Language, Information and Computation(PACLIC 11), 1996, 63-72

IMPROVING AUTOMATED ALIGNMENT IN MULTILINGUAL CORPORA

J.A. Campbell, N. Chatterjee

Dept. of Computer Science, University College London,
Gower Street, London WC1E 6BT, England
Email: {jac, nchatter}@cs.ucl.ac.uk

Alex Chengyu Fang

Dept. of Phonetics & Linguistics, University College London
Gower Street, London WC1E 6BT, England
Email: alex@phonetics.ucl.ac.uk

M. Manela

Mobile Systems International plc, 1 Harbour Exchange Square,
London E14 9GE, England
Email: mauro@msi-uk.com

Abstract

We report on methods of improving multilingual text alignments that have been produced in a simple dynamic-programming scheme, by automated detection of possible misalignments. Details of methods involving cognates, specially-identified words, and propositional contents of sentences are given, together with notable features of their performance on parallel corpora in a number of different types of European languages.

1. Background

In a previous paper (Campbell and Fang 1995) we described the background to our studies of automated alignment in multilingual corpora, and summarised our contribution to the European-Union-sponsored research project TRANSLEARN, as well as indicating some directions for future research to continue after the end of that project. The present paper reports our work since then.

The European Union (EU) has 11 official languages, with the possibility that more may be added as countries with relatively advanced economies in eastern Europe make progress in their negotiations to join the Union. Many EU documents are required to appear in all of the official languages, which puts an increasing burden on translators at a time when budgets to pay for translations seem actually to be decreasing. This is part of the explanation for the increasing emphasis given to natural-language processing within EU programmes of research and development. TRANSLEARN was one component of the LRE programme, and a successor to LRE (LE, or Language Engineering) has now been defined. While recognising that human translators will remain indispensable to the proper functioning of EU institutions, the designers of these programmes have been concerned to promote research leading to the improved automation of those parts of translators' tasks that can be automated.

Most EU documents have some conventional structure, use of formulas etc., which can be recognised and exploited in the creation of subsequent documents. Because of this, a translation of a new document from one official language to another may not require the activity of translation to be carried out ab initio: it may be more efficient to retrieve a similar document that exists already in a target-language version, and carry out editing operations on it to build the desired result. This remark is the starting-point of TRANSLEARN, which has used parallel corpora of EU documents in English, French, Greek and Portuguese as a basis for its work. In considering a translation of a document from language A to language B, the TRANSLEARN approach is to search a corpus in A for the most similar material (at paragraph or sentence level, typically; considering the document level is practicable but apparently too coarse to be interesting to translators, and we have not yet looked at sub-sentence levels systematically) to an indicated part of the source document, retrieve a corresponding block of text in B by using alignments that have already been imposed on the parallel A-B corpora, and then to offer that block to the translator for editing.

The TRANSLEARN project group has been made up of contributors from four countries: the Institute for Language and Speech Processing (Athens) and ILTEC (Lisbon), which are both organisations that combine academic and commercial aspects, the companies Knowledge S.A. (Patras, Greece) and SITE (Paris), and University College London. UCL has been concerned with the creation of alignments between corpora in different languages, and with the development of automated methods for this process.

Below, we present our latest observations on the process, concentrating particularly on the use of "special words" in texts, cognates, and simple use of tagging to detect both syntactic and semantic correspondences that can improve the quality of automated alignment. It appears also from our experiments to date that the corresponding methods are robust across a wide variety of (European) language types, and do not owe their success to any fortuitous feature of particular languages or language-pairs. It would certainly be interesting to investigate whether this remains true when at least one major Asian language is added to our set. We have not yet been able to explore that direction, because of the time required to find and preprocess (e.g. add markers for features such as ends of paragraphs, and perform tagging) suitable corpora, but we hope to do so in the future.

