Review of Statistical Word Alignment Techniques
Lambert, P.
Research Report LSI-04-42-R
Review of Statistical Word Alignment Techniques

Patrik Lambert  (lambert@lsi.upc.es)

TALP Research Center
Universitat Politècnica de Catalunya
Jordi Girona Salgado, 1-3, 08034 Barcelona, Spain
lambert@talp.upc.es

Abstract

The notion of word alignment of a bilingual parallel corpus is defined and the difficulties of the task are highlighted. Alignment evaluation procedures are described. First, straightforward, baseline methods are presented. Then state-of-the-art systems are described and qualitatively or quantitatively compared.

1 Introduction

This review intends to describe the state of the art of the techniques that employ statistical or heuristic methods to obtain a word alignment of a bilingual sentence-aligned corpus. As word alignment of a bilingual sentence pair, we refer to the set of links between lexical units that are translation of each other. Lexical units can be words or groups of words (phrases). This definition doesn’t specify a unique set of links, since the correspondence between words or phrases is subjective and thus ambiguous.

Word aligned corpora are useful in a variety of fields. An obvious one is the automatic extraction of bilingual lexica and terminology (Smadja et al., 1996; Ribeiro et al., 2001). Word alignments are also used in statistical machine translation (Brown et al., 1993; Wang and Waibel, 1998; Och and Ney, 2000). (Och and Ney, 2000) have shown that the translation quality depends directly on the word alignment quality. Word sense disambiguation is another application (Diab and Resnik, 2002), since ambiguities are distributed differently in different languages. Word aligned corpora can also help for the transfer of language tools to new languages. In (David Yarowsky and Wicentowski, 2001), text analysis tools such as morphologic analysers or part-of-speech taggers are projected to languages where such resources don’t exist. (Kuhn, 2004) presents a study on ways of exploiting statistical word alignment for grammar induction.

With this variety of applications, the notion of alignment quality depends on the application in which the alignment will be used. The required lexical units also depend on the task. In machine translation, phrases are typically preferred over single words and the recall is important. In extraction of bilingual lexica, high precision single-word pairs are usually detected. All these applications perform better if the alignment quality is improved, so achieving better alignments is a central issue in various natural language processing tasks.

However, the alignment between two sentences can be quite complicated. It can include word or phrase reorderings, omissions, insertions and word-to-phrase alignments. Therefore, a very general alignment representation is needed. In particular, an ideal representation should allow various (possibly non-contiguous) source words to be aligned with various (possibly non-contiguous) target words, as sometimes it is not possible to align words of a phrase pair. For instance, in the English-Spanish phrase pair (two o’clock, las dos), if “two” should be aligned with “dos”, it doesn’t make sense to align “o’clock” with “las”. It’s actually the entire phrase “two o’clock” that should be aligned with “las dos”.

1
A general definition of an alignment between two word strings is as follows: an alignment is a subset of the Cartesian product of the word positions, i.e. an alignment \( \mathcal{A} \) is defined as:

\[
\mathcal{A} \subseteq \{(j, i) : j = 1, \ldots, J; i = 1, \ldots, I\}
\]  

if the source and target sentences have got respectively \( I \) and \( J \) words. Some words are not aligned to any other word. The definition can be modified to take this into account explicitly by adding links of the type \((j, 0)\) or \((0, i)\) ("NULL" links). This is equivalent to adding a "NULL" word at position 0 of the source and target sentences. In practice, the representation of equation 1 is hard to implement because it induces too many alignments to be considered. There are \( IJ \) possible connections between words. Hence, there are \( 2^{IJ} \) possible alignments. A typical way to restrict the problem, proposed by (Brown et al., 1993), is to assign each source word to exactly one target word. The alignment is a mapping from source positions \( j \) to target language positions \( i = a_j \). Alignment \( a_j = 0 \) with an "empty" (Null) word \( e_0 \) is assigned to the source words that are not aligned to any target words. In this way, the number of possible alignments is limited to \((J + 1)^I\).

A good overview of the state of the art in the domain was given by the HLT-NAACL 2003 Workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond". The workshop included a word alignment evaluation task (Mihalcea and Pedersen, 2003). The participating teams were provided with training data consisting of sentence aligned parallel texts and were evaluated with manually aligned test data. Two sets of training data were made available:

- English-French parallel texts consisting of about 20 millions words, extracted from the Canadian Hansards (parliament debates).
- Rumanian-English parallel texts: 1 million words collected from the Web (mainly in Rumanian newspaper archives), Orwell's 1984 and the Rumanian Constitution.

All these resources, as well as code for evaluation of the submitted alignments, are available from the web site of the workshop (http://www.cs.unt.edu/~rada/wpt). This allows future systems to be compared to those that participated to the shared task.

