Text-Translation Alignment: Three Languages Are Better Than Two *

Michel Simard
Laboratoire de recherche appliquée en linguistique informatique (RALI)
Université de Montréal
SimardM@IRO.UMontreal.CA

Abstract

In this article, we show how a bilingual text-translation alignment method can be adapted to deal with more than two versions of a text. Experiments on a trilingual corpus demonstrate that this method yields better *bilingual* alignments than can be obtained with bilingual text-alignment methods. Moreover, for a given number of texts, the computational complexity of the multilingual method is the same as for bilingual alignment.

Introduction

While bilingual text corpora have been part of the computational linguistics scene for over ten years now, we have recently witnessed the appearance of text corpora containing versions of texts in three or more languages, such as those developed within the CRATER (McEnery et al., 1997), MULTEXT (Ide and Véronis, 1994) and MULTEXT-EAST (Erjavec and Ide, 1998) projects. Access to this type of corpora raises a number of questions: Do they make new applications possible? Can methods developed for handling bilingual texts be applied to multilingual texts? More generally: is there anything to gain in viewing multilingual documents as more than just multiple pairs of translations?

Bilingual alignments have so far shown that they can play multiple roles in a wide range of linguistic applications, such as computer assisted translation (Isabelle et al., 1993; Brown et al., 1990), terminology (Dagan and Church, 1994) lexicography (Langlois, 1996; Klavans and Tzoukermann, 1995; Melamed, 1996), and cross-language information retrieval (Nie et al., 1998). However, the case for trilingual and multilingual alignments is not as clear. True multilingual resources such as multilingual glossaries are not widely used, and most of the time, when such resources exist, the real purpose is usually to provide *bilingual* resources for multiple pairs of languages in a compact way.

What we intend to show here is that while multilingual correspondences may not be interesting in themselves, multilingual text alignment techniques can be useful as a means of extracting information on *bilingual correspondences*. Our idea is that each additional version of a text should be viewed as valuable information that can be used to produce better alignments. In other words: whatever the intended application, three languages are better than two (and, more generally: the more languages, the merrier!).

After going through some definitions and preliminary material (Section 1), we present a general method for aligning three versions of a text (Section 2). We then describe some experiments that were carried out to evaluate this approach (Section 3) and various possible optimizations (Section 4). Finally, we report on some disturbing experiments (Section 5), and conclude with directions for future work.

1 Trilingual Alignments

There are various ways in which the concept of alignment can be formalized. Here, we choose to view alignments as mathematical relations between linguistic entities:

Given two texts, A and B, seen as sets of *linguistic units*: $A = \{a_1, a_2, ..., a_m\}$ and $B = \{b_1, b_2, ..., b_n\}$, we define a **binary alignment** X_{AB} as a relation on $A \cup B$:

$$X_{AB} = \{(a_1, b_1), (a_2, b_2), (a_2, b_3), ...\}$$

* This research was funded by the Canadian Department of Foreign Affairs and International Trade (http://www.dfait-maeci.gc.ca/), via the *Agence de la francophonie* (http://www.francophonie.org)

The interpretation of X_{AB} is: (a, b) belongs to X_{AB} if and only if some translation equivalence exists between a and b, total or partial.

This definition of alignment, inspired from Kay and Röscheisen (1993), can be naturally extended to accommodate any number of versions of a text. In general, we will say that, given N versions of a text $A_1, ..., A_N$, a **N-lingual alignment** $X_{A_1^N}$ is a relation on $\cup_{i=1}^{N} A_i$.

Clearly, a N-lingual alignment can be obtained by combining pairwise bilingual alignments. For example, with three texts A, B and C, and three alignments X_{AB}, X_{BC} and X_{CA}, one can easily obtain the trilingual alignment X_{ABC} as $X_{AB} \cup X_{BC} \cup X_{CA}$. In fact, in all that follows, we indifferently refer to trilingual alignments as unique relations or as triples of bilingual alignments. Conversely, any smaller-degree alignment can be extracted as a subset of a N-lingual alignment, by projecting the relation onto a given "plane".

Another thing that becomes apparent as soon as more than two languages are involved is that text-translation alignments appear to be *equivalence relations*, which means that they generally display the properties of *reflexivity*, *symmetry* and *transitivity*:

- **reflexivity**: Any word or sequence of words aligns with itself – which is natural, insofar as we extend the notion of "translation", so as to include the translation from one language to itself...
- **symmetry**: if a in language A is aligned with b in language B, then we expect b to align with a. In other words, alignment is not "directional".
- **transitivity**: if a aligns with b, and if b itself aligns with c, then a aligns with c.

