Robust Bilingual Word Alignment for Machine Aided Translation

Ido Dagan Kenneth W. Church William A. Gale

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

Abstract

We have developed a new program called *word_align* for aligning parallel text, text such as the Canadian Hansards that are available in two or more languages. The program takes the output of *char_align* (Church, 1993), a robust alternative to sentence-based alignment programs, and applies word-level constraints using a version of Brown *et al.*'s Model 2 (Brown et al., 1993), modified and extended to deal with robustness issues. *Word_align* was tested on a subset of Canadian Hansards supplied by Simard (Simard et al., 1992). The combination of *word_align* plus *char_align* reduces the variance (average square error) by a factor of 5 over *char_align* alone. More importantly, because *word_align* and *char_align* were designed to work robustly on texts that are smaller and more noisy than the Hansards, it has been possible to successfully deploy the programs at AT&T Language Line Services, a commercial translation service, to help them with difficult terminology.

1 Introduction

Aligning parallel texts has recently received considerable attention (Warwick et al., 1990; Brown et al., 1991a; Gale and Church, 1991b; Gale and Church, 1991a; Kay and Rosenschein, 1993; Simard et al., 1992; Church, 1993; Kupiec, 1993; Matsumoto et al., 1993). These methods have been used in machine translation (Brown et al., 1990; Sadler, 1989), terminology research and translation aids (Isabelle, 1992; Ogden and Gonzales, 1993), bilingual lexicography (Klavans and Tzoukermann, 1990), collocation studies (Smadja, 1992), word-sense disambiguation (Brown et al., 1991b; Gale et al., 1992) and information retrieval in a multilingual environment (Landauer and Littman, 1990).

The information retrieval application may be of particular relevance to this audience. It would be highly desirable for users to be able to express queries in whatever language they chose and retrieve documents that may or may not have been written in the same language as the query. Landauer and Littman used SVD analysis (or Latent Semantic Indexing) on the Canadian Hansards, parliamentary debates that are published in both English and French, in order to estimate a kind of soft thesaurus. They then showed that these estimates could be used to retrieve documents appropriately in the bilingual condition where the query and the document were written in different languages.

We have been most interested in the terminology application. How does Microsoft, or some other software vendor, want "dialog box," "text box," and "menu box" to be translated in their manuals? Considerable time is spent on terminology questions, many of which have already been solved by other translators working on similar texts. It ought to be possible for a translator to point at an instance of "dialog box" in the English version of the Microsoft Windows manual and see how it was translated in the French version of the same manual. Alternatively, the translator can ask for a bilingual concordance as shown in Figure 1. A PC-based terminology reuse tool is being developed to do just exactly this. The tool depends crucially on the results of an alignment program to determine which parts of the source text correspond with which parts of the target text.

In working with the translators at AT&T Language Line Services, a commercial translation service, we discovered that we needed to completely redesign our alignment programs in order to deal more effectively with texts supplied by Language Line's customers. All too often the texts are not available in electronic form, and may need to be scanned in and processed by an OCR (optical character recognition) device. Even if the texts are available in electronic form, it may not be worth the effort to clean them up by hand. Real texts are not like the Hansards; real texts are much smaller and not nearly as clean as the ideal texts that have

```
displayed . In   the Save        As           dialog   box     , this  area is  called Save
  afficha   Dans     Enregistrer Enregistrer  dialogue boite     cette zone est        Enregistr
  ainsi que son extension . Dans la boite de  dialogue Enregistrer sous , cette zone est appele

x When    you chcose a command  button , the  dialog   box     closes and the command  is
  Lorsque              commande bouton        dialogue boite   ferme          commande execute
  sissez un bouton de commande , la boite de  dialogue se ferme et le programme execute la com
 ...
 ...
  button . Or doubl   ¯ lick the Control   -  menu box  . Or press ESC . If a dialog   box     d
r bouton      cliquer   fois    Systeme       menu case                  Si   dialogue boite   p
ouvez aussi cliquer deux fois sur la case du  menu Systeme . Il est egalement possible d ' a
 ...
 ...
¯ ee ' aa , ' When    you move   to an empty  text  box  , an iiisertion point     ( flastung ve
              Lorsque     placez              texte zone                 insertion (
de Lorsque vous vous placez dans une zone de  texte vide , un point d ' insertion ( barre vertic
```

Figure 1: A small sample of a bilingual concordance, based on the output of *word_align*. Four concordances for the word "box" are shown, identifying three different translations for the word: *boite, case, zone*. The concordances are selected from English and French versions of the Microsoft Windows manual (with some errors introduced by OCR). There are three lines of text for each instance of "box": (1) English, (2) glosses, and (3) French. The glosses are selected from the French text (the third line), and are written underneath the corresponding English words, as identified by *word_align*.

been used in previous studies.