Before comparing various systems, we will see in section 2 the evaluation measures that allow a quantitative comparison of the alignments that are produced. In section 3, we will present very simple methods based on direct computation of various type of distances between two words, without any training phase. In sections 4 and 5, we will describe state-of-the-art systems.

The different statistical word alignment systems are distinguished in the way they accomplish the following tasks: modelling the alignment problem, estimate the parameters of the model and find, for a given sentence pair, the optimal alignment according to the model and its parameters. There are generally two approaches, called association approach and estimation approach in (Tiedemann, 2003). In the association approach, the alignment is based on association scores between lexical units. In the estimation approach, the parameters are usually modelled as hidden parameters in a statistical translation model. (Och and Ney, 2003) call heuristic models the most simple models of the association approach, and statistical alignment models the models of the estimation approach. Actually both approaches use statistics and a form of association score between lexical units. However, in the association approach, the scores are evaluated counting the associations throughout the parallel training corpus, whereas in the estimation approach, they result from statistical estimation theory and are adjusted such that the likelihood on the parallel training corpus is maximal. So the parameter estimation method is one of the differences between the two approaches. We will refer to these approaches as “association approach” (section 4) and “statistical alignment models” (section 5).

2 Evaluation

The evaluation methods employed in the evaluation of the HLT-NAACL 2003 workshop (Mihalcea and Pedersen, 2003) reflected the measures typically used in the literature. Submitted
alignments are compared to a manually aligned reference corpus (gold standard) and scored with respect to precision, recall, F-measure and Alignment Error Rate (AER). An inherent problem of the evaluation is the ambiguity of the manual alignment task. The annotation criteria depend on each annotator. Therefore, (Och and Ney, 2003) introduced a reference corpus with explicit ambiguous (called P or Possible) links and unambiguous (called S or Sure) links. Given an alignment $A$, and a gold standard alignment $G$, we can define sets $A_S$, $A_P$ and $G_S$, $G_P$, corresponding to the sets of Sure and Possible links of each alignment. The set of Possible links is also the union of S and P links, or equivalently $A_S \subseteq A_P$ and $G_S \subseteq G_P$.

The following measures are defined (where $T$ is the alignment type, and can be set to either S or P):

$$P_T = \frac{|A_T \cap G_T|}{|A_T|}, \quad R_T = \frac{|A_T \cap G_T|}{|G_T|}, \quad F_T = \frac{2P_T R_T}{P_T + R_T}$$

$$AER = 1 - \frac{|A_P \cap G_S| + |A_P \cap G_P|}{|A_P| + |G_S|}$$

2.1 Unlinked words representation

The scores are greatly affected by the representation of NULL links (between a word and no other word: whether they are assigned an explicit link to NULL or removed from the alignments). Explicit NULL links contribute to a higher error rate because in this case the errors are penalised twice: for the incorrect link to NULL and for the missing link to the correct word. Thus both submitted and answer alignments must have the same alignment mode, which can be one of the following:

- null-align, where each word is enforced to belong to at least one alignment; if a word doesn’t belong to any alignment, a NULL Possible link is assigned by default.

- no-null-align, where all NULL links are removed from both submission and gold standard alignments.

2.2 Link weights

In the evaluations of (Och and Ney, 2000; Mihalcea and Pedersen, 2003), each link contributes with the same weight to the count of the various sets. This tends to give more importance to the words aligned in groups than to the words linked only once. To correct this effect, (Melamed, 1998) proposes to attach a weight to each link. The weight $w(x, y)$ of a link between two words $x$ and $y$ would be inversely proportional to the number of links ($num\_links$) in which $x$ and $y$ are involved:

$$w(x, y) = \frac{1}{2} \left[ \frac{1}{num\_links(x)} + \frac{1}{num\_links(y)} \right]$$

(2)

3 Baselines

They are very simple methods based on geometric distance, word edit distance or frequency ratios. Before relating a study where these techniques were applied to word alignment, we will recall how they were introduced for the alignment of sentences.

Early text alignment systems focused on finding corresponding sentences. Sentences were an easier starting point because their order is usually conserved during translation. In 1991, (Brown et al., 1991) and (Gale and Church, 1991b) independently presented systems aligning sentences with high accuracy just by matching sentence sequences with similar lengths. These two systems were improved (or new ones were proposed) to work better with noisy parallel corpora and with languages that don’t use phonetically based alphabet, like Chinese or Japanese (for more details see (Melamed, 2001), section 1.2.3: “Previous Work”). Most of these systems took advantage of anchor points (points at which there is evidence of a correspondence between both sides of the corpus). These anchor points can be provided by:
cognates (Church, 1993). Cognates are word of similar spelling.
a seed translation lexicon (Wu, 1994). It is a set of word pairs with high probability to be
translation of each other.
a statistical translation model (Chen, 1993).
dot-plot geometry (Fung and McKeown, 1994). More precisely in this case: distribution
vectors associated to each word.
(Melamed, 1996) generates candidate points of correspondence using cognates or a seed trans-
lation lexicon or otherwise with other applicable methods. Then the chain of anchor points
whose geometric arrangement most resembles the typical arrangement is selected.
Since these approaches have been used with great success in both finding anchor points
and aligning sentences, (Henderson, 2003) studied their application to word alignment. It
also provides a starting point for measuring the technological advances.