Although there are limits to the applicability of these mathematical properties to real-life translations, the case of transitivity is particularly interesting, as we will see later on.

Translation equivalences can be viewed at different levels of *resolution*, from the level of documents to those of structural divisions (chapters, sections, etc.), paragraphs, sentences, words, morphemes and eventually, characters. In general, it seems quite clear that the smaller the units, the more interesting an alignment is likely to be (although we can question the interest of a character-level alignment). However, in the experiments described here, we focus on alignment at the level of sentences, this for a number of reasons: First, sentence alignments have so far proven their usefulness in a number of applications, e.g. bilingual lexicography (Langlois, 1996; Klavans and Tzoukermann, 1995; Dagan and Church, 1994), automatic translation verification (Macklovitch, 1995; Macklovitch, 1996) and the automatic acquisition of knowledge about translation (Brown et al., 1993). Also, the sentence alignment problem has been widely studied, and we could even say that at this point in time, a certain consensus exists regarding how the problem should be approached.

On the other hand, not only is the computation of finer-resolution alignments, such as phrase- or word-level alignments, a much more complex operation, it also raises a number of difficult problems related to evaluation (Melamed, 1998), which we wanted to avoid, at least at this point. Finally, we believe that the concepts, methods and results discussed here can be applied just as well to alignments at other levels of resolution.

2 A General Method for Aligning Multiple Versions of a Text

Existing alignment algorithms that rely on the optimality principle and dynamic programming to find the best possible sentence alignment, such as those of Gale and Church (1991), Brown et al. (1991), Simard et al. (1992), Chen (1993), Langlais and El-Bèze (1997), Simard and Plamondon (1998), etc. can be naturally extended to deal with three texts instead of two, or more generally to deal with N texts. While the resolution of the bilingual problem is analogous to finding an optimal path in a rectangular matrix, aligning N texts is analogous to the same problem, this time in a N-dimensional matrix. Normally, these methods produce alignments in the form of *parallel segmentations* of the texts into equal numbers of segments. These segmentations are such that 1) segmentation points coincide with sentence boundaries and 2) the k^{th} segment of one text and the k^{th} segment of all others are mutual translations. We refer to such alignments as *non-crossing alignments* (see Figure 1 for an example).

La vraie question posée par cette controverse est la suivante: ¶ qu'est ce que la pensée? ¶	Behind this debate lies the question, What does it mean to think? ¶
Elle mystifie l'humanité (seule, apparemment, à pouvoir penser) depuis des millénaires. ¶	The issue has intrigued people (the only entities known to think) for millennia. ¶
Des ordinateurs qui ne pensent pas ont cependant réorienté la question et éliminé diverses réponses. ¶	Computers that so far do not think have given the question a new slant and struck down many candidate answers. ¶
La vraie réponse reste cependant inconnue. ¶	A definitive one remains to be found. ¶
L'esprit est-il un programme d'ordinateur? ¶	Is the brain's Mind a Computer Program? ¶
JOHN SEARLE ¶ Non : ¶ un programme manipule seulement des symboles, mais le cerveau leur donne un sens. ¶	No. ¶ A program merely manipulates symbols, whereas a brain attaches meaning to them. ¶ by John R. Searle ¶

Figure 1: Bilingual Non-crossing Sentence Alignment – "¶"'s denote sentence boundaries; horizontal lines represent segmentation points.

This definition covers a subset of alignments as defined in Section 1. It is therefore always possible to represent a non-crossing alignment as an equivalence relation. However, the converse is not true: in particular, and as their name suggests, one cannot explicitly represent *inversions* with such alignments, i.e. situations where the order of sentences is not preserved across translation. In spite of this limitation, these alignments cover the vast majority of situations encountered in real-life texts, at least at the level of sentences (Gale and Church, 1991).