To deal with these robustness issues, Church (1993) developed a character-based alignment method called *char_align*. The method was intended as a replacement for sentence-based methods (e.g., (Brown et al., 1991a; Gale and Church, 1991b; Kay and Rosenschein, 1993)), which are very sensitive to noise. This paper describes a new program, called *word_align*, that starts with an initial "rough" alignment (e.g., the output of *char_align* or a sentence-based alignment method), and produces improved alignments by exploiting constraints at the word-level. The alignment algorithm consists of two steps: (1) estimate translation probabilities, and (2) use these probabilities to search for most probable alignment path. The two steps are described in the following section.

2 The alignment Algorithm

2.1 Estimation of translation probabilities

The translation probabilities are estimated using a method based on Brown *et al.*'s Model 2 (1993), which is summarized in the following subsection, 2.1.1. Then, in subsection 2.1.2, we describe modifications that achieve three goals: (1) enable *word_align* to accept input which may not be aligned by sentence (e.g. *char_align*'s output), (2) reduce the number of parameters that need to be estimated, and (3) prepare the ground for the second step, the search for the best alignment (described in section 2.2).

2.1.1 Brown *et al.*'s Model

In the context of their statistical machine translation project (Brown et al., 1990), Brown *et al.* estimate $\Pr(\mathbf{f}|\mathbf{e})$, the probability that $\mathbf{f}$, a sentence in one language (say French), is the translation of $\mathbf{e}$, a sentence in the other language (say English). $\Pr(\mathbf{f}|\mathbf{e})$ is computed using the concept of *alignment*, denoted by $\mathbf{a}$, which is a set of *connections* between each French word in $\mathbf{f}$ and the corresponding English word in $\mathbf{e}$. A connection, which we will write as $con_{j,i}^{\mathbf{f},\mathbf{e}}$, specifies that position j in $\mathbf{f}$ is connected to position i in $\mathbf{e}$. If a French word in $\mathbf{f}$ does not correspond to any English word in $\mathbf{e}$, then it is connected to the special word *null* (position 0 in $\mathbf{e}$). Notice that this model is directional, as each French position is connected to exactly one position in the English sentence (which might be the *null* word), and accordingly the number of connections in an alignment is equal to the length of the French sentence. However, an English word may be connected to several words in the French sentence, or not connected at all.

Using alignments, the translation probability for a pair of sentences is expressed as

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}\in\mathcal{A}} \Pr(\mathbf{f},\mathbf{a}|\mathbf{e}) \tag{1}$$

where $\mathcal{A}$ is the set of all combinatorially possible alignments for the sentences $\mathbf{f}$ and $\mathbf{e}$ (calligraphic font will be used to denote sets).

In their paper, Brown *et al.* present a series of 5 models of $\Pr(\mathbf{f}|\mathbf{e})$. The first two of these 5 models are summarized here.

Model 1

Model 1 assumes that $\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$ depends primarily on $t(f|e)$, the probability that an occurrence of the English word e is translated as the French word f. That is,

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a} \in \mathcal{A}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \sum_{\mathbf{a} \in \mathcal{A}} C_{\mathbf{f},\mathbf{e}} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2}$$

where $C_{\mathbf{f},\mathbf{e}}$, an irrelevant constant, accounts for certain dependencies on sentence lengths, which are not important for our purposes here. Except for $C_{\mathbf{f},\mathbf{e}}$, most of the notation is borrowed from Brown *et al.*. The variable, j, is used to refer to a position in a French sentence, and the variable, i, is used to refer to a position in an English sentence. The expression, f_j, is used to refer to the French word in position j of a French sentence, and e_i is used to refer to the English word in position i of an English sentence. An alignment, $\mathbf{a}$, is a set of pairs (j, i), each of which connects a position in a French sentence with a corresponding position in an English sentence. The expression, a_j, is used to refer to the English position that is connected to the French position j, and the expression, e_{a_j}, is used to refer to the English word in position a_j. The variable, m, is used to denote the length of the French sentence and the variable, l, is used to denote the length of the English sentence.