3.1 Length Ratios
The hypothesis of (Gale and Church, 1991b; Brown et al., 1991) transposed from sentences to
words would be that short words such as stop words tend to align with short words and long
words such as names tend to align with long words. (Henderson, 2003) codify the observation
as a distance between the word \(l_i\) at position \(i\) on the left hand side (LHS), and the word
\(r_j\), at position \(j\) on the right hand side (RHS). The distance is based on the length \(L(l_i)\) and
\(L(r_j)\) of the two words:

\[
D_{len}(i, j) = 1 - \frac{4L(l_i) * L(r_j)}{[L(l_i) + L(r_j)]^2}
\]  

(3)

3.2 Edit Distance
A method to find cognates (in languages that share character sets) is by comparing the
edit distance between words. The edit distance is the minimum number of character edits
(insertions, deletions, substitutions) required to transform one word into another, normalised
by lengths. The edit distance rate (edits per character) writes:

\[
D_{edit}(i, j) = \frac{edits(l_i, r_j)}{L(l_i) + L(r_j)}
\]  

(4)

3.3 Dot-plot Geometry
Transposing the approach of (Fung and McKeown, 1994; Melamed, 1996) to word alignment,
(Henderson, 2003) created distance metrics to incorporate the knowledge that all of the aligned
pairs use roughly the same word order. The distance of the pair of words from a diagonal in the
dot-plot was used. The distance \(L_1\) from a point \((i, j)\) to the diagonal from \((0, 0)\) to \((I, J)\)
was taken to be:

\[
d_{L_1}(i, j, J) = \frac{|i - j|}{J}
\]  

(5)

For instance, the normalised distance of the \((i, j)\) pair of tokens to the diagonal on the word
dot-plot is

\[
d_{edit}(i, j) = d_{L_1}(i, L_w(l), j, L_w(r))
\]  

(6)

3.4 Likelihoods
Three likelihood-based distance metrics were investigated. The “best”-scoring metric uses the
relative likelihood of the aligned pairs of words

\[
D_{freqatio}(i, j) = 1 - \frac{\min(c(l_i, LHS), c(r_j, RHS))}{\max(c(l_i, LHS), c(r_j, RHS))}
\]  

(7)

where \(c(l_i, LHS)\) is the number of times the word \(l_i\) was seen in the LHS of the aligned corpus.
3.5 Experiments and Conclusions

A threshold was used to turn the distance metrics into a classification rule. The various baseline systems were evaluated on the same test corpora as in the HLT-NAACL 2003 evaluation task. The individual scores of the different baseline systems give insight into the relative contributions of the features they exploit. Moreover, the typical performance of the baseline aligners could be compared to that of state-of-the-art systems.

Concerning the relative contribution of the various features, (Henderson, 2003) concludes that “Word length matching appears to be the least important feature, followed by character edit distance, and geometric dot-plot distances appear to contribute most strongly to alignment performance”. The frequency ratios perform poorer. On the other hand, the alignment error rate (AER) lies in the range 55%-75% (English-French data) and 65%-85% (Rumanian-English data). In the 2003 evaluation, the participating systems received scored in the range 6%-38% (English-French data) and 30%-45% (Rumanian-English data). So the baseline metrics are very basic compared to the techniques used in more advanced systems.

Actually a useful information to detect correspondences between words, which is present in a sentence-aligned parallel corpus but not taken into account in the baseline methods, are the co-occurrence counts. All of the more refined word or phrase alignment models take advantage in some way of co-occurrence information. In the next section we will discuss how to count co-occurrences and describe some of the association measures presented in the literature.