2. The Key Issues in Automated Alignment

The initial view of automated multilingual alignment was that it was likely to require non-trivial attention to quite deep linguistic features in texts: in other words, that it would pose a difficult problem. Thanks to work by Brown et al. (1991) and Gale and Church (1991) on parallel corpora in French and English, a very different view replaced it. These authors observed that a simple method relying on counting (sentences per paragraph, and/or characters per sentence) was remarkably effective in establishing alignments, especially if these counts were used to determine payoffs or qualities in a dynamic- programming algorithm that added some elasticity or robustness in local searches and adjustments on the way to alignments.

Although effective, dynamic programming based on simple counting does not give the kinds of "industrial strength" results that could be used routinely in the daily work of, for example, the administration of the European Union. By making use of simple features of text as well as the counts of characters and sentences, we have improved the performance of automated alignment (Campbell and Fang 1995) to a point where the situation may seem highly satisfactory. It is not unusual to achieve better than 99% accuracy in the computations.

There is nevertheless a penalty for this success. The EU processes extremely large volumes of text, and even 0.5% to 1% of parallel multilingual corpus material that is wrongly aligned adds up to a large absolute number of misalignments. Moreover, the large volumes require computationally very efficient methods of establishing the alignments. The key issues of alignment, in our perspective, are therefore:

* development of methods that are efficient for large amounts of data;
* development of adequately efficient methods for automated detection of possible misalignments that the first types of methods have produced.

It is unrealistic to expect that we can trust all of the alignments that follow from use of an automatic method. The outputs of such methods will need to be inspected for accuracy, if they are to be used in a large-scale practical exercise in support of human translators' activities. But which parts of an automatically-aligned pair of corpora are correct, and which parts should be corrected? In principle there is no way of telling, short of human checking of all of the alignments. But this loses many of the gains of automation. Is there some better way to deal with the checking?

This reasoning lies behind the second "key issue" above. An automated checker can at least flag possible misalignments for attention by a human inspector, and thus reduce the amount of work involved in the final stage of polishing an inter-corpus alignment before any large-scale practical applications of the corpora.

There is one likely feature of the applications that may work in our favour. We have studied informally the behaviour of a translator using automatically-aligned corpora, and observed this feature in action. If an alignment error means that a retrieved block of text in a target language is not appropriate for editing to produce a translation of a given block of source-language text, a translator using a terminal can select a "discard" option and ask for the next-best apparent alignment to be displayed. If it is clear that this process will not produce something suitable in the target language, the translator can abandon the attempt to retrieve material and revert to manual translation of the source block. If the frequency of occurrence of this behaviour in a computing system is low enough, the translator will accept it because it is not seen as distracting and because the frequency of hits will ensure that the overall exercise of translation goes more smoothly and efficiently than if it had been purely manual.

Thus, it is unnecessary to aim at perfect alignment. It remains to be determined what is a good trade-off position between the amount of checking that a human inspector is asked to perform on possible misalignments that a previous checking program has flagged for attention, the quality (e.g. as measured by information-retrieval criteria such as completeness, accuracy and relevance) of the flagged items, and the frequency of occurrence of mistaken retrievals when a translator is looking for target-language texts to assist in translation of a fragment in a source language. But this is out of the scope of the present paper, as it is first a matter of experimental psychology.

3. Special Words

Some words, by themselves, are reliable indicators of position and are therefore good anchors for alignments. For our EU corpora, the best candidates are usually words with some specialised meaning in the relevant documents. Also (related to this consideration for the EU corpora, but a good general condition in our practice with other documents, e.g technical manuals, promotional literature in the travel industry), a strong condition for reliability for corpora in languages X and Y is that a candidate word W in X should be represented by one word in a corresponding dictionary of Y, and that word should have a unique translation to W in X. In a weaker form, it would be acceptable for the dictionary entries to be non-unique, provided that a specialist in the material being examined could certify that the X-Y and Y-X translations for a candidate special word in this material should be unique.

A further property of candidate special words is that they should not be so rare that they occur only once or twice in a document, because then the chance that they would ever occur in a misaligned portion of the text would be small. On the other hand, the text should not use them so frequently that they would be likely to occur in most of the sentences near the starting-point of a misalignment.

