability of all possible fragments of S2 to be a mutual translation with the given source fragment, or

$$\hat{F}_T = \text{argmax}_{\{F_T\}} \left(p(s_i, ..., s_{i+k} \leftrightarrow t_j, ..., t_{j+l})\right) \tag{3}$$

All possible segments can be checked in $O(m^2)$ time, where m is the target language length, because we will check m 1-word segments, $m-1$ two-word segments, and so on. If we bound the target language n-grams to a maximal length k, then the complexity is linear, i.e. $O(km)$.

The score of the best possible alignment is computed as follows: Let L_T be the Target Language Vocabulary, s a source word, t_i be target segment words, and $V = \{t_i \in \{L_T\} | i \geq 1\}$ the translation word set of s,

We define the *translation relation probability* $p(Tr(s) \in \{t_0, t_1, ..., t_k\})$ as follows:

1. $p(Tr(s) \in \{t_0, t_1, ..., t_k\}) = \max(p(t_i|s))$ for all $t_i \in \{t_0, t_1, ..., t_k\}$ when $\{t_i | t_i \in \{t_0, t_1, ..., t_k\}\}$ is not empty.
2. $p(Tr(s) \in \{t_0, t_1, ..., t_k\}) = 0$ otherwise.

Then the score of the best alignment is

$$S_{\hat{F}_T} = \max_{\{F_T\}} S_{F_T} \tag{4}$$

where the score can be written as two components

$$S_{F_T} = P_1 \times P_2 \tag{5}$$

which can be further specified as

$$P_1 = \left(\prod_{m=0}^{k} \max\left(p\left(Tr(s_{i+m}) \in \{t_{j...j+l}\}\right), \epsilon\right) \right)^{\frac{1}{k+1}} \tag{6}$$

$$P_2 = \left(\prod_{n=0}^{l} \max\left(p\left(Tr(t_{j+n}) \in \{s_{i...i+k}\}\right), \epsilon\right) \right)^{\frac{1}{l+1}} \tag{7}$$

where ϵ is a very small probability used as a *smoothing value*.

1.2 Length Penalty

The ratio between source and target segment (n-gram) lengths should be comparable to the ratio between the lengths of the source and target sentences, though certainly variation is possible. Therefore, we add a penalty function to the alignment probability that increases with the discrepancy between the two ratios.

Let the length of the source language segment be i and the length of a target language segment under consideration be j. Given a source language sentence length of n (in the corpus sentence containing the fragment) and its corresponding target language length of m. The *expected target segment length* is then given by $\hat{j} = i \times \frac{m}{n}$. Further defining an *allowable difference* AD, our implementation calculates the length penalty LP as follows, with the value of the exponent determined empirically:

$$LP_{F_T} = \min\left(\left(\frac{|j - \hat{j}|}{AD}\right)^4, 1\right) \tag{8}$$

The score for a segment including the penalty function is then:

$$S_{F_T} \leftarrow S_{F_T} \times (1 - LP_{F_T}) \tag{9}$$

Note that, as intended, the score is forced to 0 when the length difference $|j - \hat{j}| > AD$.

1.3 Distortion Penalty

For closely-related language pairs which tend to have similar word orders, we introduce a distortion penalty to penalize the alignment score of any candidate target fragment which is out of the expected position range. First, we calculate C_E, the expected center of the candidate target fragment using C_{F_S}, the center of the source fragment and the ratio of target- to source-sentence length.

$$C_E = C_{F_S} * \frac{m}{n} \tag{10}$$

Then we calculate an allowed distance limit of the center $D_{allowed}$ using a constant distance limit value DL and the ratio of actual target sentence length to average target sentence length.

$$D_{allowed} = DL * \frac{m}{m_{average}} \quad (11)$$

Let D_{actual} be the actual distance difference between the candidate target fragment's center and the expected center, and set

$$S_{F_T} \leftarrow \begin{cases} 0, \text{if} D_{actual} \geq D_{allowed} \\ \frac{S_{F_T}}{(D_{actual} - D_{allowed} + 1)^2}, \text{otherwise} \end{cases} \quad (12)$$

Furthermore, we think that we can apply this penalty to language pairs which have lower word-order similarities than e.g. French-English. Because there might exist certain positional relationships between such language pairs, if we can calculate the expected position using each language's sentence structure, we can apply a distortion penalty to the candidate alignments.

1.4 Anchor Context

If the adjacent words of the source fragment and the candidate target fragment are translations of each other, we expect that this alignment is more likely to be correct. We boost S_{F_T} with the anchor context alignment score S_{AC_p},

$$S_{AC_p} = P(s_{i-1} \leftrightarrow t_{j-1}) * P(s_{i+k} \leftrightarrow t_{j+l}) \quad (13)$$

$$S_{F_T} \leftarrow (S_{F_T})^{\lambda} * (S_{AC_p})^{1-\lambda} \quad (14)$$

Empirically, we found this combination gives the best score for French-English when $\lambda = 0.6$ and for Romanian-English when $\lambda = 0.8$, and leads to better results than the similar formula

$$S_{F_T} \leftarrow \lambda * S_{F_T} + (1 - \lambda) * S_{AC_p} \quad (15)$$

2 Experimental Design

In previous work (Kim et al., 2005), we tested our alignment method on a set of French-English sentence pairs taken from the Canadian Hansard corpus and on a set of English-Chinese sentence pairs, and compared the results to human alignments. For the present workshop, we chose to use the Romanian-English data which had been made available.