4 Association approach

4.1 Co-occurrence counts

Since we suppose that the corpus we work with is aligned at the sentence level, two words obviously co-occur if they are found in each side of the same sentence pair. If freq(l) and freq(r) are the frequencies of occurrence of the words l in the LHS and r in the RHS, common ways to count co-occurrences are

\[
\text{cooc}(l,r) = \text{freq}(l) \times \text{freq}(r) \tag{8a}
\]

\[
\text{cooc}(l,r) = \min[\text{freq}(l), \text{freq}(r)] \tag{8b}
\]

\[
\text{cooc}(l,r) = \max[\text{freq}(l), \text{freq}(r)] \tag{8c}
\]

Method (8a) works fine when the frequencies are not greater than one, but can produce undesirable results if they are greater than one. For example, if \( \text{freq}(l) = \text{freq}(r) = 3 \), the number of co-occurrences would be 9. The co-occurrences are used as cues of the correspondence between words or phrases, which implies a one-to-one relationship. Thus it seems reasonable to count them with relation (8b).

4.2 Association Measures

There exists many measures that use the co-occurrence counts to compute association scores, a value that indicates the amount of statistical association between two words in a corpus (see http://www.collocations.de/AM/index.html). These association (or co-occurrence) measures are often expressed in terms of the cells of contingency tables, like table 1. In this table, \( a \) is the number of sentence pairs that contain both \( l \) and \( r \), \( b \) is the number of sentences that contain \( l \) but not \( r \), etc. If \( N \) is the total number of sentence pairs, the table can be calculated in the following way:

\[
a = \text{cooc}(l,r) \\
b = \text{freq}(l) - \text{cooc}(l,r) \\
c = \text{freq}(r) - \text{cooc}(l,r) \\
d = N - a - b - c \tag{9}
\]
<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>not r</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>not l</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Table 1: Contingency table for words \( l \) and \( r \).

The interpretation of the measures can be more intuitive if they are expressed in terms of occurrence probabilities, so both ways will be shown. The probability of seeing the two words/phrases in the same sentence pair is

\[
p(l, r) = \frac{\text{cooc}(l, r)}{N} = \frac{a}{a + b + c + d}
\]

(10)

And the marginal probabilities are

\[
p(l) = \frac{\text{freq}(l)}{N} = \frac{a + b}{a + b + c + d}, \quad p(r) = \frac{\text{freq}(r)}{N} = \frac{a + c}{a + b + c + d}
\]

(11)

We will describe only the measures cited further in this report. In (Gale and Church, 1991a), a \( \chi^2 \)-like statistic is used to identify word correspondences:

\[
\phi^2 = \frac{[p(l, r) - p(l)p(r)]^2}{p(l)p(r)[1 - p(l)][1 - p(r)]} = \frac{(ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)}
\]

(Fung and McKeown, 1994) use a t-score of word distribution vectors to generate a bilingual lexicon in a corpus divided in \( N \) segments:

\[
t = \frac{p(l, r) - p(l)p(r)}{\sqrt{\frac{1}{Np(l, r)}}} = \frac{ad - bc}{(a + b + c + d)^{\frac{1}{2}}}
\]

(13)

Other association measures used are variations of the Dice coefficient (Dice, 1945):

\[
dice = \frac{2\text{cooc}(l, r)}{\text{freq}(l) \text{freq}(r)} = \frac{2p(l, r)}{Np(l)p(r)} = \frac{2a}{(a + b)(a + c)}
\]

(14)

4.3 Alignment Methods

For each sentence pair, a matrix including the association scores between every word at every position is obtained. (Melamed, 2000) presented a method to obtain a word alignment from the association score matrix, the competitive linking algorithm. In a first step, the highest ranking word pair \( (i_j, r_j) \) is aligned. Then, the corresponding row and column are removed from the association score matrix. This procedure is iteratively repeated until every LHS or RHS language word is aligned. The resulting alignment contains only one-to-one links. (Och and Ney, 2003) have evaluated alignments obtained with a Dice-score association matrix and the competitive linking algorithm, and have compared them with alignments obtained with so-called statistical algorithms (HMM model and IBM models, see section 5.1). In the Hansards corpus (similar to the English-French data used in the HLT-NAACL 2003 evaluation mentioned in section 1), the AER is about 34% for the Dice coefficient, 26% for IBM model 1 and less than 10% for more refined models. Both IBM model 1 and the heuristic model (Dice coefficient + competitive linking) use only co-occurrence information. However, the statistical model performs better.

A great advantage of the association approach is its flexibility. It is easy to use another association metric or to add more information to the model. So many variations and extensions of this basic method have been done, and further improvements will be probably achieved in the future.
A first refinement of the co-occurrence score function is Melamed’s explicit noise model (Melamed, 2000). First, word pairs are scored with an association metric and linked. Then, they are re-scored considering the ratio of the number of times a word pair is linked compared to the number of times it co-occurs. Two words which are linked nearly every time they co-occur seem to a better translation than two words that are linked only half of the time they co-occur. Based on these ratios, the explicit noise model evaluates the probability $\lambda^+$ of a link occurring between words that are a true translation and the probability $\lambda^-$ of a link between words that are not a true translation. The final scoring function uses these two probabilities to compute the strength of the association between two word types.