There are quite a number of constraints that could be used to estimate $\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$. Model 1 depends primarily on the translation probabilities, $t(f|e)$, and does not make use of constraints involving the positions within an alignment. These constraints will be exploited in Model 2.

Brown *et al.* estimate $t(f|e)$ on the basis of a training set, a set of English and French sentences that have been aligned at the sentence level. Those values of $t(f|e)$ that maximize the probability of the training set are called the maximum likelihood estimates. Brown *et al.* show that the maximum likelihood estimates satisfy

$$t(f|e) = \frac{\sum_{con_{j,i}^{\mathbf{f},\mathbf{e}} \in \mathcal{CON}_{f,e}} \Pr(con_{j,i}^{\mathbf{f},\mathbf{e}})}{\sum_{con_{j,i}^{\mathbf{f},\mathbf{e}} \in \mathcal{CON}_{\cdot,e}} \Pr(con_{j,i}^{\mathbf{f},\mathbf{e}})} \tag{3}$$

where $\mathcal{CON}_{f,e}$ and $\mathcal{CON}_{\cdot,e}$ denote sets of connections: the set $\mathcal{CON}_{f,e}$ contains all connections in the training data between f and e, and the set $\mathcal{CON}_{\cdot,e}$ contains all connections between some French word and e. The probability of a connection, $con_{j,i}^{\mathbf{f},\mathbf{e}}$, is the sum of the probabilities of all alignments that contain it. Notice that equation 3 satisfies the constraint: $\sum_f t(f|e) = 1$, for each English word e.

It follows from the definition of Model 1 that the probability of a connection satisfies:

$$\Pr(con_{j,i}^{\mathbf{f},\mathbf{e}}) = \frac{t(f_j|e_i)}{\sum_{k=0}^{l} t(f_j|e_k)} \tag{4}$$

Recall that f_j refers to the French word in position j of the French sentence $\mathbf{f}$ of length m, and that e_i refers to the English word in position i of the English sentence $\mathbf{e}$ of length l. Also, remember that position 0 is reserved for the *null* word.

Equations 3 and 4 are used iteratively to estimate $t(f|e)$. That is, we start with an initial guess for $t(f|e)$. We then evaluation the right hand side of equation 4, and compute the probability of the connections in the training set. Then we evaluate equation 3, obtain new estimates for the translation probabilities, and repeat the process, until it converges. This iterative process is known as the EM algorithm and has been shown to converge to a stationary point (Baum, 1972; Dempster et al., 1977). Moreover, Brown *et al.* show that Model 1 has a unique maximum, and therefore, in this special case, the EM algorithm is guaranteed to converge to the maximum likelihood solution, and does not depend on the initial guess.

Model 2

Model 2 improves upon model 1 by making use of the positions within an alignment. For instance, it is much more likely that the first word of an English sentence will be connected to a word near the beginning of the corresponding French sentence, than to some word near the end of the French sentence. Model 2 enhances Model 1 with the assumption that the probability of a connection, $con_{j,i}^{\mathbf{f},\mathbf{e}}$, depends also on j and i (the positions in $\mathbf{f}$ and $\mathbf{e}$), as well as on m and l (the lengths of the two sentences). This dependence is expressed through the term $a(i|j, m, l)$, which denotes the probability of connecting position j in a French sentence of length m with position i in an English sentence of length l. Since each French position is connected to exactly one English position, the constraint $\sum_{i=0}^{l} a(i|j, m, l) = 1$ should hold for all j, m and l. In place of equation 2, we now have:

$$\begin{aligned}\Pr(\mathbf{f}|\mathbf{e}) &= \sum_{\mathbf{a} \in \mathcal{A}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \qquad (5) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} C'_{\mathbf{f},\mathbf{e}} \prod_{j=1}^{m} t(f_j|e_{a_j}) \cdot a(a_j|j, m, l).\end{aligned}$$

where $C'_{\mathbf{f},\mathbf{e}}$ is an irrelevant constant.

As in Model 1, equation 3 holds for the maximum likelihood estimates of the translation probabilities. The corresponding equation for the maximum likelihood estimates of $a(i|j,m,l)$ is:

$$a(i|j,m,l) = \frac{\sum_{con^{\mathbf{f},\mathbf{e}}_{j,i} \in \mathcal{CON}^{m,l}_{j,i}} \Pr(con^{\mathbf{f},\mathbf{e}}_{j,i})}{\sum_{con^{\mathbf{f},\mathbf{e}}_{j,i} \in \mathcal{CON}^{m,l}_{j,\cdot}} \Pr(con^{\mathbf{f},\mathbf{e}}_{j,i})} \quad (6)$$

where $\mathcal{CON}^{m,l}_{j,i}$ denotes the set of connections in the training data between positions j and i in French and English sentences of lengths m and l, respectively. Similarly, $\mathcal{CON}^{m,l}_{j,\cdot}$ denotes the set of connections between position j and some English position, in sentences of these lengths.

