Using Cognates to Align Sentences in Bilingual Corpora

Michel Simard
George F. Foster
Pierre Isabelle[1]

Canadian Workplace Automation Research Centre
1575 Chomedey Blvd.
Laval, Québec
CANADA H7V 2X2

Abstract

In a recent paper, Gale and Church describe an inexpensive method for aligning bitext, based exclusively on sentence lengths [Gale and Church, 1991]. While this method produces surprisingly good results (a success rate around 96%), even better results are required to perform such tasks as the computer-assisted revision of translations. In this paper, we examine some of the weaknesses of Gale and Church's program, and explain how just a small amount of linguistic knowledge would help to overcome these weaknesses. We discuss how *cognates* provide for a cheap and reasonably reliable source of linguistic knowledge. To illustrate this, we describe a modification to the program in which the criterion is cognates rather than sentence lengths. Finally, we show how better and more efficient results may be obtained by combining the two criteria — length and "cognateness". Our method can be generalized to accommodate other sources of linguistic knowledge, and experimentation shows that it produces better results than alignments based on length alone, at a minimal cost.

Introduction

Recent years have seen a surge of interest in bilingual and multilingual corpora, i.e. corpora composed of a source text along with translations of that text in different languages. One very useful organization of bilingual corpora, that we will call *bitext* (or *multitext*) [Harris, 1988], requires that the different versions of the same text be *aligned*: Given a text and its translation, an alignment is a segmentation of the two texts such that the *n*th segment of one text is the translation of the *n*th segment of the other (as a special case, empty segments are allowed, and either correspond to translator's omissions or additions). We call "couples" such pairs of segments that

1. email: simard@ccrit.doc.ca, foster@ccrit.doc.ca, isabelle@ccrit.doc.ca

are mutual translations. The appearance of an alignment depends on its resolution, i.e. on the nature of the units on which the segmentation is done. For example, an alignment that simply puts paragraphs in correspondence would be considered a “gross” alignment, compared to one that shows word correspondences. In any case, given its resolution, a correct alignment should be “maximal”, i.e. it should be composed of the smallest possible couples. The type of alignment we will be discussing takes the sentence to be the segmentation unit.

Figure 1 illustrates such an alignment.

<table>
<tr><td>couple 1</td><td>The crisis our farmers are in right now will affect all of us at a certain point in time.</td><td>La crise que vivent en ce moment nos agriculteurs se répercutera sur tous et chacun de nous à un certain moment.</td></tr>
<tr><td rowspan="2">couple 2</td><td rowspan="2">We are all consumers and we all need a strong and healthy agricultural sector.</td><td>Nous sommes des consommateurs.</td></tr>
<tr><td>Nous avons tous besoin d’une agriculture saine et forte.</td></tr>
<tr><td rowspan="2">couple 3</td><td>I am glad that the Hon. Member for Algoma (Mr. Foster) mentioned figures in his remarks.</td><td rowspan="2">Heureusement que le député d’Algoma (M. Foster) a mentionné des chiffres dans ses remarques, sans cela ce gouvernement s’en serait sorti en douce encore une fois.</td></tr>
<tr><td>Otherwise , the Government might have eluded the problem once again .</td></tr>
<tr><td>couple 4</td><td>The Hon. Member for Algoma suggested Tuesday night that the Government had to take a clear position and make a commitment to assist our farmers before it is too late.</td><td>Le député d’Algoma suggérait mardi soir qu’il fallait que le gouvernement se prononce clairement et s’engage à aider nos agriculteurs avant qu’il ne soit trop tard .</td></tr>
</table>

Figure 1: An alignment between a pair of English and French paragraphs.