There is a problem with extending this class of alignment algorithms to deal with the general N-dimensional case, however: the computational complexity of the algorithm increases multiplicatively with each new language. For instance, the space and time complexity of the trilingual version of the Gale and Church (1991) program would be $\Theta(N^3)$. The use of such an algorithm quickly becomes prohibitive (for example: 1,000,000 computation steps for texts of 100 sentences each). Of course, in the case of bilingual alignment, it is common practice to restrict the search, for instance to a narrow corridor along the main "diagonal" (see Simard and Plamondon (1998), for example). But even with such heuristics, it is quite clear that in general, the search-space will grow multiplicatively with each new language.

Nevertheless, the idea of aligning multiple versions of a text simultaneously is intuitively appealing: while the alignment operation will no doubt be more complex than with two languages, every new version brings some additional information, which we should be able to make good use of (see Figure 2). Therefore, we will want to find a way to overcome the complexity issues.

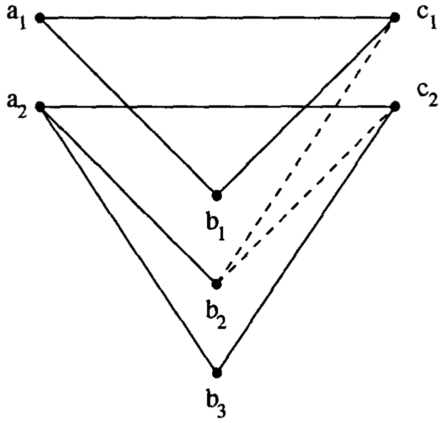

Figure 2: Three texts are easier to align than two – In the face of an uncertainty regarding correspondences between b_2, c_1 and c_2: the absence of evidence for (a_2, c_1) or (a_1, b_2) correspondences suggests rejecting (b_2, c_1), while a similar reasoning supports (b_2, c_2).

We know that the multilingual alignment problem is related to a number of other sequence comparison problems, with applications in various domains. In particular, molecular biologists are concerned with relating sequences of nucleotides (in DNA or RNA molecules) and of amino acids (in proteins) (Sternberg, 1996). The methods used to attack these problems are very similar to those used in translation alignment, and rely largely on dynamic programming. In practice, researchers in molecular biology have observed that, insofar as the input sequences are not excessively dissimilar, the greater the number of sequences, the better the alignments obtained. Therefore, numerous strategies have been proposed to alleviate the complexity issues related to multiple sequence comparison (Chan et al., 1992). One common heuristic approach is to reduce the search-space, either in width (i.e. by concentrating the search around the "diagonal"), or in depth (i.e. by first segmenting the input sequences at judicious points, and then aligning the subsequences). Of course, these strategies are also widely used in text-translation alignment.

However, the most widespread approach is to construct multiple alignments by iteratively combining smaller-degree alignments. While these methods are not generally optimal, they still produce good results in most situations. More importantly, for a given number of sequences, they usually work in quadratic time and space. The general idea is to first compare sequences two-by-two, so as to measure their pairwise similarity; based on the result of this operation, an order of alignment is determined — typically, the most similar pairs will be aligned first; the final multiple alignment is produced by gradually combining alignments (see, for example, Barton and Sternberg (1987)).

This approach can be directly adapted to the trilingual text alignment problem. The idea is simple: given three versions of a text A, B and C, in three different languages, we first determine which of the three pairs AB, BC or AC is the most "similar". Let us suppose that this is the AB pair. We then align this pair, using whatever bilingual alignment program we have at hand, producing X_{AB}; we then align text C with this alignment, thus producing X_{ABC}.

To implement this idea, we need to answer two questions: First, how to measure the similarity between different versions of a text? And second, what does it mean to align a text with an alignment?

There are certainly numerous possible answers to the first question. But actually, statistical alignment methods such as those derived from Gale and Church (1991) provide us with a simple solution: to find the best alignment, these methods explore different alignment hypotheses, and select the one with the highest probability with regard to a certain statistical model of translation. Therefore, at the end of the operation, a statistical alignment program has at its disposal an overall score for the best alignment, in the form of a global probability. In practice, we observe that this score is a good indicator of the similarity between two texts. For instance, Gale and Church used this score to identify dubious regions in their alignments[1].

Therefore, to determine the most similar pair of texts, we propose to compute the bilingual alignments X_{AB}, X_{BC} and X_{AC}, and to compare the final alignment scores. Of course, for this exercise to be meaningful, we must make sure that the scores associated with the bilingual alignments are indeed comparable. In general, if the same alignment method is used with comparable translation models for all pairs of languages, this should not be a problem.