Instead of equation 4, we obtain the following equation for the probability of a connection:

$$\Pr(con^{\mathbf{f},\mathbf{e}}_{j,i}) = \frac{t(f_j|e_i) \cdot a(i|j,m,l)}{\sum_{k=0}^{l} t(f_j|e_k) \cdot a(k|j,m,l)}. \quad (7)$$

Notice that Model 1 is a special case of Model 2, where $a(i|j,m,l)$ is held fixed at $\frac{1}{l+1}$.

As before, the EM algorithm is used to compute maximum likelihood estimates for $t(f|e)$ and $a(i|j,m,l)$ (using first equation 7, and then equations 3 and 6). However, in this case, Model 2 does not have a unique maximum, and therefore the results depend on the initial guesses. Brown *et al.* therefore use Model 1 to obtain estimates for $t(f|e)$ which do not depend on the initial guesses. These values are then used as the initial guesses of $t(f|e)$ in Model 2.

2.1.2 Our model

As mentioned in the introduction, we are interested in aligning corpora that are smaller and noisier than the Hansards. This implies severe practical constraints on the word alignment algorithm. As mentioned earlier, we chose to start with the output of *char_align* because it is more robust than alternative sentence-based methods. This choice, of course, requires certain modifications to the model of Brown *et al.* to accommodate as input an initial rough alignment (such as produced by *char_align*) instead of pairs of aligned sentences. It is also useful to reduce the number of parameters that we are trying to estimate, because we have much less data and much more noise. The paragraphs below describe our modifications which are intended to meet these somewhat different requirements. The two major modifications are: (a) replacing the sentence-by-sentence alignment with a single global alignment for the entire corpus, and (b) replacing the set of probabilities $a(i|j,m,l)$ with a small set of *offset probabilities*.

Word_align starts with an initial rough alignment, $\mathbf{I}$, which maps French positions to English positions (if the mapping is partial, we use linear extrapolation to make it complete). Our goal is to find a global alignment, $\mathbf{A}$, which is more accurate than $\mathbf{I}$. To achieve this goal, we first use $\mathbf{I}$ to determine which connections will be considered for $\mathbf{A}$. Let $con_{j,i}$ denote a connection between position j in the French corpus and position i in the English corpus (the super-scripts in $con^{\mathbf{f},\mathbf{e}}_{j,i}$ are omitted, as there is no notion of sentences). We assume that $con_{j,i}$ is a possible connection only if i falls within a limited window which is centered around $\mathbf{I}(j)$, such that:

$$\mathbf{I}(j) - w \leq i \leq \mathbf{I}(j) + w \quad (8)$$

where w is a predetermined parameter specifying the size of the window (we typically set w to 20 words). Connections that fall outside this window are assumed to have a zero probability. This assumption replaces the assumption of Brown *et al.* that connections which cross boundaries of aligned sentences have a zero probability. In this new framework, equation 3 becomes:

$$t(f|e) = \frac{\sum_{con_{j,i} \in \mathcal{CON}_{f,e}} \Pr(con_{j,i})}{\sum_{con_{j,i} \in \mathcal{CON}_{\cdot,e}} \Pr(con_{j,i})} \quad (9)$$

where $\mathcal{CON}_{f,e}$ and $\mathcal{CON}_{\cdot,e}$ are taken from the set of possible connections, as defined by (8).

Turning to Model 2, the parameters of the form $a(i|j,m,l)$ are somewhat more problematic. First, since there are no sentence boundaries, there are no direct equivalents for i, j, m and l. Secondly, there are too many parameters to be estimated, given the limited size of our corpora (one parameter for each combination of i, j, m and l). Fortunately, these parameters are highly redundant. For example, it is likely that $a(i|j,m,l)$ will be very close to $a(i+1|j+1,m,l)$ and $a(i|j,m+1,l+1)$.