Clearly, a corpus of properly aligned bitext constitutes an extremely valuable source of information, not only to researchers in bilingual lexicography and terminology, but also for a range of applications. While producing alignments by hand is extremely time-consuming and requires the skills of individuals with a good knowledge of both languages, there exist programs that produce relatively reliable alignments at a minimal cost [Brown et al. 1991, Gale and Church 1991]. And in fact, for some applications, it is sufficient that the alignment for a given bitext be only partially correct, as long as there is a way of automatically extracting a subset of that bitext for the alignment of which there is a high level of confidence.

For other applications however, much can be gained from a program that is capable of producing high-quality alignments for an entire piece of bitext. This is the case for translation revision and evaluation [Isabelle, 1991]. The first of these gains is obvious: to allow one to visualize a text and its translation side-by-side, with explicit connections between individual components.

An alignment may also constitute the basis of deeper automatic analyses of translations. For example, it could be used to flag possible omissions in a translation, or to signal common translation mistakes, such as terminological inconsistencies and the use of *faux amis*.

Yet another possibility is to have an alignment process at work while a translation is being done. In addition to the error detection mechanisms mentioned above, such a process could provide translation ‘suggestions’ when the same piece of text appears more than once in the source text.

It is clear that, in order to be a useful basis for a translation tool, an alignment process must ultimately have access to some language-specific knowledge; what we have done represents a preliminary step in that direction.

1 Length-based Alignment Program

Following an idea which first appeared in [Brown et al. 1990], Gale and Church suggest a method for aligning pairs of texts. It relies on two hypotheses: a) the lengths (in number of characters) of segments which are translations of one another are highly correlated; and b) all translations are done using one of six "translation patterns": (1) one sentence translates into one, (2) two consecutive sentences translate into one, (3) one sentence translates into two, (4) two sentences translate into two[2], (5) a sentence is not translated at all or (6) a new sentence that has no equivalent in the source text is introduced by the translator.

Assuming that paragraphs are already aligned (i.e. the *n*th paragraph of the first text is the translation of the *n*th paragraph of the second text), the program works as follows. For each pair of aligned paragraphs, consider all possible couples constructed using one of the translation patterns. Assign each couple a score, intended to reflect how well the two segments relate to one another. Based on these scores, and using a dynamic programming scheme, determine the best sequence of couples leading to a valid alignment.

Gale and Church's scoring function is based on a probabilistic model. It produces an approximation of the probability that two segments are mutual translations, given the lengths of the two segments and the likelihood of the translation pattern that connects them.

The success rate of this method is surprisingly high: the program finds almost 96% of the couples of the correct alignment. The remaining couples — alignment errors — are either pairs of unrelated or partially related segments, or pairs of segments that could have been further segmented.

One possible explanation for this high rate of success is that most of the time the program is actually solving easy problems. Obviously, for two paragraphs containing five sentences each, chances are the correct alignment is the trivial one (five one-to-one alignments), and as expected, this is the alignment the program tends to produce.

But as soon as the problems get a little harder, the program becomes more likely to make mistakes. For example, when two paragraphs contain a different number of sentences, one has to assume that either the translator did not translate all of the source text or, more likely, that he used some contraction or expansion translation pattern. Except in the most straightforward situations (e.g. two short sentences that translate into one long sentence, all other couples in the alignment being highly length-correlated), quite often the program incorrectly introduces some irregular alignment (expansion or contraction), and misaligns everything between that point and the actual troublesome spot. Figure 2 shows an example of such a situation.

2. That is: the first sentence of language A and the first sentence of language B are not mutual translations, nor are the second sentence of language A and the second sentence of language B, but together, the first and second sentences of language A constitute a translation of the first and second sentences of language B.