Once the most similar pair of versions has been identified, say A and B, and we have computed a bilingual alignment for that pair, we are ready to tackle the problem of aligning the remaining text C with the X_{AB} alignment. In practice, this will amount to aligning the elements of C (in our case, sentences) with individual "couples" of the X_{AB} relation: whenever we align some sentence $c \in C$ with a sentence $a \in A$, then this implies that c must also be aligned with all other sentences to which a is related within the transitive closure of X_{AB}. In other words, this alignment method is "inherently transitive".

In practice, the alignment of X_{AB} and C is dealt with just like a bilingual alignment: the X_{AB} alignment is viewed as a sequence of items, and dynamic programming is used to find the best alignment with the sentences of C. The only real difference lies in how individual "triples" are scored. Here again, we turn to molecular biology, where experience seems to show that the "joint similarity" of multiple items can be measured as the linear combination of all pairwise comparisons:

$$s(a_1, ..., a_N) = \sum_{i<j} s(a_i, a_j)$$

This sort of combination supposes that all binary scoring functions $s(a_i, a_j)$ are comparable (Carillo and Lipman, 1988). Once again, this will not be a problem if we plan to use analogous translation models for all language pairs.

To sum up, given three versions of a text A, B and C, we propose the following trilingual alignment method:

1. Compute initial bilingual alignments X_{AB},

[1] Also recall that the dynamic programming approach to text alignment actually derives from a classic algorithm to measure the "edit distance" between two strings (Wagner and Fischer, 1974)

X_{BC} and X_{AC};

2. Using the final alignment score, identify the most similar pair (say, AB);
3. Align the remaining text (C) with the initial alignment of the retained pair (X_{AB}); the result is a trilingual alignment X_{ABC};

The computational complexity of this method is essentially the same as that of the underlying bilingual alignment method, both in terms of time and space. In practice, aligning three texts this way takes about the same amount of memory as aligning one pair, and about four times as much computation time.

3 Evaluation

We have implemented a trilingual sentence-alignment program called *trial*, based on the approach presented in Section 2 and on a bilingual sentence-alignment program called *sfial*, which implements a modified version of the method of Simard et al. (1992). In *sfial*, we essentially combine into a statistical framework two criteria: the length-similarity criterion proposed by Gale and Church (1991) and a "graphemic resemblance" criterion based on the existence of cognate words between languages. This method was chosen because it is simple, it requires a minimum of language-specific knowledge, and because it is representative of the kind of approaches that are typically used for this task, at least for aligning between closely-related languages such as German, English, French, Spanish, etc. Furthermore, in a recent sentence-alignment "competition" held within the ARCADE project (Langlais et al., 1998), the three top-ranking systems relied at least partially on cognates, and two of them were derived directly from the Simard et al. (1992) method.

To test the performance of the *trial* program, we needed a performance metric and a test corpus. Following the work of the ARCADE project, we decided to measure performance in terms of *alignment recall*, *precision* and *F-measure*, computed on the basis of sentence lengths (measured in terms of characters). In our experience, this set of metrics is the most generally useful.

Our test corpus was *The Gospel According to John*, in English (*New International Version*), French (Louis Segond translation) and Spanish (*Reina Valera* version). All versions were obtained via the *Bible Gateway* (`http://www.gospelcom.net`). For the needs of the evaluation, we manually segmented all three versions of the text into sentences, and then produced reference sentence alignments, using the *Manual* system (Simard, 1998c). This corpus and its preparation are described in more details in Simard (1998a).

Languages		*sfial*	*trial*
Spanish -French	precision	0.997	0.997
	recall	0.989	0.989
	F	0.993	0.993
French -English	precision	0.956	0.962
	recall	0.941	0.941
	F	0.948	0.951
English -Spanish	precision	0.952	0.950
	recall	0.936	0.943
	F	0.944	0.947

Table 1: Precision, recall and *F-measure* of alignments produced by *sfial* and *trial*, on *The Gospel According to John*, French, English and Spanish versions.

The test-corpus was submitted to both *sfial* and *trial*; the results of this experiment are reproduced in Table 1. The Spanish-French pair was identified by *trial* as being the most similar (not surprisingly, English-Spanish was the most dissimilar). Since the alignment of the most similar pair is used as the basis of the trilingual alignment, the results obtained by *sfial* and *trial* for this pair are identical. On the other hand, for the two other pairs, the *trial* method seems to improve the quality of the alignments, but the gains are minimal.