In order to deal with these concerns, we replace probabilities of the form $a(i|j,m,l)$ with a small set of *offset probabilities*. We use k to denote the offset between i, an English position which corresponds to the French position j, and the English position which the input alignment $\mathbf{I}$ connects to j: $k = i - \mathbf{I}(j)$. An offset probability, $o(k)$, is the probability of having an offset k for some arbitrary connection. According to (8), k ranges between $-w$ and w. Thus, instead of equation 6, we have

$$o(k) = \frac{\sum_{con_{j,i} \in \mathcal{CON}_k} \Pr(con_{j,i})}{\sum_{con_{j,i} \in \mathcal{CON}} \Pr(con_{j,i})} \quad (10)$$

where $\mathcal{CON}$ is the set of all connections and $\mathcal{CON}_k$ is the set of all connections with offset k. Instead of equation 7, we have

$$\Pr(con_{j,i}) = \frac{t(f_j|e_i) \cdot o(i - \mathbf{I}(j))}{\sum_{h=\mathbf{I}(j)-w}^{\mathbf{I}(j)+w} t(f_j|e_h) \cdot o(h - \mathbf{I}(j))}. \quad (11)$$

The last three equations are used in the EM algorithm in an iterative fashion as before to estimate the translation probabilities and the offset probabilities. Table 1 and Figure 2 show some values that were estimated in this way. The input consisted of a pair of Microsoft Windows manuals in English (125,000 words) and its equivalent in French (143,000 words). Table 1 shows four French words and the four most likely translations, sorted by $t(e|f)$[1]. Note that the correct translation(s) are usually near the front of the list, though there is a tendency for the program to be confused by collocates such as "information about". Figure 2 shows the probability estimates for offsets from the initial alignment **I**. Note that smaller offsets are more likely than larger ones, as we would expect. Moreover, the distribution is reasonably close to normal, as indicated by the dotted line, which was generated by a Gaussian with a mean of 0 and standard deviation of 10^2.

We have found it useful to make use of three filters to deal with robustness issues. Empirically, we found that both high frequency and low frequency words caused difficulties and therefore connections involving these words are filtered out. The thresholds are set to exclude the most frequent function words and punctuations, as well as words with less than 3 occurrences. In addition, following a similar filter by Brown *et al.*, small values of $t(f|e)$ are set to 0 after each iteration of the EM algorithm because these small values often correspond to inappropriate translations. Finally, connections to *null* are ignored. Such connections model French words that are often omitted in the English translation. However, because of OCR errors and other sources of noise, it was decided that this phenomenon was too difficult to model.

Some words will not be aligned because of these heuristics. It may not be necessary, however, to align all words in order to meet the goal of helping translators (and lexicographers) with difficult terminology.

2.2 Finding the most probable alignment

The EM algorithm produces two sets of maximum likelihood probability estimates: translation probabilities, $t(f|e)$, and offset probabilities, $o(k)$. Brown *et al.* select their preferred alignment simply by choosing the most probable alignment according to the maximum likelihood probabilities, relative to the given sentence alignment. In the terms of our model, it is necessary to select the alignment **A** that maximizes:

$$\prod_{con_{j,i}\in \mathbf{A}} t(f_j|e_i)\cdot o(i-\mathbf{I}(j)). \qquad (12)$$

Unfortunately, this method does not model the dependence between connections for French words that are near one another. For example, the fact that the French position j was connected to the English position i will not increase the probability that $j+1$ will be connected to an English position near i. The absence of such dependence can easily confuse the program, mainly in aligning adjacent occurrences of the same word, which are common in technical texts. Brown *et al.* introduce such dependence in their Model 4. We have selected a simpler alternative defined in terms of offset probabilities.

2.2.1 Determining the set of relevant connections

The first step in finding the most probable alignment is to determine the *relevant connections* for each French position. Relevant connections are required to be reasonably likely, that is, their translation probability ($t(f|e)$) should exceed some minimal threshold. Moreover, they are required to fall within a window between $\mathbf{I}(j)-w$ and $\mathbf{I}(j)+w$ in the English corpus, as in the previous step (parameter estimation). We call a French position *relevant* if it has at least one relevant connection. Each alignment **A** then consists of exactly one connection for each relevant French position (the irrelevant positions are ignored).