couple 1	*Notwithstanding the fact that we were going through some rough economic times, clearly spending was absolutely and totally out of control of that previous Government which now sits as the official Opposition.*	*Il est certain que nous traversions une période difficile sur le plan économique mais, néanmoins, les dépenses de ce gouvernement qui nous a précédé, et que nous avons renvoyés sur les bancs de l'opposition officielle, étaient totalement hors de contrôle.*
	Who is left to deal in a frugal way with taxpayers' dollars?	
couple 2	*Who is left to get things under control gradually so that we do not pass that albatross of increasing debt down to future generations?*	*À qui appartient-il maintenant d'être parcimonieux avec l'argent du contribuable?*
couple 3	*We are in the situation now where a tremendous amount of our tax revenues goes just to service the cost of that debt.*	*Qui doit essayer de remettre les choses à leur place pour qu'on ne transmette pas aux autres générations une dette croissante?*
couple 4	*In terms of expenditures, a large chunk of our current tax revenues go just to service the debt.*	*Nous en sommes à un stade où une quantité énorme de recettes fiscales ne sert qu'à financer la dette.*

Figure 2: Erroneous alignment resulting from an incorrect guess as to where the actual contraction occurs.

A striking characteristic of these mistakes is that even a very small amount of linguistic knowledge would help prevent them: e.g. the fact that a question (identified by a terminating question mark) is very likely to translate into another question; or that *taxe* is a likely translation for *tax*.

The intuition that underlies our work is that the notion of 'cognate words' does provide such a source of knowledge for a minimal price.

2 Translation and Cognates

Informally speaking, cognates are pairs of tokens of different languages which share "obvious" phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. The pairs *generation/génération* and *error/erreur* constitute typical examples for English and French. One might want to extend the notion so as to include such things as proper nouns (*Paris*; *London* and *Londres*), numerical expressions and even punctuation (question marks, parentheses, etc.).

We need a way to measure how two pieces of text are related in terms of cognates. Given a pair of text segments from different languages, one can compute their level of "cognateness" in the following way. We first count the numbers n and m of tokens in each segment; then match these tokens so as to obtain the largest possible number c of pairs of cognates, without using the same token twice. The cognateness γ of that pair of segments is defined as

$$\gamma = \frac{c}{(n+m)/2}.$$

This measure is useful, because it is independent of the lengths of the segments involved. A null cognateness ($\gamma = 0$) for a pair of texts means that the two are totally unrelated in terms of cognates. On the other hand, a cognateness equal to 1 denotes a "maximal" relation.

Our fundamental assumption is that translation (seen as a relation) and cognateness are correlated: we expect to find a significantly higher number of cognates between pairs of sentences which are mutual translations than between random pairs of sentences.

To verify this conjecture, we went through the process of hand-aligning a small extract of the Canadian Hansards (13 pairs of paragraphs: 102 English and 94 French sentences), and identifying pairs of cognates between aligned segments. A similar operation was performed on a "random" alignment of the same texts[3]. For each couple, we computed the level of cognateness.

The results are quite convincing: an average cognateness $\bar{\gamma} = 0.21$ for pairs of segments which are mutual translations and $\bar{\gamma} = 0.06$ for random pairs. Cognates would therefore appear to be a reasonable criterion for aligning sentences.

Now, how does an alignment program compute the level of cognateness of a given pair of text segments? It appears that for this task, it is not essential to resort to a list of cognates specific to a particular pair of languages. An automatic cognate-matching mechanism can be devised that relies on an "operational" definition of cognates instead of such a list, and that produces very acceptable results. Consider the following definition:

Given a pair of sentences S_1 and S_2, we identify two lists T_1 and T_2 of tokens t, to be used as candidates in cognate pairs; these are "maximal" substrings of S_1 or S_2 which belong to one of the following categories:

(1) t is entirely composed of letters and digits, but contains at least one digit;

(2) t is exclusively composed of letters, and is at least four letters long.

(3) t is a single punctuation character;

The first category is intended to catch numerical expressions, which in most cases are language-independent and preserved across translations, thereby constituting very interesting candidates. The second category is defined so as to exclude most "functional" words which tend to be short and seldom 'cognated'. Finally, we included the third category on the intuition that the translation process has a tendency to preserve punctuation.