In (Ahrenberg et al., 1998), the t-score is used as a basis, but the word-to-word approach is extended by considering also multi-word candidates. Some morphological information is also used (expressions with different suffixes might be treated equally), and weights are distributed depending on positions.

In (Tiedemann, 2003), the association score is obtained by the combination of various clues. An alignment clue $C_i$ is defined as a probability $P$ which represents a weighted association $A$ between two lexical units (words or phrases) $s$ and $t$:

$$C_i(s, t) = P_i(a_i) = w_i A_i(s, t)$$  \hspace{1cm} (15)

The association $A$ can be the Dice coefficient or a string similarity measure. Other clues can be estimated from word aligned training data: $C_j(s, t) = w_j \frac{f_{\text{frequency}}(f_s, f_t)}{f_{\text{frequency}}(f_s) + f_{\text{frequency}}(f_t)}$, where $f_s$ and $f_t$ are sets of features of $s$ and $t$, respectively. The features are based on POS labels, chunk labels and relative word position. The various clues are combined using the following addition rule:

$$C_{\text{all}}(s, t) = P(a_{\text{all}}) = P(a_1 \cup a_2 \cup \ldots \cup a_n)$$

$$P(a_1 \cup a_2) = P(a_1) + P(a_2) - P(a_1)P(a_2)$$  \hspace{1cm} (16)

The final alignment is obtained with a type of competitive linking algorithm that takes into account overlapping lexical units. This alignment can in turn be used to improve the feature-based clues. Thus various features have been combined to obtain a refined association matrix.

Another model using various type of features is the probability model of (Cherry and Lin, 2003). Contrary to the other models seen in this section, it calculates the probability of the entire alignment $A = \ell_T^T$, considered as a sequence of $T$ links:

$$P(A|E, F) = \prod_{t=1}^{T} P(\ell_t|E, F; \ell_{t-1}^t)$$  \hspace{1cm} (17)

$T$ is the total number of links, $E$ and $F$ are the two aligned sentences, $\ell_t$ is an individual link between two words. In this model, the probability of each link given its context is given by the product of two terms. One is the probability $Pr(\ell_t | e_t, f_t)$ of a link given the co-occurrence count of its words. The other term is a product of feature contributions. The features are link adjacency features and features extracted from a parse tree. The probability tables are calculated from a word aligned parallel corpus. The initial alignment is obtained with the $\phi^2$ association score and a competitive linking algorithm restricted to one-to-one links. The highest probability alignment is searched with a constrained best-first search algorithm. The constraints are the prohibition of one-to-many links and a cohesion constraint, which prohibits some crossing combinations in the alignment. This system obtained very good scores in the HLT-NAACL 2003 evaluation.

(de Gispert et al., 2004) use a similar method but eliminate the principal drawbacks of the one-to-one constraint, which can’t model phrases that must be treated as a whole. First they detect verb groups (with a few rules) and idiomatic expressions (with small monolingual dictionaries) and align them. The remaining words are aligned with an algorithm similar to that of (Cherry and Lin, 2003). They obtain alignments with a quality comparable to those produced by IBM model 4 (see section 5.1).
In conclusion, association measures, despite of their simplicity, give strong evidence of the correspondences between words/phrases in a parallel corpus. However, the most basic models don’t perform as well as the simplest statistical models. Anyway, co-occurrence measures are not sufficient to model the whole alignment complexity. Models taking advantage of the flexibility of the association approach combine the co-occurrence measures to other techniques, or other type of linguistic information, and can produce high quality alignments.

5 Statistical Alignment Models

In this section we focus on techniques in which the alignments are obtained as a by-product of a statistical translation process. The translation model typically tries to estimate the probability \( Pr(e | f) \) that a source sentence \( f \) be translated by the target sentence \( e \). A “hidden” alignment \( a \) is introduced to describe the connections between the words (or phrases):

\[
Pr(e | f) = \sum_a Pr(e, a | f)
\]  (18)

In practice, the alignment is usually restricted to a mapping from a source position \( j \) to a target position \( a_j \). In general, the statistical alignment model \( Pr(e, a | f) \) depends on a set of unknown parameters \( \theta \) which are trained by maximising the likelihood on a parallel corpus. If the training corpus consists of \( S \) sentence pairs \( \{(e_s, a_s) : s = 1, \ldots, S\} \), and using the notation \( Pr_\theta (e, a | f) \) to express the dependence of the alignment model on the unknown parameters:

\[
\hat{\theta} = \arg \max_\theta \prod_{s=1}^S \sum_a Pr_\theta (e_s, a | f_s)
\]  (19)

Typically, the training corpus is not aligned at a word level, and the EM algorithm (Dempster et al., 1977) or some approximate EM algorithm is used to perform this maximisation.