A close examination of the results quickly reveals what is going on here: As mentioned earlier, our trilingual alignment method is "inherently transitive"; in fact, it naturally produces alignments which are transitively closed. In doing so, it sometimes run into some natural limitations of the applicability of transitivity to real-life translations. Take the following example: suppose that the word *weak* in an English text is rendered in French as *sans force* ("without strength") and in Spanish as *sin fortaleza*. A transitively closed trilingual alignment will contain the correct correspondences (*sans*, *sin*) and (*force*, *fortaleza*), but also the correspon-

Languages		before	after
French -English	precision	0.962	0.979
	recall	0.941	0.938
	F	0.951	0.958
English -Spanish	precision	0.950	0.970
	recall	0.943	0.938
	F	0.947	0.954

Table 2: Impact of re-segmenting couples involving more than a single pair of sentences in the alignment produced by *trial.*

dences (*sans*, *fortaleza*) and (*force*, *sin*), which are superfluous. Such *contractions* and *expansions* happen all the time in real-life translations, not only at the level of words, but at the level of sentences as well. As a result, transitively closed alignments of three texts or more will usually display a lower precision than bilingual alignments.

To compensate for this "transitivity noise", we decided to apply a final post-processing step: for each pair of languages, whenever the trilingual alignment produced by *trial* connects two pairs of sentences or more, we evaluate the impact of re-segmenting the corresponding region of text (in other words, we perform a local bilingual alignment). Typically, this operation can be carried out in near-linear time and space.

Table 2 shows the impact of this procedure on the *trial* alignments of *The Gospel According to John* (the initial bilingual alignment is not submitted to re-alignment, and so the results for the French-Spanish pair is not reproduced here). What we observe is a significant improvement of precision, and a slight decrease in recall. Compared to the *sfial* bilingual alignment, the overall improvement (*F*-measure) is approximately 1%: all figures being in the 0.95 area, this corresponds to a 20% reduction in the total error. Therefore, it would indeed seem that our final post-processing is effective, and that in the end, "three languages are better than two".

4 Optimizations

In addition to all the usual optimizations to bilingual alignment methods, various things can be done to reduce computation times in the trilingual alignment method of Section 2: for instance, individual bilingual scores from step 1 can be recorded in memory, to be later reused in step 3. Also, if multiple processors are available, the three initial alignments of step 1 can be done in parallel. By combining these optimizations, it is possible to align three texts in less than twice the time necessary to align a single pair.

Another possible optimization is to initially segment the three texts in parallel, so as to perform step 3 on smaller pieces. Of course, this idea is not new, but what makes it particularily appealing for trilingual alignment, in addition to the usual reduction in the needed time and space, is the potential for further improvements in the quality of the resulting alignments: In the method outlined above, we have chosen to base the re-alignment on the initial alignment that connected the two most similar versions of the text. In reality, nothing proves that this similarity is "evenly distributed" on the totality of the texts. In fact, if we segmented the input texts at arbitrary points, we might very well discover that the most similar pair of languages is not always the same. If this is the case, then we could improve our results by doing the re-alignment in small chunks, each time basing the re-alignment on the pair of languages that locally displays the best similarity.

On the other hand, this approach also carries its own risks. Indeed, by pre-segmenting the three texts in parallel, we will be fixing points in the alignment *a priori*, namely those points at the boundaries between segments. This is why it is crucially important to select segmentation points judiciously: we will want these to lie in areas where all three initial alignments agree and each display a high level of confidence.

In practice, such "points of agreement" between the initial bilingual alignments can be found by computing their *transitive closure*, i.e. by adding to the union of the three alignments all couples whose existence is predicted by transitivity (a simple procedure for this can be found in Hopcroft and Ullman (1979)). From such transitively closed trilingual alignments emerge "islands of correspondence", i.e. groups of sentences that are all related to one another. In between these islands lie natural segmentation points, that can be viewed as points of agreement between the three initial alignments.

We also found that to obtain the best possible segmentation of the texts, it was necessary to select among such points of agreement only those lying between pairs of islands of correspondence for which we have a high degree of confidence. To measure this "confidence", we currently use two criteria: first, the number of sentences of each language in the surrounding islands; and second, the alignment program's own scoring function. The first criterium is based on the simple observation that most alignment errors happen when the translation diverges from the usual pattern of "one sentence translates to one sentence" (Simard, 1998b); so we only consider points of agreement lying between "1-to-1-to-1" islands. The second criterium is based on the observation by Gale and Church (1991) that good alignments usually coincide with high scoring regions of text.