Given two such candidates t_1 and t_2 from token lists T_1 and T_2 respectively,

- if both are members of categories (1) or (3), t_1 and t_2 are cognates iff they are completely identical[4].
- if they are members of category (2), t_1 and t_2 are cognates iff their *four* first characters are identical.

According to this definition, English's *financed* and French's *financier* are cognates, and so are English's and French's *opposition*, but *government* and *gouvernement* are not. On the other hand, numerical expressions and punctuations can only be cognates of themselves.

---

3. To obtain a random alignment, we used a variant of the alignment program which replaces the scoring function with a random function. The idea was to obtain an approximation of the expected number of cognates for arbitrary pairs of segments that the program does consider as candidates for alignment. It should be noted that the probability of two segments being mutual translations in such a random alignment varies with the numbers of sentences per paragraph. If we assume that cognateness and translation are correlated, then the average cognateness of random alignments will also vary with the sizes of paragraphs. The samples we used were relatively long, with the average numbers of sentences per paragraph at 7.8 for English and 7.2 for French. So the results should be taken with caution.
4. In this context, character identity is independent of capitalization and accent marks: as far as we are concerned, characters *é* and *E* are "identical".

Needless to say such a definition makes it very easy to devise a simple program that automatically identifies all pairs of cognates for a given pair of text segments.

Crude as it is, this definition produces results that compare with those obtained with the previous intuition-based selection: $\bar{\gamma} = 0.30$ for mutual translations and 0.09 for random pairs of segments. In both cases, the number of cognates for segments which are mutual translations is at least three times as high as for randomly selected segments. That this definition produces higher figures may be accounted for by the fact that it excludes shorter tokens.

3 Cognate-based Alignments

The easiest way to illustrate how cognates may be used to produce automatic alignments is to modify the standard length-based program so that it uses cognateness instead of segment lengths as its main criterion. This may be done by changing the scoring function.

The statistical analysis of our hand-aligned portion of the Canadian Hansards revealed that the number of pairs of cognates, c, that can be obtained from a pair of aligned segments of average size n (number of candidate tokens per segment) approximately follows a binomial distribution $B(n, p_t)$, where p_t is the probability that an individual token of one segment has a cognate in the other segment when the two segments are mutual translations (notice that this is the same as the expected cognateness $E(\gamma)$). This means that in practice, if two segments of average size n are mutual translations (an event denoted by t), then we can estimate the probability of obtaining c pairs of cognates as:

$$P(c|n, t) \approx \binom{n}{c} \cdot p_t^c \cdot (1 - p_t)^{n-c}$$

The same type of distribution is observed when random couples are examined instead of mutual translations, the only difference being the expected cognateness, which is simply denoted by p in this case.

An interesting way of measuring how well two segments of average size n relate to one another is to compute the probability of the observed number of cognates, c, under the hypothesis that the two segments are mutual translations, and compare it with the probability of that same number of cognates under the hypothesis that the two segments are the result of a random choice:

$$\frac{P(c|n, t)}{P(c|n)}$$

This ratio takes values greater than 1 when the observed cognateness is closer to that of mutual translations than to that of random pairs of segments, and values smaller than 1 in opposite situations.

Our scoring function is based both on this ratio and on the likelihood of the translation pattern a that connects the two segments of the given couple: it is defined as minus the log of their product.

$$Score(a, c, n) = -\log\left[\frac{P(c|n, t)}{P(c|n)} \cdot P(a)\right]$$

$$= -\log\left[\left[\frac{p_t}{p}\right]^c \cdot \left[\frac{1-p_t}{1-p}\right]^{n-c} \cdot P(a)\right]$$

$$= -\left[c \cdot \log\frac{p_t}{p}\right] - \left[(n-c) \cdot \log\frac{1-p_t}{1-p}\right] - \log P(a)$$