2.2.2 Determining the most probable alignment

To model the dependency between connections in an alignment, we assume that the offset of a connection is determined relative to the preceding connection in **A**, instead of relative to the initial alignment, **I**. For this purpose, we define $\mathbf{A}'(j)$ as a linear extrapolation from the preceding connection in **A**:

$$\mathbf{A}'(j)=\mathbf{A}(j_{prev})+(j-j_{prev})\frac{N_E}{N_F} \qquad (13)$$

where j_{prev} is the last French position before j which is aligned by **A** and N_E and N_F are the lengths of the English and French corpora. $\mathbf{A}'(j)$ thus predicts the connection of j, knowing the connection of j_{prev} and assuming that the two languages have the same word order. Instead of (12), the most probable alignment maximizes

$$\prod_{con_{j,i}\in \mathbf{A}} t(f_j|e_i)\cdot o(i-\mathbf{A}'(j)). \qquad (14)$$

[1] In this example, French is used as the source language and English as the target.

[2] The center of the estimated distribution seems more flat than in a normal distribution. This might be explained by a higher tendency for local changes of word order within phrases than for order changes among phrases. This is merely a hypothesis, though, which requires further testing.

French word	English translations (with probabilities)
zone	box (0.58) area (0.28) want (0.04) In (0.02)
fermer	close (0.44) when (0.08) Close (0.07) selected (0.06)
informations	information (0.66) about (0.15) For (0.12) see (0.04)
insertion	insertion (0.61) point (0.23) Edit (0.06) To (0.05)

Table 1: Estimated translation probabilities

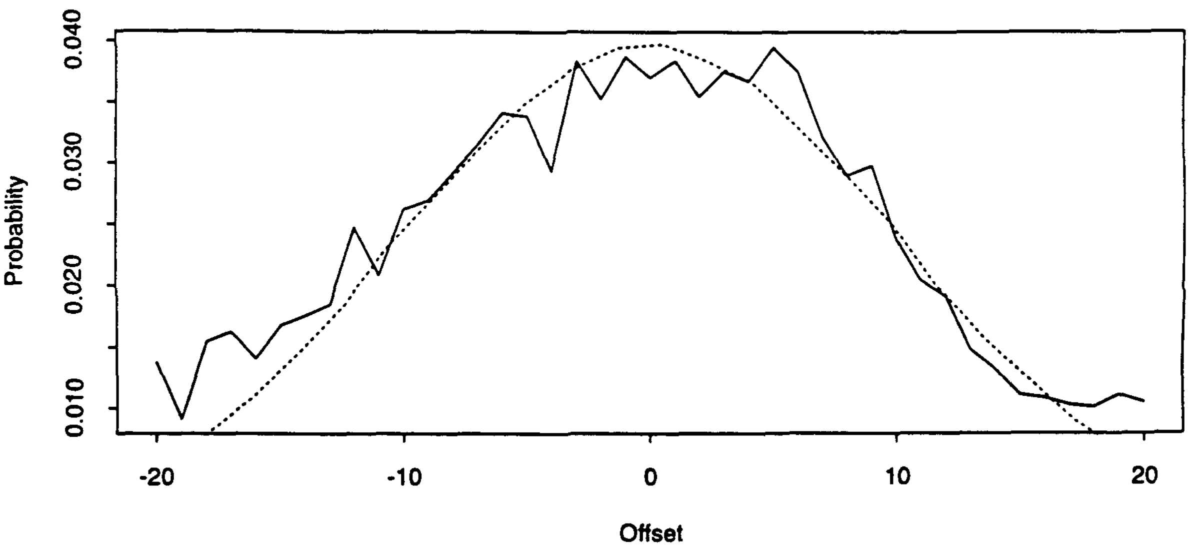

Figure 2: Estimated offset probabilities (solid line) along with a Gaussian (dashed line) for comparison.

We approximate the offset probabilities, $o(k)$, relative to $\mathbf{A}'$, using the maximum likelihood estimates which were computed relative to $\mathbf{I}$ (as described in Section 2.1.2).

We use a dynamic programming algorithm to find the most probable alignment. This enables us to know the value $\mathbf{A}(j_{prev})$ when dealing with position j. To avoid connections with very low probability (due to a large offset) we require that $t(f_j|e_i)\cdot o(i-\mathbf{A}'(j))$ exceeds a pre-specified threshold T[3]. If the threshold is not exceeded, the connection is dropped from the alignment, and $t(f_j|e_i)\cdot o(i-\mathbf{A}'(j))$ for that connection is set to T when computing (14). T can therefore be interpreted as a global setting of the probability that a random position will be connected to the *null* English word[4]. A similar dynamic programming approach was used by Gale and Church for word alignment (Gale and Church, 1991a), to handle dependency between connections.

3 Evaluation

Word_align was first evaluated on a representative sample of Canadian Hansards (160,000 words in English and French). The sample was kindly provided by Simard *et al.*, along with alignments of sentence boundaries as determined by their panel of 8 judges (Simard et al., 1992).