At present the software that we have produced for TRANSLEARN simply accepts nominations from a user (the person who is overseeing the creation of the alignments, or someone who can advise this person on the significance of the material that makes up the corpora) for special words, and later looks for these when checking alignments that have been established by the basic dynamic-programming method. We have found that specialists in the subject-matter of particular corpora are rather economical in their suggestions of possible special words, i.e the relatively few words that they nominate do not turn up often in or near misalignments. Therefore it is desirable to supplement a specialist's recommendations with some automatic means of nominating special words. The nominations should be subject to the "unique translation" and "not too frequent, not too infrequent" heuristics that we have mentioned.

On the question of "How frequent?", there is a good informal argument for choosing special words that occur, on a rough average, once in every e (the base of natural logarithms: about 2.7) sentences. In looking for local clues to possible misalignments, one should have a fixed search horizon (not too large, for reasons of efficiency), and avoid backtracking to reconsider clues (again to promote efficiency) once they have been used. This is analogous to the problem of finding the best example in a unidirectional search of a known number of examples, where of course the finding of the best example is not guaranteed because one may have passed over it in the hope of finding something better. The technique that maximises the chance of finding the best of N items is, except for very small N, to look at but not choose the first N/e items, to obtain the best rating R in that set, and then to choose the first subsequent item with a rating that is at least R (or to take the last item if nothing fits this criterion). The analogy can be stretched far enough to suggest the "once in e sentences" proposition for good special words.

What is interesting about our experiments, as summarised in Table 1, is that they do not confirm the hypothesis but that they support a value not far from e.

When we have examined samples of text manually for good candidates for special words, the words chosen are overwhelmingly nouns and verbs. (Adverbs might figure prominently in certain types of literary texts, e.g. poems, but translators who deal with such texts are unlikely to want any computer-based help).

The method of selecting candidate special words automatically should be simple and computationally efficient. The simplest approach is to select the longest word in the region where a search for candidates is carried out. This ignores the possibility of prefixes and suffixes surrounding a root that can be checked for the unique translation property in a dictionary. However, if we discard any long selected word that does not have an immediate dictionary entry (which usually rules out verbs), the method quickly arrives at nouns that satisfy the criterion, when no language where specialised nouns are constructed routinely by compounding is involved. For example, it is satisfactory for English-French, but not for either one when paired with German.

The approach above is the best that we have been able to construct when we are using no explicit syntactic or semantic information. If tagged text is available, selection of candidates that are nouns or verbs is somewhat faster, but the main advantage over the simple approach is that suitable short words are found also, i.e. the set from which candidates are drawn is larger and better. We would expect large corpora in a genuine "applied" setting to have been tagged, but in some areas (and some languages, in the present state of the art in automated parsing) no suitable tagger may be available. The method that relies on lengths of words is then a possible substitute, except for languages like German where compound noun construction is common.

Table 1 summarises results from an experiment on parallel corpora of mixed material: technical manuals, notes accompanying compact discs of classical music, promotional texts from the travel industry, and abstracts of scientific papers. The English corpus contained 1260 sentences. Tags for text in each language were added manually to indicate nouns and verbs, so that the method mentioned immediately above could be tested. Alignments were created by the dynamic-programming method based on sentence (at paragraph level) and character (at sentence level) counts. These were then inspected for misalignments, and misalignments were added where the initial quality of the alignment was too good to offer any demanding test for methods of automatic detection of possible misalignments. The starting-point for each test was a pair of corpora with 95% of correct alignments. For each column with a language-pair heading, the numbers refer first to the percentage of alignments that would be correct following correction of rightly-detected misalignments, and second (following a minus sign) to the effect of spurious detection of possible misalignments. "Frequency" indicates that special words (except those provided by a user) nominated by the programs have been ruled out unless they occur, on the average, within 5% of once in n sentences, where n is the number in the column. A - in the column indicates that this criterion has not been used. In the "Method" column, we use U to indicate that the user has supplied the special words, L to indicate the length-based identification mentioned above, and S to indicate the syntactic (tag-based) method of identification.