The behavior of function *Score* is compatible with the dynamic programming scheme used in the program: it is such that smaller values indicate better alignments. To illustrate this, we can re-express *Score* as a function *Score*' of a, γ and n:

$$Score'(a, \gamma, n) = n \cdot (A\gamma + B) - \log P(a).$$

First, it is clear that for fixed values of n and γ, more likely translation patterns (a) yield smaller values for *Score*'. Second, if $0 \leq p < p_t \leq 1$, then $A < 0$ and $B > 0$, so that higher levels of cognateness γ also produce smaller values of *Score*' when n and a are fixed. Finally, the "size" of the couple n has the effect of determining the relative importance of γ in the computation of *Score*': the degree of cognateness will play a greater role in the scoring function when the segments considered are relatively long. Intuitively, this may be taken to reflect the fact that cognateness, as a measure of how well two segments relate to one another, is not as significant for short pairs of segments as it is for long ones.

While the observed number of cognates per token varies slightly with the size of the segments in the Hansards, we found that our alignment program was fairly insensitive to these small variations, so the overall average values of Section 2: 0.30 and 0.09 were used as estimations for p_t and p respectively. As for $P(a)$, we used the values proposed by Gale and Church (Table 1).

Translation pattern	$P(a)$
1-1	*0.89*
1-0 or 0-1	*0.0099*
1-2 or 2-1	*0.089*
2-2	*0.011*

Table 1: a priori probabilities of translation patterns.
Source: [Gale and Church, 1991].

The program was tested on a fairly large sample of bitext. The manner in which the tests were conducted and the quantitative results are detailed in section 5. For now, let us simply say that, not very surprisingly, the results we obtain with this program are not as impressive as those obtained with a scoring function based on lengths alone. We believe that this is attributable to the large variance in cognateness levels: our scoring function accounts for the fact that it is not at all uncommon to find average size pairs of sentences (say, 10 words each) which are perfect translations of one another, but that do not share a single cognate. On the other hand, it is quite frequent to see unrelated pairs of sentences that share a few cognates, especially if they appear in the same context.

Another observation is that this program is not nearly as efficient as the standard length-based program (on our test corpus, it was 9 times slower): obviously, finding pairs of cognates is much more costly than simply comparing lengths.

What the results do show however is that an approximate measure of the level of cognateness such as the one described above is a valid, albeit weak, criterion for aligning sentences.

4 An Algorithm using Cognates to improve a Length-based Alignment

While cognates alone cannot produce better alignments than length differences, an appealing possibility is to use the cognateness criterion only in situations where the length-based method alone runs into trouble. Gale and Church suggest that in such cases their scoring function is likely to have assigned poor scores, and that this information may be used to locate potential errors. The following observation suggests a more convenient way of sensing trouble:

The length-based scoring function is such that it produces only positive integers, that smaller scores indicate a better fit between pairs of segments, and that the overall score of an alignment is obtained by adding the individual scores of its constituent couples. If for a pair of paragraphs, instead of identifying the alignment with the best overall score, we keep the *two* best alignments, we observe that the overall score of the second best is typically much larger than that of the best: approximately 100 times as large on average. When looking only at paragraphs where the program fails to find the correct alignment, we find that figure to be much closer to 2. This means that in many difficult paragraphs, the program is actually making decisions based on relatively small scoring differences: in a third of all paragraphs where the program produces an incorrect alignment, the overall score of the correct alignment is within 15% of that of the best scoring alignment.

This suggests ways for locating an interesting number of difficult paragraphs and for identifying alternative alignments in these cases. The method we propose proceeds in two passes: the first pass is essentially identical to the length-based algorithm, except that instead of producing the single best solution, it outputs a list of “best alignments”, i.e. a list of alignments whose overall score is relatively “good”. If this does not produce a unique solution, the program then proceeds with the second pass, and uses the cognate-based scoring function described in the previous section to select the best alignment of the list.

In our implementation, an alignment is considered a valid candidate for the second pass if its overall score falls within a certain percentage of the absolute best scoring alignment. Actually, finding *exactly* all of these alignments involves a computation that is exponential in time with the total number of sentences in the paragraphs. We use a heuristic which, while it does find all the interesting alignments in polynomial time, typically slightly over-generates.