To sum up, our optimized trilingual alignment method follows these lines:

Given three versions of a text A, B and C,

1. Compute initial bilingual alignments X_{AB}, X_{BC} and X_{AC};
2. Segment the texts:
   (a) Identify points of agreement between X_{AB}, X_{BC} and X_{AC}, by computing the transitive closure X^{+}_{ABC} of $X_{AB} \cup X_{BC} \cup X_{AC}$;
   (b) Among these points, select those points that lie between pairs of 1-1-1 triples within which individual bilingual alignment scores do not exceed some threshold T;
   (c) Segment A, B and C at these points, thus producing sub-segments $A_1 ... A_n$, $B_1 ... B_n$ and $C_1 ... C_n$.
3. Jointly align each triple of segments (A_i, B_i, C_i) as with the *trial* method (Section 2), and obtain the final trilingual alignment as the union of all partial alignments $X_{ABC} = \cup X_{A_i B_i C_i}$;

This optimization was implemented into the *trial* program, thus producing a program we call *trial++*. *The Gospel According to John*, in French, English and Spanish was then submitted to this new program. To a certain degree, the results of this experiment were a disappointment, since they turned out to be virtually identical to those obtained with the *trial* program.

step	*trial*	*trial++*
bilingual alignment	113	113
pre-segmentation	–	12
re-alignment	69	18
Total	182	143

Table 3: Execution times in seconds for individual computation steps of *trial* and *trial++* (French, Spanish and English versions of *The Gospel According to John*).

A closer examination reveals what is going on: in 201 out of the 279 segments produced by the pre-segmentation procedure, *trial++* chose to base the re-alignment on the same alignment as *trial*, i.e. the Spanish-French. No differences could therefore be expected between the two programs on these segments. Of the 78 remaining segments, 60 contained exactly one sentence per language, so not much improvement could be expected for those either. In the end, only 18 segments remained where *trial* and *trial++* had the potential to produce different alignments; but even here, both programs produced birtually identical results.

The main difference between *trial* and *trial++* was execution times. Table 3 shows the times required for each of the computation steps of the two programs. What we observe is that pre-segmentation reduces execution times significantly, without hampering the quality of the alignments. We can therefore consider that this is a useful step, especially if we are dealing with long texts[2].

5 A Disturbing Experiment...

We mentioned earlier that to compute trilingual alignments, directly extending dynamic programming bilingual alignment methods was not a realistic approach from the point of view of computational complexity. However, it seems that the optimization described in Section 4 for pre-segmenting the three texts into small segments before performing the trilingual alignment could actually help resolve the problem: if we manage to segment the input texts into small enough chunks, then a cubic-order algo-

[2] Also worth noting here is that pre-segmentation is currently carried out by a Perl script. With a proper C implementation, execution times for this step would probably become negligible.

Languages		*sfial*	*trial*	*trial*−−
Spanish	precision	0.997	0.997	0.992
-French	recall	0.989	0.989	0.984
	F	0.993	0.993	0.988
French	precision	0.956	0.979	0.959
-English	recall	0.941	0.938	0.919
	F	0.948	0.958	0.938
English	precision	0.952	0.970	0.952
-Spanish	recall	0.936	0.938	0.914
	F	0.944	0.954	0.933

Table 4: Precision, recall and *F-measure* of *sfial*, *trial* and *trial*−−.

rithm may not be so problematic after all.

To test this conjecture, we implemented the following method into a program called *trial*−− (the origin of the name will become clear in a moment):
Given three versions of a text A, B and C,

1. Compute initial bilingual alignments X_{AB}, X_{BC} and X_{AC};
2. Pre-segment the texts, as in Section 4;
3. Align each triple of sub-segments (A_i, B_i, C_i), using dynamic programming to find the optimal alignment $X_{A_iB_iC_i}$ in the $A_i \times B_i \times C_i$ space;
4. Obtain the final trilingual alignment as the union of all partial alignments $X_{ABC} = \cup_i = 1...nX_{A_iB_iC_i}$;

Once again, and as in the *trial* and *trial++* methods, we finish up with a final bilingual alignment pass to compensate for "transitivity noise".