Ten iterations of the EM algorithm were computed to estimate the parameters of the model. The window size was set to 20 words in each direction, and the minimal threshold for $t(f|e)$ was set to 0.005. We considered connections whose source and target words had frequencies between 3 and 1700 (1700 is the highest frequency of a content word in the corpus. We thus excluded as many

[3] In fact, the threshold on $t(f_j|e_i)$, which is used to determine the relevant connections (described in the previous subsection), is used just as an efficient early application of the threshold T. This early application is possible when $t(f_j|e_i)\cdot o(k_{max}) < T$, where k_{max} is the value of k with maximal $o(k)$.

[4] As mentioned earlier, we do not estimate directly translation probabilities for the *null* English word.

function words as possible, but no content words). In this experiment, we used French as the source language and English as the target language.

Figure 3 presents the alignment error rate of *word_align*. It is compared with the error rate of *word_align*'s input, i.e. the initial rough alignment which is produced by *char_align*. The errors are sampled at sentence boundaries, and are measured as the relative distance between the output of the alignment program and the "true" alignment, as defined by the human judges[5]. The histograms present errors in the range of -20–20, which covers about 95% of the data[6]. It can be seen that *word_align* decreases the error rate significantly (notice the different scales of the vertical axes). In 55% of the cases, there is no error in *word_align*'s output (distance of 0), in 73% the distance from the correct alignment is at most 1, and in 84% the distance is at most 3.

A second evaluation of *word_align* was performed on noisy technical documents, of the type typically available for AT&T Language Line Services. We used the English and French versions of a manual of monitoring equipment (about 65,000 words), both scanned by an OCR device. We sampled the English vocabulary with frequency between three and 450 occurrences, the same vocabulary that was used for alignment. We sampled 100 types from the top fifth by frequency of the vocabulary (quintile), 80 types from the second quintile, 60 from the third, 40 from the fourth, and 20 from the bottom quintile. We used this stratified sampling because we wanted to make more accurate statements about our error rate *by tokens* than we would have obtained from random sampling, or even from equal weighting of the quintiles. After choosing the 300 types from the vocabulary list, one token for each type was chosen at random from the corpus. By hand, the best corresponding position in the French version was chosen, to be compared with *word_align*'s output.

Table 2 summarizes the results of the second experiment. The figures indicate the expected relative frequency of each offset from the correct alignment. This relative frequency was computed according to the word frequencies in the stratified sample. As shown in the table, for 60.5% of the tokens the alignment is accurate, and in 84% the offset from the correct alingment is at most 3. These figures demonstrate the usefulness of *word_align* for constructing bilingual lexicons, and its impact on the quality of bilingual concordances (as in Figure 1). Indeed, using bilingual concordances which are based on *word_align*'s output, the translators at AT&T Language Line Services are now producing bilingual terminology lexicons at a rate of 60-100 terms per hour! This is compared with the previous rate of about 30 terms per hour using *char_align*'s output, and an extremely lower rate before alignment tools were available.

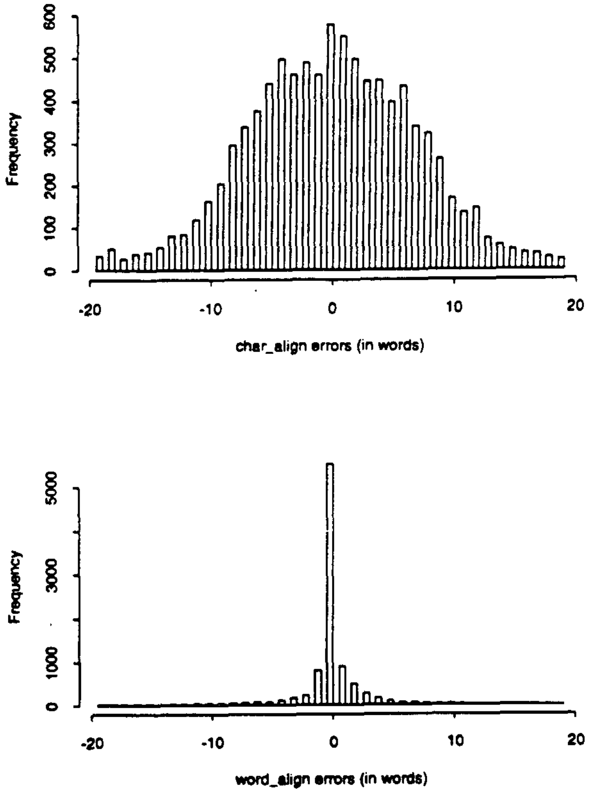

Figure 3: *Word_align* reduces the variance (average square error) by a factor of 5 over *char_align* alone (notice the vertical scales).