Experimentation shows that the best results are obtained by retaining for the second pass all the alignments whose score falls within 30% of the overall best scoring alignment.

5 Evaluation

In evaluating the different alignment methods discussed in the previous sections, we were interested in two things: first, in measuring their overall performance, both in terms of efficiency and of correctness; second, in identifying the respective strengths and weaknesses of each.

Both of these objectives required the existence of a test corpus for which a reference ("correct") alignment was available. Our first concern was to construct such a corpus.

The Test Corpus

The Canadian Hansards (parliamentary proceedings) were chosen as the source for the test corpus because of their wide availability and common use as a testbed for bitextual techniques. For reasons to be discussed later, two distinct corpora were set up: the first corpus consists of 2775 pairs of paragraphs (approximately 160 000 words of each language) and may be considered fairly representative of the Hansard proceedings in terms of difficulty of alignment; the second one is shorter (790 pairs of paragraphs) and was chosen for its relatively large proportion of asymmetric pairs of paragraphs (we call two paragraphs "asymmetric" if they do not contain the same number of sentences). In what follows, we will refer to these as the "base" corpus and the "hard" corpus respectively.

The reference alignments had to be done by hand. All 3565 paragraphs were equally split among 8 judges, all of which speak and read both English and French fluently. With the help of a special-purpose interactive program, these judges were asked to verify and correct an initial alignment produced automatically following a "dumb" method: each pair of paragraphs was segmented into a series of one-to-one alignments, followed if need be by a series of one-to-zero or of zero-to-one alignments. The resulting manual alignment of the base corpus contained 7123 couples and that of the hard corpus, 2693.

Overall Performance Tests

In our first experiment, the base corpus was submitted to each program and the resulting alignment compared to the reference alignment. The results of these tests are summarized in table 3. In this table, error counts are reported first as the number of paragraphs where the machine alignment disagrees with the reference alignment, then as the number of couples of the reference alignment not found in the corresponding machine alignment ("missing" couples). Error percentages are given as the number of missing couples over the total number of couples in the reference alignment. Processing times (where applicable) are in seconds. All tests were done on a Sun SPARCstation 1+ with 24 Megs of RAM.

For reference, we also provide the results of the "dumb" alignment, i.e. the initial alignment from which the reference alignment was produced. As it turns out, the success rate of this alignment is probably the most striking result in table 3: an impressive 90.4%! To a certain extent, this could be said to support the claim that aligning text is an "easy" problem. However, we take it more as an indication that this specific corpus was particularly easy to align. This interpretation is supported by the high rate of success of the length-based method: while Gale and Church report a rate of 95.8%, in our experiment, it scored 98.2%. The two-pass method of section 4 was just slightly better, with a success rate of 98.4%. Separate experiments on other (non-Hansard) corpora seem to confirm this tendency to reduce the absolute number of alignment errors by 10%.

	length-based alignment	*cognate-based alignment*	*two-pass alignment*	*"dumb" alignment*
Paragraphs in which program and reference disagree	*57*	*85*	*58*	*290*
Missing couples	*128*	*171*	*114*	*681*
Error percentages	*1.8%*	*2.4%*	*1.6%*	*9.6%*
Processing time	*99.2*	*908.1*	*111.4*	—

Table 3: Results of alignment programs on the base corpus.

So we conclude that the two-pass program does produce better results than the simple length-based alignment, at a minimal cost (a 12% increase in processing time), but that the improvement remains modest.

Error analysis

A quick look at the errors made by the three programs on the base corpus reveals that a large number of these were "unavoidable" errors: 27 pairs of paragraphs of that corpus featured "unorthodox" translation patterns, i.e. patterns other than the six enumerated in section 2 (e.g. three sentences that translate into one, or two that translate into four), and which therefore could never have been caught by any of the programs we tested. We felt that in order to better assess the behavior of each program, we needed a test corpus for which the rate of success of the length-based method was closer to Gale and Church's own predictions. Hence the "hard" corpus.