Frequency	Method	English-French	English-Czech	French-Czech
-	U	95.7-0.2	95.6-0.2	95.6-0.2
-	L	95.5-0.3	95.5-0.2	95.4-0.2
-	S	95.6-0.3	95.5-0.2	95.5-0.2
-	U+L	95.9-0.3	95.8-0.2	95.8-0.2
-	U+S	95.9-0.3	95.9-0.2	95.8-0.2
2	L	95.3-0.2	95.3-0.3	95.4-0.3
2	U+L	95.9-0.4	95.8-0.3	95.8-0.3
4	L	95.7-0.2	95.7-0.2	95.8-0.1
4	U+L	96.2-0.2	96.1-0.2	96.1-0.2
6	L	95.3-0.1	95.3-0.0	95.1-0.1

Table 1: Percentages of correct alignments after treatment of a basic 95% alignment by special-word methods

Table 1 is an extract from a longer set of data. The entries shown there are chosen to indicate the main trends of the experiment. Firstly, when the Frequency was varied, the best values peaked quite sharply near 4: if e is involved, one could say that this was about 3e/2 rather than the value e that was suggested by naive reasoning. Values significantly below or above 4 were less effective, and in some instances they led to worse results (less correct identifications and more spurious ones) than when the Frequency criterion was not used at all. Secondly, user choice of special words, supplemented by automatic selection of additional special words, was better than any one method used in isolation. An automatic method that accessed simple syntactic information explicitly was better than one that relied only on word lengths - but not by a large amount. Thirdly, each method, including the one that took special words only from the user, was prone to indicate that alignments that were in fact correct were possible misalignments.

The behaviour we have reported for 3 European languages in Table 1 was repeated for other European languages in tests, with the German exception mentioned above. Swedish, which might have been expected to show some "German" behaviour, in fact performed in the tests in essentially the same way as English.

4. Cognates

In European languages, partly because of a common influence of classical Latin and Greek roots, root structure of many words that are translations of each other share the same sequence of characters. A consequent index that can be used in considering alignment of a sentence x in X with a sentence y in Y is the number of words of x whose first n characters occur in a word in y. We call the words that have this correspondence "cognates". We have conducted detailed experiments on correct and useful cognate detection in many pairs of languages to determine whether the hypothesis has any value, and (if it has) to select a best value of n. In the present paper

there is no space to give detailed results, but we have found for each pair that we have tried that cognates are indeed useful, if (and almost only if) n=4. In other words, the choice n=4 is significantly more likely to lead to a correct identification of real cognates in aligned pairs of language corpora than any other choice of n. The situation is least clear-cut for languages that encourage the use of compounds (e.g. Finnish; also German, but the effect was not as marked there as we had expected).

In Table 2, we give an indication of how the experiments on cognate length behaved, for the English-French pair of EU corpora that were prepared for TRANSLEARN. This Table displays the cognate key length n along one axis, and the percentage of text blocks (basically paragraphs) containing at least 1 and at most b automatically-proposed cognates of the relevant length n, in its body. A user can decide on a value of b in actual use: informally, we want to avoid values that are either too small or too large to give good discriminatory power. What we did not expect in advance, and what Table 2 shows, is that this decision is not influenced by n: whatever the choice of b, n=4 gives the best "raw material".

b	n				
-	3	4	5	6	7
2	35.6	44.7	37.2	37.7	31.5
3	48.0	54.0	45.9	45.9	36.9
4	54.4	59.3	51.9	51.9	40.8
5	59.2	65.1	58.0	57.3	43.0
6	63.5	68.9	61.9	60.5	43.9
7	67.2	72.2	64.7	63.0	49.5

Table 2: Percentages of text blocks containing automatically-proposed cognates of key (initial) length n

In our 1995 work we observed that cognates were significant, but not as effective as special words in highlighting possible misalignments. By not searching merely for words whose first n characters matched, but applying a character-matching algorithm that attempted to skip prefixes and some compound material, we have been able to raise the level of effectiveness of the use of cognates to roughly that of special words. Moreover, for experiments of the kind leading to Table 1 that we have been able to repeat so far, with both cognate and special-word searches included, a fair summary of the behaviour is that the improvements shown in Table 1 are bettered by at least 50% while the percentage of spurious identifications goes up only slightly.