This new program was succesful in aligning *The Gospel According to John* , using amounts of time and memory comparable to the *trial++* program. However, the resulting alignments were quite different, as can be seen in the performance figures in Table 4.

The performance of this new program on *The Gospel According to John* turns out to be not only poorer than that of the *trial* program, but also poorer than that of the *sfial* program! What is going on, here? After all, we would expect *trial*−− to be *The* optimal method, of which all others are heuristic approximations. Although we have no definitive answer to this question, we see two different lines of explanation.

The first and most obvious possibility is that our initial assumption, namely that three languages are better than two, is false. In other words: aligning three texts is at least as hard as aligning two, and possibly harder. Of course, this would also imply that the results obtained with the *trial* and *trial++* methods were mere accidents. Although this explanation for the failure of the *trial*−− approach clearly contradicts our initial intuitions, we cannot entirely reject it. We do not, however, pursue this line any further. (Besides, it doesn't go any further!)

The second line of explanation leads us straight to the scoring function of our alignment methods. As in the *trial* and *trial++* methods, the scoring function used in *trial*−− is the sum of all pairwise (bilingual) alignment scores. It could simply be that this way of measuring "trilingual fit" is inadequate.

However, we believe that our problems run deeper. To begin with, it could be argued that what a trilingual alignment program really needs is a true "trilingual translation model"; it is not at all clear that three bilingual translation models are an adequate substitute for this. Even if it were, we know that there are numerous problems with the "length" and "cognate" stochastic models on which the *sfial* scoring function is based. For instance, we know that while these models usually describe the phenomena observed in translations adequately, they are not necessarily as good when it comes to things that *are not* translations.

While these weaknesses do not appear to cause too many problems when computing bilingual alignments, it would not be surprising that a third language is all it takes for incoherences to creep out and performance to degrade. If this is indeed the case, then this is one more argument in favor of the *trial* approach: by treating multilingual alignments the same way as bilingual alignments, this approach may let us get away with poor translation models, at least until we come up with something better!

Conclusion and Future Work

We have showed how an existing bilingual text alignment method could be adapted to align three versions of a text simultaneously. The computational complexity of the resulting trilingual alignment method is the same as that of the underlying bilingual method, and various optimizations are possible. In experiments on English, French and Spanish versions of *The Gospel According to John*, this approach produced sentence alignments significantly better than those obtained using a bilingual alignment program.

All tests reported here were conducted on a single, relatively small corpus of text. The contradictory results reported in Section 5 highlight the need for more work with regard to evaluation. Such an evaluation exercise would normally imply putting together a much larger and more varied test corpus, segmented and aligned by hand, a process which is known to be costly. However, since the goal is to measure improvements relative to existing bilingual alignment methods, an interesting alternative would be to perform a relative evaluation instead: the programs could then be tested on a much larger test-corpus, and the performance of each system would be measured on only those portions of text where the alignments differ. The details of such an evaluation need to be worked out.

Also, it remains to be seen how the *trial* approach would adapt to the general multilingual case (three languages or more) on the one hand, and to the more challenging problem of finer-grained alignments on the other hand. Here again, we will likely encounter numerous complications, most notably regarding questions of evaluation. Ongoing work in the word-alignment track of the ARCADE project is likely to bring interesting results regarding this question (Langlais et al., 1998).

Finally, and probably more importantly, working with more than two languages has highlighted weaknesses in the modeling of translation that underlies our alignment methods. Much work remains to be done in this direction.

Acknowledgements

Many of the ideas expressed here emerged from informal exchanges with Fathi Debili and Pierre Isabelle; I am greatly indebted to both for their constant support throughout this project. I also wish to thank the anonymous reviewers for their constructive comments on the paper.

References

1991. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berkeley, California, June.

1996. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA)*, Montréal, Canada.

Geoffrey J Barton and Michael J E Sternberg. 1987. A Strategy for the Rapid Multiple Alignment of Proteine Sequences. *Journal of Molecular Biology*, 198:327–337.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

Peter Brown, Jennifer C. Lai, and Robert Mercer. 1991. Aligning Sentences in Parallel Corpora. In ACL-29 (ACL, 1991).

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2).

Humberto Carillo and David Lipman. 1988. The Multiple Sequence Alignment Problem in Biology. *SIAM Journal of Applied Mathematics*, 48(5).