It would have been possible to "cook up" such a test corpus, but we discovered that large portions of the Hansards exist that are significantly harder to align than the base corpus used in the first part of the evaluation. This is often true of sections that have a large proportion of asymmetric pairs of paragraphs, as is the case in the sample that we used for our "hard" corpus: 14% of its pairs of paragraphs are asymmetric, while the average is below 10%. On this sample, the length-based method missed 80 of the 2693 couples of the reference alignment, a success rate of 97.0%. The performance of the two-pass method on the same corpus was significantly higher, with only 50 errors (a 37.5% reduction in the number of errors).

We examined the errors that these two programs produced on the "hard" corpus, and determined for each one the assumed source of the error, cataloguing them accordingly. For each type of error, we recorded both the number of "regions" (series of contiguous misaligned couples) in which a certain type of error occurred and the total number of couples affected by the error. The results of this classification of alignment errors appear in Table 4.

The first category concerns errors attributable to unorthodox translation patterns. The second deals with what we call "decomposition" errors: the task of aligning a pair of paragraphs depends on a previous decomposition of the text into sentences. Errors sometimes occur during that process, such as for example when a sentence is split in two because the program takes the period in an abbreviation for an end of sentence marker. As much as possible, an alignment program should correct these errors by regrouping the pieces that composed the original sentence. When it fails to do so, we have a "decomposition" error.

Error type	*length-based alignment*		*cognate-based alignment*		*two-pass alignment*	
	regions	*couples*	*regions*	*couples*	*regions*	*couples*
Unorthodox patterns	*7*	*23*	*8*	*15*	*7*	*18*
Decomposition	*5*	*5*	*5*	*5*	*5*	*5*
Missed omissions	*5*	*16*	*5*	*14*	*5*	*14*
Misplaced contraction/expansion	*10*	*28*	*28*	*58*	*6*	*13*
Others	*3*	*8*	*7*	*10*	*0*	*0*
Total	*30*	*80*	*53*	*102*	*23*	*50*

Table 4: Types of alignment errors.

The third category concerns situations where the alignment program failed to locate an omission or an addition (1-to-0 or 0-to-1 alignments), the fourth deals with misplaced contractions and expansions such as the one of Figure 2 (Section 1), and the last category groups together various other errors.

The three programs exhibited similar behaviors on the first three types of errors:

As expected, they all failed when confronted to unorthodox translation patterns. There were actually 8 couples of this type in the reference alignment: six 3-to-1's and two 4-to-1's. But two of these were very close to one another, and therefore appeared within one large region of error in the length-based and the two-pass alignments. Only the cognate-based program managed to contain each error within a single, relatively small region of error.

All three programs made the same 5 decomposition errors, which in fact all followed the same pattern: two sentences which were mutual translations were both incorrectly split on a period not marking an end of sentence; all programs produced two one-to-one couples, when the correct solution was a single two-to-two. Such errors are very difficult to locate and we may assume that to do so would require much more language-specific information (e.g. enough syntactic knowledge to recognize what constitutes an acceptable sentence).

As for missed omissions, they still constitute the most embarrassing category of errors: all programs missed all 5 omissions that the test corpus contained. Gale and Church suggest that it may be necessary to consider language-specific methods in these cases. Obviously, cognates do not provide enough information to solve this problem.

Where the real differences appear is in the last two categories: on these, the two-pass program managed to get three times less errors than the length-based, which itself was twice as good as the cognate-based. Actually, considering the fairly poor performance of the cognate-based approach in these situations, it is surprising that we could obtain such good results simply by combining it with the length-based method. The most probable explanation is that the length-based and cognate-based methods do not normally make the same mistakes. So when our two-pass strategy effectively locates the length-based method's weaknesses and manages to propose interesting alternatives, then the cognate-based method is likely to come up with the correct alignment in the second pass.