5. Use of Semantic Information

Our work in this area last year (Campbell and Fang 1995) was based on the idea that sentence pairs which are mutual translations have a constant minimal difference in terms of "semantic weight" expressed as the number of lexical items, assuming that running words can be divided into functional items (closed class words, e.g.

auxiliaries, pronouns and prepositions) and lexical items (open class words, e.g. nouns, verbs and adjectives). The semantic weight was thus conveniently calculated as the difference between the numbers of running words and of functional items. We tested the method on our full corpora of 49 aligned EU documents in English and French. The results were promising: 42.1% of the sentence pairs had no difference in terms of semantic weight expressed as the number of lexical items, and more than 93% of sentence pairs had a difference of fewer than 3 words.

However, there has been doubt whether semantic weight calculated this way has an even performance with other languages that make heavy use of noun compounding. German, for instance, collates all the items in its compound nouns into one solid single item, while in English they are kept as separate; for instance, Bundesbank vs. federal bank. If this doubt is justified on a general scale, then the use of semantic weight may mean simply a methodology that is language-specific. Instead of conducting investigations in other language pairs to find out whether this is true, which might require experimenting with every single EU official language, we have found it worth attempting to discover some other measure that is less language-dependent. Our most recent work suggests that "propositional content" may be such a candidate.

We may safely assume that the information structure can be divided into what is talked about - the referential property, and what is said about it - the predicational property (Croft 1984). These two properties combine to correspond to the concept called "proposition", which forms the core meaning of the sentence. The propositional content is held to be a constant under translation from one language to another, though the social context of any speech situation may radically influence the interpretation of a sentence (Lyons 1995; Jacobs 1995).

In linguistics, there are several distinct treatments for propositions. The most popular one is based on predicate logic, with predicates, (unlabelled) arguments and quantifiers. This is especially associated with truth-conditional semantics, but also used by others. Then there is a tradition that is exclusive to linguistics, which addresses situation-types with their associated (labelled) participants and circumstances. The labels on the participants distinguish actors, instruments and the like (Richard Hudson, personal communication). Though different, the various frameworks agree that propositions comprise a predicate and at least one argument, commonly expressed as P(x), where P is any predicate and x any argument. For our purposes, we are not concerned with the truth value or the meaningfulness of propositions. We are concerned with the structure or the linguistic realisation of propositions so that some structural component can be pinned down that indicates the existence of propositions as such.

By definition, predicates are realised through the use of [1] verbs, [2] noun phrases, [3] prepositional phrases, and [4] adjectival phrases. The inclusion of non-verbal predicates is mainly due to the understanding that the semantic nuclei in [2]-[4] are not expressed by any verbs, but by the complementing prepositional and adjectival phrases. Arguments, on the other hand, are mostly realised through the use of

[5] clauses, as in To be early is not difficult; and also via nouns.

Since clauses may be expanded into propositions themselves, we may conclude that propositions are realised through such grammatical properties as verbs, noun phrases, prepositional phrases, and adjectival phrases. Note that propositions defined as such have roughly the same grammatical properties as what is circumscribed by what we defined as semantic weight. To suit our practical needs, we propose that all complements to verbs may be treated as arguments despite their possible grammatical subcategorisations, so that appropriate examples of [1]-[4] may be taken as:

[1b] LOVES(John, Mary)

[2b] BE(John, a great lover)

[3b] BE(John, in love with Mary)

[4b] BE(John, very loving)

and [5] rewritten as

[5b] BE(BE(~,early), not difficult)

Thus we may assume that verbs always realise predicates, and it is well-accepted that propositions are semantic counterparts of clauses (Jacobs 1995). The number of arguments to a particular predicate is then decided by the valency of the verb.