S.C. Chan, A.K.C. Wong, and D.K.U. Chiu. 1992. A Survey Of Multiple Sequence Comparison Methods. *Bulletin of Mathematical Biology*, 54(4):563–598.

Stanley F. Chen. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, Ohio, June.

Ido Dagan and Ken W. Church. 1994. TERMIGHT: Identifying and Translating Technical Terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP)*.

Tomaž Erjavec and Nancy Ide. 1998. The MULTEXT-East Corpus. In LREC-1 (LRE, 1998).

William A. Gale and Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. In ACL-29 (ACL, 1991).

John E. Hopcroft and Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Lan-*

guages and Computation. Computer Science. Addison-Wesley Publishing Company.

Nancy Ide and Jean Véronis. 1994. MULTEXT (Multilingual Text Tools and Corpora). In *Proceedings of the International Conference on Computational Linguistics (COLING) 1994*, Kyoto, Japan, August.

Pierre Isabelle, Marc Dymetman, George Foster, Jean-Marc Jutras, Elliott Macklovitch, François Perrault, Xiaobo Ren, and Michel Simard. 1993. Translation Analysis and Translation Automation. In *Proceedings of the 5th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Kyoto, Japan.

Martin Kay and Martin Röscheisen. 1993. Text-translation Alignment. *Computational Linguistics*, 19(1).

Judith Klavans and Evelyne Tzoukermann. 1995. Combining Corpus and Machine-readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, 10(3).

Philippe Langlais and Marc El-Bèze. 1997. Alignement de corpus bilingues : algorithmes et évaluation. In *Proceedings of 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF (JST)*, Avignon, April.

Philippe Langlais, Michel Simard, Jean Véronis, Susan Armstrong, Patrice Bonhomme, Fathi Debili, Pierre Isabelle, Emna Souissi, and P. Théron. 1998. ARCADE: A Cooperative Research Project on Parallel Text Alignment Evaluation. In LREC-1 (LRE, 1998).

Lucie Langlois. 1996. Bilingual Concordances: A New Tool for Bilingual Lexicographers. In AMTA-2 (AMT, 1996).

1998. *Proceedings of the First International Conference on Language Resources & Evaluation (LREC)*, Granada, Spain.

Elliott Macklovitch. 1995. TransCheck — or the Automatic Validation of Human Translations. In *Proceedings of the Fifth Machine Translation Summit*, Luxembourg.

Elliott Macklovitch. 1996. Peut-on vérifier automatiquement la cohérence terminologique? *META*, 41(3).

A.M. McEnery, A. Wilson, F. Sánchez-Leon, and A. Nieto-Serrano. 1997. Multilingual Resources for European Languages: Contributions of the CRATER Project. *Literary and Linguistic Computing*, 12(4).

I. Dan Melamed. 1996. Automatic Construction of Clean Broad-coverage Translation Lexicons. In AMTA-2 (AMT, 1996).

I. Dan Melamed. 1998. Manual Annotation of Translational Equivalence: The Blinker Project. Technical Report 98-06, IRCS, Philadelphia PA.

Jian-Yun Nie, Pierre Isabelle, Pierre Plamondon, and George Foster. 1998. Using a Probabilistic Translation Model for Cross-language Information Retrieval. In *Proceedings of the 6th ACL Workshop on Very Large Corpora (WVLC)*, Montréal, Canada.

Michel Simard and Pierre Plamondon. 1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation*, 13(1):59–80.

Michel Simard, George Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the 4th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Montréal, Québec.

Michel Simard. 1998a. Projet TRIAL : Appariement de texte trilingue. URL: `http://www-rali.iro.umontreal.ca/Trial`.

Michel Simard. 1998b. RALI-ARCADE : Analyse des erreurs d'alignement commises par Salign sur les corpus BAF et JOC. URL: `http://www-rali.iro.umontreal.ca/-arc-a2/analyse-erreurs`.

Michel Simard. 1998c. The BAF: A Corpus of English-French Bitext. In LREC-1 (LRE, 1998).

M. J. E. Sternberg, editor. 1996. *Protein Structure Prediction – A Practical Approach.* Oxford University Press, Oxford.

Robert A. Wagner and Michael J. Fischer. 1974. The String-to-string Correction Problem. *Journal of the ACM*, 21(1):168–173.