For a given sentence pair, among all possible alignments, the best one is obtained by maximisation of the alignment probability:

\[
\hat{a} = \arg \max_a Pr_\theta (e, a | f)
\]  (20)

This best alignment is called Viterbi alignment.

5.1 Single-Word Based Statistical Alignment models

(Och and Ney, 2003) review the main statistical translation models used that are based on word-by-word alignment and individual word translation, that is IBM models 1 to 5 (Brown et al., 1993), HMM model (Vogel et al., 1996) and a log-linear combination of HMM model and IBM model 4 called model 6.

Model 1 is a simple word translation model, which only makes use of co-occurrence of word pairs. The information it uses corresponds exclusively to that of cell \( a \) in the contingency table seen in section 4.2. It is initialised with a uniform distribution of word translation probabilities. Model 2 adds local dependencies by introducing position parameters to the translation model. In model 3, fertility parameters are introduced. Fertility parameters represent the probability of words to be aligned to a certain number of corresponding word. Using fertility probabilities, multiple connections to one word can be penalised or supported. Model 4 includes additional dependencies on the previous alignment and on the word classes of surrounding words in order to handle word groups, which tend to stick together. Model 3 and 4 have deficiency problems, which means that parts of the probability distribution are reserved for impossible events. Model 5 gets rid of the deficiency problems. The HMM model is similar to model 2, but the position parameters depend on the previous word. Actually IBM’s models 1, 2 and 3 can be formalised in a zero-order HMM (where model 3 has additional fertility parameters) and model 4 and 5 can be expressed as a first order HMM with additional fertility parameters (Och and Ney, 2000). The HMM model predicts the distance between subsequent
source language positions, whereas Model 4 predicts the distance between subsequent target language positions. Thus the HMM makes use of the locality in the source language whereas model 4 makes use of locality in the target language. Model 6, which combines HMM and model 4, takes into account both types of dependencies.

The models are trained on a bilingual corpus with EM algorithm, “bootstrapping” from a simpler model to a more complex model. For the models from 3 on, EM algorithm runs on a selection of alignments close to the Viterbi alignment.

(Och and Ney, 2003) systematically compare the results obtained with training schemes using those models. In general, very important ingredients of a good model seem to be a first-order dependence of word positions and a fertility model. It is important to bootstrap the refined models with good initial parameters (HMM instead of model 2). The best results are obtained with model 6 and 1°H3°4°6° training scheme (that is 5 iterations of the EM algorithm with model 1 followed by 5 iterations with HMM model, etc.).

The same authors have also measured the effect of various techniques that, independently of the models used, can improve the results. These techniques are:

**number of alignment in training** The effect measured here is that of including more alignments in the training of the fertility-based alignment models: the neighbourhood of the Viterbi alignment, or an even larger set of high probability alignments, called pegged alignment in (Brown et al., 1993). In general, the effect of adding the neighbourhood of the Viterbi alignment is greater than the effect of adding pegged alignments. In any case, the improvement is more significant if the quality of the starting point used for training the fertility-based alignment models is lower (model 2 instead of HMM model). So (Och and Ney, 2003) conclude that using more alignments in training is a way to avoid a poor local optimum. Regarding performance, including the neighbourhood of the Viterbi alignment doesn’t increase much the training time, whereas pegging increases it considerably.

**smoothing** Alignment and fertility probabilities are smoothed to solve the problems of rare words and over-fitting.

**word classes** The alignment parameters of HMM, Model 4 and Model 5 include a dependence on the word classes of the surrounding words. These word classes can be learned automatically from bilingual corpora using clustering techniques (Och, 1999).

**conventional bilingual dictionary** Entries of a conventional bilingual dictionary are added to the corpus.

**symmetrisation** : The Viterbi alignment is calculated in both source-target and target-source directions. An alignment matrix is calculated putting together information from both alignments.

Increasing the corpus size can decrease dramatically the alignment error rate. Smoothing and symmetrisation have a significant effect on the alignment quality. Adding entries of a bilingual dictionary, making the alignment models depending on word classes (Model 4 and 5) or increasing the number of selected alignments only have a minor effect on the alignment quality.

Several extensions to IBM models are tested in (Toutanova et al., 2002), namely introducing a “tag translation” probability as a product to alignment and lexicon probabilities, simulating fertility for the HMM model with a new probability, and a new treatment of the NULL word. Its combinations shows promising results (in Alignment Error Rate), when compared to HMM model and IBM4, especially when there are few training sentences. Other dependencies from part-of-speech tags are described, but no results are reported.