4 Conclusions

Compared with other word alignment algorithms (Brown et al., 1993; Gale and Church, 1991a), *word_align* does not require sentence alignment as input, and was shown to produce useful alignments for small and noisy corpora. Its robustness was achieved by modifying Brown *et al.*'s Model 2 to handle an initial "rough" alignment, reducing the number of parameters and introducing a dependency between alignments of adjacent words. Taking the output of *char_align* as input, *word_align* produces significantly better, word-

[5] As explained earlier, *word_align* produces a partial alignment. For the purpose of the evaluation, we used linear interpolation to get alignments for all the positions in the sample.

[6] Recall that the window size we used is 20 words in each direction, which means that *word_align* cannot recover from larger errors in *char_align*.

Offset from correct alignment	Percentage	Accumulative percentage
0	60.5%	60.5%
1	10.8%	71.3%
2	7.5%	78.8%
3	5.2%	84%
4	1.6%	85.6%

Table 2: *Word_align*'s precision on noisy input, scanned by an OCR device.

level, alignments on the kind of corpora that are typically available to translators. This improvement increased the rate of constructing bilingual terminology lexicons at AT&T Language Line Services by a factor of 2-3. In addition, the alignments may also be helpful to developers of lexicons for machine translation systems. *Word_align* thus provides an example how a model such as Brown *et al.*'s Model 2, that was originally designed for research in statistical machine translation, can be modified to achieve practical, though less ambitious, goals in the near term.

REFERENCES

L. E. Baum. 1972. An inequality and an associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3:1-8.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R.L. Mercer, and Roossin P.S. 1990. A statistical approach to language translation. *Computational Linguistics*, 16(2):79–85.

P. Brown, J. Lai, and R. Mercer. 1991a. Aligning sentences in parallel corpora. In *Proc. of the Annual Meeting of the ACL.*

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1991b. Word sense disambiguation using statistical methods. In *Proc. of the Annual Meeting of the ACL.*

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of machine translation: parameter estimation. *Computational Linguistics.* to appear.

Kenneth W. Church. 1993. Char_align: A program for aligning parallel texts at character level. In *Proc. of the Annual Meeting of the ACL.*

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum liklihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38.

William Gale and Kenneth Church. 1991a. Identifying word correspondence in parallel text. In *Proc. of the DARPA Workshop on Speech and Natural Language.*

William Gale and Kenneth Church. 1991b. A program for aligning sentences in bilingual corpora. In *Proc. of the Annual Meeting of the ACL.*

William Gale, Kenneth Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proc. of the International Conference on Theoretical and Methodolgical Issues in Machine Translation.*

P. Isabelle. 1992. Bi-textual aids for translators. In *Proc. of the Annual Conference of the UW Center for the New OED and Text Research.*

M. Kay and M. Rosenschein. 1993. Text-translation alignment. *Computational Linguistics.* to appear.

J. Klavans and E. Tzoukermann. 1990. The bicord system. In *Proc. of COLING.*

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proc. of the Annual Meeting of the ACL.*

Thomas K. Landauer and Michael L. Littman. 1990. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proc. of the Annual Conference of the UW Center for the New OED and Text Research.*

Yuji Matsumoto, Hiroyuki Ishimoto, Takehito Utsuro, and Makoto Nagao. 1993. Structural matching of parallel texts. In *Proc. of the Annual Meeting of the ACL.*

William Ogden and Margarita Gonzales. 1993. Norm – a system for translators. Demonstration at ARPA Workshop on Human Language Technology.

V. Sadler. 1989. *Working with analogical semantics: Disambiguation techniques in DLT.* Foris Publications.

M. Simard, G. Foster, and P. Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proc. of the International Conference on Theoretical and Methodolgical Issues in Machine Translation.*

Frank Smadja. 1992. How to compile a bilingual collocational lexicon automatically. In *AAAI Workshop on Statistically-based Natural Language Processing Techniques*, July.

S. Warwick, J. Hajic, and G. Russell. 1990. Searching on tagged corpora: linguistically motivated concordance analysis. In *Proc. of the Annual Conference of the UW Center for the New OED and Text Research.*