We have thus isolated one grammatical property that is a definite component in the propositional structure, viz., the verb. It follows that by counting the number of verbs we may reliably establish the number of propositions. The number of propositions in a sentence can then be employed to test the hypothesis that sentence pairs that are translations of each other have a minimal difference in terms of verb counts.

The identification of predicates requires grammatically tagged material so that verbs are explicitly indicated as such. For this, we used tagged versions of documents in the EU corpora in Portuguese and English. In order to attach tagging information to the aligned version, we subsequently tagged it for the English material. Regarding Portuguese, we generated a list of verbs by automatically extracting all the word forms analysed as such from the original tagged version in Portuguese. We discovered that past participial verbs were treated differently for English and Portuguese, e.g.:

[6a] COUNCIL REGULATION EEC No 252/87 of 19 January .....
signed at Malabo on 15 June 1984.

[6b] REGULAMENTO CEE N 252/87 DO CONSELHO de 19 de Janeir .....
assinado em Malabo em 15 de Junho de 1984.

Signed in [6a] is treated as a past participle verb by the English tagger while its Portuguese equivalent assinado is analysed as an adjective by the Portuguese tagger. As a result, we found it necessary to include such adjectives in the verb list for the Portuguese material. Past participial verbs analysed as adjectives were then extracted and subsequently included in the verb list. The automatic flagging program counted as verbs any lexical item included in the verb list for Portuguese. For English, lexical verbs (excluding auxiliaries) as well as -ed adjectives were counted as verbs.

We can summarise our results as follows. Over 53% of all the sentence pairs had a zero difference in propositional content, which was calculated as the absolute difference between English and Portuguese verb counts. English sentences with a zero difference had a mean sentence length of 16.32 in terms of graphic words, while the

corresponding figure for Portuguese is 14.79. Sentences with zero difference had standard deviation 17.92 and 18.46 respectively for English and Portuguese. Within the same group, the shortest English sentence length was 4 and the longest was 157, while their Portuguese counterparts ranged from 1 word to 162 words. As in other experiments reported above, full tabular details require more space than is available for this paper, but are available on request.

Given that the statistics regarding propositional content revealed a comparable or greater central tendency than that of semantic weight, our investigation seemed to indicate that the measure of propositional content through verb counts could be a better discriminating factor than semantic weight. We nevertheless reserve any definitive conclusion at this stage as the results could be much more accurate if the English and Portuguese taggers produced analyses on a more uniform basis and if the aligned material incorporated explicit tagging information: statistics could be easily distorted when we tried to match items in the aligned material with those in the Portuguese verb list because of the problem of homographs. We plan to test this approach further given more favourable conditions. Before then, we conclude that propositional content is more reliable than semantic weight as a discriminating factor for the automatic detection of alignment errors. We have not yet been able to combine it with the other methods that we have described in previous sections, but we would expect the effect to be a further improvement over the results of the kind shown in Table 1 as augmented by the cognate technique described in Section 4.

References

Brown, P.F. et al. 1991. Aligning Sentences in Parallel Corpora. *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, 169-176.

Campbell, J.A. and A.C. Fang. 1995. Automated Alignment in Multilingual Corpora. *Proc. 10th Pacific Asia Conference on Language, Information and Computation (PACLIC10)*, 185-193. Hong Kong: Language Information Sciences Research Centre, City University.

Croft, W. 1984. Semantic and Pragmatic Correlates to Syntactic Categories. In D. Testen, V. Mishra and J. Drogo, eds., *Papers from the Parasession on Lexical Semantics*, 53-70. Chicago: Chicago Linguistic Society.

Gale, W.A. and K.W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, 177-184.

Jacobs, R. 1995. *English Syntax: A Grammar for English Language Professionals*, 9. New York: Oxford University Press.

Lyons, J. 1995. *Linguistic Semantics*, 141. Cambridge: Cambridge University Press.