We saw methods to improve the quality of statistical word alignment models, but word-based models have inherent limitations. Translation of multiple word phrases which do not decompose easily into word for word translations are not possible. If the parameters are tied
to words, it is also more difficult to model large-scale re-orderings. In the next section, we will see techniques that allow many-to-many alignments in order to address these problems.

5.2 Phrase Based Statistical Alignment models

A first idea to consider phrases in statistical alignment models is to group phrases in the corpus and train them with the word based statistical models, treating each group as a word. (Watanabe et al., 2002) use Hierarchical Phrase Alignment (HPA) (Inamura, 2001) to segment both sides of the corpus into chunks, and use IBM models 1 to 4, treating each chunk as a word. HPA finds the correspondence of sub-trees between source and target language parse trees based on partial parse (non-terminal) results. The chunks were obtained by separating the sequence of words before and after the phrase alignments. This method is expected to model successfully many-to-many word relations. It should also improve the distortion model since the length of the sentences are shortened. An advantage of this method for the lexical model ($Pr(e_i | f_j)$) is that correspondences between phrases are higher than between single words. However, this model suffers from a data sparseness problem as the vocabulary size is increased by the chunked words.

In (Wang and Waibel, 1998), the IBM models are modified to add phrase based parameters to the set of word based parameters. The sentences are decomposed into a sequence of n phrases (by applying iteratively a bilingual mutual information clustering algorithm and a phrasing operator). The translation model has got phrase-based parameters, that describe the alignment between phrases, and single-word-based parameters, that account for alignments within phrases and for individual word translations. The parameters are an inter-phrase alignment probability, a phrase beginning position probability and the single-word-based parameter, which are the same as in IBM model 4 (fertility, word translation and word permutation probabilities). All the parameters are trained together with the EM-algorithm. An interesting result of this work shows that this model captures better the phrase order in a sentence than a typical word-based model. For model 2, the first German word has a 95% probability to be aligned with one of the 3 first positions ($j=0,1,2$) of the English string (with 86% for $j=1$). In the present model, the probability mass of alignment of the first German phrase is more scattered (29% for $j=1$, 25% for $j=2$).

The model of (Wang and Waibel, 1998) takes phrases into account at the alignment and reordering level, but not at a lexical level. It assumes indeed that source words are individually translated into target words, which can yield unintuitive translation probabilities.

In the translation model of (Marcu and Wong, 2002), lexical correspondences can be established not only at the word level, but at the phrase level as well. The parameters of the model are the phrase translation probabilities $t(\tilde{e}, \tilde{f})$ and the probabilities $d(pos(f^n), pos_cm(\tilde{e}))$ of distortion between the $k^{th}$ word of $f$ and the centre of mass of $\tilde{e}$. For the $(\tilde{e}, f)$ table, all the unigrams are selected and only a selection of high-frequency n-grams (the value of n is not indicated). This approach allows the lexicon probabilities to be established between groups of words, so that the translation of phrases are trained at the same time as the translation of words. The inconvenient is the resulting size of the $t$ table and the cost of the training procedure.

Unfortunately these models have been designed for machine translation and the alignments have not been evaluated, so they can’t be quantitatively compared. However some features can be highlighted. Phrase based statistical alignment models solve the problems inherent to single-word based statistical alignment models: they can model various source words aligned to various target words and improve the reordering model. However, considering groups of words increases the vocabulary size and the number of possible alignments. Thus there is risk to have data sparseness problems and inefficient algorithms. It depends on how the sentences are segmented into phrases, and how many phrases are produced, which are key issues in phrase based statistical alignment modelling. Typical methods use co-occurrences, clustering or high probability n-grams. (Watanabe et al., 2002) employ partial parse results. Whatever the segmentation method, a reasonable strategy seems to be to select phrases only when the
alignment fails to be represented with single words (like for complex noun phrases, phrasal verbs, idiomatic expressions or other phrasal constructions that should not be split up in an alignment process). In this way the vocabulary is increased only with a restricted set of phrases that can fix the main errors of single-word based models, and inefficiency and data sparseness problems can be limited.

6 Conclusions

Word alignment can be used in a variety of applications. Each application can have slightly different requirements for the alignment task, so the first duty is to precisely these requirements.

Another point to keep in mind is that alignments can be ambiguous, which makes their evaluation more difficult. Alignment evaluation has to be carried out with care, since various factors can modify the scores, like the representation of NULL links.