Also worthy of notice is the average number of missing couples per region of error: 2.58 for the length-based alignment, 2.17 for the two-pass alignment, and 1.92 for the cognate-based alignment. This indicates that while the cognate-based method produces substantially more errors than the other two, it is less prone to producing large alignment errors, i.e. errors involving several (say, 3 or more) contiguous couples. This tendency can be observed in all categories but the decomposition errors, where that ratio is already minimal. To a certain extent, the two-pass method inherits this highly desirable property.

Conclusions

In this paper, we outlined some of the weaknesses of Gale and Church's program for aligning sentences in bilingual text, and suggested that a small amount of linguistic knowledge could be used to overcome these weaknesses. Cognates were proposed as such a source of knowledge, and we described methods both to efficiently identify pairs of cognates and to estimate how well two pieces of text related to one another given their "level of cognateness".

We then described how the length-based program could be modified to take advantage of this new information. The new program proceeds in two passes. In the first pass, it uses the length criterion to filter out all unlikely alignments. In the second pass, cognates are used to identify the overall best alignment of those candidates that remain.

Experimentation shows that this method yields better results than the length-based program, but that this improvement remains modest. We believe that the main problem with our approach is that we are trying to improve a fairly reliable method with one that is not as reliable.

However, we observe that while cognates are less precise than length as an alignment criterion, they are probably more robust: the two-pass program is less likely to misalign large pieces of text in pathological situations such as those described in section 1. In this sense, we believe the method could be used to produce finer alignments, e.g. aligning segments smaller than sentences, or considering more translation patterns (3-to-1's, etc.).

On the other hand, the method we describe to locate difficult regions and identify alternatives in the first pass is remarkably reliable. Obviously, scoring differences could be used in a length-based alignment program as the basis for an error-detection mechanism, or to purge a sample of bitext of dubious pairs of segments.

Another interesting aspect of the two-pass strategy is that it allowed us to use a relatively "expensive" alignment criterion, without sacrificing efficiency. This idea can be generalized so that other sources of language-specific knowledge are used in the alignment process. For example, one could complement or replace the cognate-matching mechanism with a bilingual dictionary, or with some device capable of evaluating the probability of two words "trans"-occurring in segments which are mutual translations. Such additions would probably make the second pass more reliable, and in turn allow us to filter out less candidates in the first pass, therefore becoming less dependent on the length criterion.

We plan to explore these avenues in the near future.

References

Brown, Peter F., Jennifer C. Lai and Robert L. Mercer, 1991, "Aligning Sentences in Parallel Corpora", in *COLING 91*, pp. 169-176.

Brown, Peter F., John Cocke, Stephen A. DellaPietra, Vincent J. DellaPietra, Frederick Jelinek, John D. Rafferty, Robert L. Mercer and Paul S. Roossin, 1990, "A Statistical Approach to Machine Translation", *Computational Linguistics*, 16(2), pp. 79-85, June 1990.

Gale, William A. and Kenneth W. Church, 1991, "A Program for Aligning Sentences in Bilingual Corpora", in *COLING 91*, pp. 177-184.

Harris B., 1988, "Bi-Text, a New Concept in Translation Theory", in *Language Monthly*, #54, pp. 8-10.

Isabelle, Pierre, 1991, "Une nouvelle génération d'aides à la traduction et la terminologie", paper presented at the *Terminologie et documentation* colloquium, Hull, October 1991.

Acknowledgments

Many thanks go to Marc Dymetman for his constructive comments on the more technical aspects of the paper. Many thanks also to Peter Brown for pointing out important omissions in the original draft.

Finally, we want to thank Marie-Louise Hannan, Jean-Marc Jutras, Elliott Macklovitch, François Perrault and Xiaobo Ren for their contribution in producing a 600 000-word hand-aligned corpus of bitext.