Due to a lack of time prior to the period of the shared task, we merely re-used the parameters which had been tuned for French-English, rather than tuning the alignment parameters specifically for the development data.

SPA was run under three experimental conditions. In the first, labeled "SPA (c)" in Tables 1 and 2, SPA was instructed to examine only contiguous target phrases as potential alignments for a given source phrase. In the second, labeled "SPA (n)", a noncontiguous target alignment consisting of two contiguous segments with a gap between them was permitted in addition to contiguous target alignments. The third condition ("SPA (h)") examined the impact of a small amount of manual alignment information on the selection of contiguous alignments. Unlike the first two conditions, the presence of additional data beyond the training corpus forces SPA(h) into the Unlimited Resources track.

We had a native Romanian speaker hand-align 204 sentence pairs from the training corpus, and extracted 732 distinct translation pairs from those alignments, of which 450 were already present in the automatically-generated dictionaries. The new translation pairs were added to the dictionaries for the SPA(h) condition and the translation probabilities for the existing pairs were increased to reflect the increased confidence in their correctness. Had more time been available, we would have investigated more sophisticated means of integrating the human knowledge into the translation dictionaries.

3 Results and Conclusions

Table 1 compares the performance of SPA on what is now the development data against the submissions with the best AER values reported by (Mihalcea and Pedersen, 2003) for the participants in the 2003 workshop, including CMU, MITRE, RALI, University of Alberta, and XRCE [1]. As SPA generates only SURE alignments, the values in Table 1 are SURE alignments under the NO-NULL-Align scoring condition for all systems except Fourday, which did not generate SURE alignments.

Despite the fact that SPA was designed specifically for phrase-to-phrase alignments rather than the

[1]Citations for individual participants' papers have been omitted for space reasons; all appear in the same proceedings.

Method	Prec%	Rec%	F1%	AER
SPA (c)	64.47	62.68	63.56	36.44
SPA (n)	64.38	62.70	63.53	36.47
SPA (h)	64.61	62.55	63.56	36.44
Fourday	52.83	42.86	47.33	52.67
UMD.RE.2	58.29	49.99	53.82	46.61
BiBr	70.65	55.75	62.32	41.39
Ralign	92.00	45.06	60.49	35.24
XRCEnolm	82.65	62.44	71.14	28.86

Table 1: Romanian-English alignment results (Development Set, NO-NULL-Align)

word-to-word alignments needed for the shared task and was not tuned for this corpus, its performance is competitive with the best of the systems previously used for the shared task. We thus decided to submit runs for the official 2005 evaluation, whose resulting scores are shown in Table 2.

On the development set, noncontiguous alignments resulted in slightly lower precision than contiguous alignments, which was not unexpected, but recall does not increase enough to improve F1 or AER. The modified dictionaries improved precision slightly, as anticipated, but lowered recall sufficiently to have no net effect on F1 or AER.

The evaluation set proved to be very similar in difficulty to the development data, resulting in scores that were very close to those achieved on the dev-test set. Noncontiguous alignments again proved to have a very small negative effect on AER resulting from reduced precision, but this time the altered dictionaries for SPA(h) resulted in a substantial reduction in recall, considerably harming overall performance.

After the shared task was complete, we performed some tuning of the alignment parameters for the Romanian-English development test set, and found that the French-English-tuned parameters were close to optimal in performance. The AER on the development test set for the SPA(c) contiguous alignments condition decreased from 36.44% to 36.11% after the re-tuning.

4 Future Work

Enhancements in the extraction of word-to-word alignments from what is fundamentally a phrase-to-phrase alignment algorithm could probably further improve results on the Romanian-English data. We also intend to investigate principled, seamless integration of manual alignments and dictionaries with probabilistic ones, since the *ad hoc* method proved detrimental. Finally, a more detailed performance analysis is in order, to determine whether the close balance of precision and recall is inherent in the bidirectionality of the algorithm or merely coincidence.

Method	Prec%	Recall%	F1%	AER%
SPA (c)	64.96	61.34	63.10	36.90
SPA (n)	64.91	61.34	63.07	36.93
SPA (h)	64.60	60.54	62.50	37.50

Table 2: Evaluation results (NO-NULL-Align)

5 Acknowledgements

We would like to thank Lucian Vlad Lita for providing manual alignments.

References

Ralf D. Brown. 2004. A Modified Burrows-Wheeler Transform for Highly-Scalable Example-Based Translation. In *Machine Translation: From Real Users to Research, Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, volume 3265 of *Lecture Notes in Artificial Intelligence*, pages 27–36. Springer Verlag, September-October. `http://www.cs.cmu.edu/~ralf/papers.html`.

Jae Dong Kim, Ralf D. Brown, Peter J. Jansen, and Jaime G. Carbonell. 2005. Symmetric Probabilistic Alignment for Example-Based Translation. In *Proceedings of the Tenth Workshop of the European Assocation for Machine Translation (EAMT-05)*, May. (to appear).

Daniel Marcu and William Wong. 2002. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, July. `http://www.isi.edu/~marcu/papers.html`.

Rada Mihalcea and Ted Pedersen. 2003. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10. Association for Computational Linguistics, May.