Word alignment is a complicated task and is still an intense research field. We saw that simple baseline methods produce very poor alignments and need to be refined. We reviewed systems that we classified with respect to two types of approach: the models of the association approach and the statistical alignment models. The strength of the statistical alignment models may be that their parameter estimation is based on a rigorous mathematical theory. The parameters are trained by maximising the likelihood on the parallel training corpus. The main inconvenient of statistical alignment models is that adding new types of parameters implies reconsidering the whole model. So introducing additional information to the model is expensive, which makes them more difficult to improve. In the association approach, the training of parameters is much simpler since it doesn’t involve any optimisation on the whole training corpus. However it is straightforward to add new features to the model. In particular, valuable linguistic information like part-of-speech or syntactic analysis can be easily taken into account.

The size of the considered lexical unit is also an important issue. Single word can be treated considerably more easily, but they fail to model fixed multi-word expressions. So the model should contain at least some parameters based on larger structures. However considering larger lexical units introduces two challenges. First, to extract the phrases. Second, to carry out the training in spite of the increase of the vocabulary size, of the number of possible alignment combinations and the higher data sparseness.

7 Acknowledgements

This work has been (partially) granted by the Spanish Government under grant TIC2002-04447-C02.

The author want to thank Adria de Gispert for his valuable contribution to this paper.

References


Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

Research Reports - 2004

- LSI-04-2-R : Comparison of Methods to Predict Ozone Concentration, Orozco, J.
- LSI-04-3-R : Towards the definition of a taxonomy for the cots product’s market, Ayala, Claudia P.
- LSI-04-4-R : Modelling Coalition Formation over Time for Iterative Coalition Games, Mérida-Campos, C. and Willmott, S.
- LSI-04-6-R : An Analysis Pattern for Electronic Marketplaces, Queralt, A. and Teniente, E.
- LSI-04-7-R : Exploring Dopamine-Mediated Reward Processing through the Analysis of EEG-Measured Gamma-Band Brain Oscillations, Vellido, A. and El-Deredy, W.
- LSI-04-9-R : Similarity and Dissimilarity Concepts in Machine Learning, Orozco, J.
- LSI-04-10-R : A Framework for the Definition of Metrics for Actor-Dependency Models, Quer, C. and Grau, G. and Franch, X.
- LSI-04-15-R : Evaluation and symmetrisation of alignments obtained with the Giza++ software, Lambert, P. and Castell, N.
- LSI-04-16-R : A note on the use of topology extensions for provoking instability in communication networks, Blesa, M.J.
- LSI-04-17-R : An ISO/IEC-compliant Quality Model for ER Diagrams, Costal, D. and Franch, X.
- LSI-04-19-R : Adding Efficient and Reliable Access Paths to the JCF, Marco, J. and Franch, X.
- LSI-04-20-R: *Exploiting Simple Corporate Memory in Iterative Coalition Games*, Mérida-Campos, C. and Willmott, S.
- LSI-04-23-R: *Chromatic number in random scaled sector graphs*, Díaz, J. and Sanwalani, V. and Serna, M. and Spirakis, P.
- LSI-04-25-R: *Open Source environment to define constraints in route planning for GIS-T*, Pérez, L. and Silveira, A. da M.
- LSI-04-26-R: *A basic repository of operations for the refinement of general ontologies*, de Palol, X.
- LSI-04-27-R: *Tetrahedral mesh subdivision based on underlying volume data*, Rodríguez, L. and Navazo, I. and Vinacua, A.
- LSI-04-29-R: *Smaller kernels for hitting set problems of constant arity*, Nishimura, N. and Ragde, P. and Thilikos, D.M.
- LSI-04-30-R: *Searching Spatial Sense in the Ontological World: Discovering Spatial Objects*, Morocho, V and Pérez, L. and Saltor, F.
- LSI-04-34-R: *SVMTool: A general POS tagger generator based on Support Vector Machines*, Giménez, J. and Márquez, LL.
- LSI-04-35-R: *A distributed and mobile component system based on the ambient calculus*, Mylonakis, N. and Orejas, F.
- LSI-04-37-R: *The AlignmentSet Toolkit*, Lambert, P.
- LSI-04-38-R: *Integración de Fuentes de Datos espaciales: análisis e implementación de una Ontología de términos espaciales: Primera Parte - Creación de una Ontología*, Ramos, Erik G.
- LSI-04-40-R: *Kernels on Structured Domains*, valentin, L.
• LSI-04-42-R: *Review of Statistical Word Alignment Techniques*, Lambert, P.
• LSI-04-43-R: *Algoritmos genéticos en el problema de la solución deseada: Optimización de parámetros*, Barreiro, E. and Joan-Arinyo, R. and Luzón, M.V.

Hardcopies of reports can be ordered from:

Núria Sanchez
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Campus Nord, Mòdul C6
Jordi Girona Salgado, 1-3
08034 Barcelona, Spain
nurias@lsi.upc.es

See also the Departament WWW pages, http://wwwlsi.upc.es/