AUTOMATED ALIGNMENT IN MULTILINGUAL CORPORA

J.A. Campbell

Dept. of Computer Science, University College London,
Gower Street, London WC1E 6BT, England
(jac@cs.ucl.ac.uk)

and

Alex Chengyu Fang

Survey of English Usage, University College London,
Gower Street, London WC1E 6BT, England
(ucleacf@ucl.ac.uk)

Abstract

Experiences in computing alignments at the paragraph and sentence level within a project TRANSLEARN in the European Union's "LRE" programme of research and development in language engineering are reported. About 98% of the sentences in pairs of corpora in different languages have been aligned correctly by a method that uses dynamic programming on numbers of characters per sentence. This parallels the experience of previous researchers for English-French alignment. We have used Portuguese and Greek material in addition to these languages, from a set of 49 European Union official documents. It is argued that the key issue of automated alignment is now the automated improvement of the quality of alignment achieved by methods that rely only on character counts. Cues that are helpful to support such an improvement are identified: special words, cognates, syntactic fragments, and a simple measure of semantic weight. A short account of their use in experiments is given.

Introduction and Motivation

There are many possible uses for multilingual corpora of parallel texts where alignments are indicated between corresponding parts of the textual material in different languages. In Europe, where the number of official languages of the European Union (EU) has recently risen to 11 and where documents recording formal acts of the EU have to be produced in all of them, there is no more pressing need than the need to give automated support to the human translators who are required to do this work.

The project TRANSLEARN within the EU's LRE programme of research and development in language engineering addresses an important part of the needs of EU translators. Because of the repetition of various structures and formal details in EU documents, it is quite possible that a translator who is making a version of a document in language Y from a source in language X will be generating text that is very close to something in Y that is already part of the EU's files or archives. Differences may be just in inessential details like dates, serial numbers, and names of institutions or countries. In this situation, the translator will clearly be more cost-effective if it is possible to find and display candidate pieces of text from the files that are

very close to the language-Y version of the current piece of text in X that the translator is processing. (Here, "piece" refers typically to a sentence or a paragraph). Rather than having to spend significant amounts of time on tackling a complete translation from X to Y, the translator should then be able to produce the desired material much more quickly by localised editing of the retrieved text in Y.

Alignment is always likely to be an important foundation-stone for any method that automates this form of assistance to translators. In the TRANSLEARN approach, the piece of text in X that is to be translated is first matched with a corpus of EU material in X which has also been aligned with a parallel corpus in Y, and the existing alignment is then followed to retrieve items in Y that can be offered as candidate translations. But in order for this method to be useful, high-quality alignments must first be produced for the parallel corpora.

We have been responsible for the "alignment" activity, and the production of alignment software, within the TRANSLEARN project.

Alignment has been the target of well-known studies in the past, e.g. by W. Gale's AT&T Bell group and in related work from the IBM Thomas J. Watson Research Center, on parallel text material in English and French. Initially alignment may have been regarded as a problem whose solution required use of quite deep semantic information, but those studies revealed that this was not so. Their lesson was that a very good level of performance could be achieved with the use of surface features only. In particular, it was possible to rely on counts of numbers of characters per sentence and numbers of characters and/or sentences per paragraph, and to make a robust one-pass method of scanning the texts, with the help of dynamic programming, to obtain what amounted to a best-fit match of such counts at the sentence level.

We have found the same behaviour during our work on TRANSLEARN, over a set of parallel corpora of EU documents in four languages (English, French, Greek, Portuguese) and over smaller sets of parallel texts involving English, French, German, Portuguese, Swedish and Czech taken from operators' manuals for electronic equipment or from translated textbooks.

The main lesson of our work, following on from the work of others, is that the key issue of automated alignment has now clearly changed from the initial question of "How, if at all, is it possible to make good alignments by computer- based methods?". The answer is that a conceptually rather pedestrian method does an extremely good job. But there is a penalty for success: potential users such as the EU with massive files of textual material will request that good results from projects such as TRANSLEARN be put into regular service for translators to exploit. A success rate of over 99% for a simple method of automated alignment at the laboratory level may be good news, but even just 1% of an EU-sized stock of parallel texts represents an embarrassingly large absolute number of sentences or paragraphs if these 1% of the items are misaligned. Correcting such misalignments at present would be a task for humans, whose time is expensive. Moreover, as they would not know a priori which 1% of the texts that they were about to scan happened to be misaligned, they would need to search all of the texts, checking alignments along the way, and therefore performing a task that would have the same amount of effort and tedium (except for the overhead of actually placing alignment indicators in position) as the job of completely manual alignment.

We believe (having tried at length to find such a way forward, without having any achievements that are worth reporting) that it is very unlikely that a new method of alignment different from the one we have mentioned above could produce higher-quality alignments - at least not without making computationally intensive use of semantic information and therefore being too inefficient for large-scale practical applications. In our opinion, the key issue in automated alignment should be "How can we automate an efficient process of scanning and

correction of the alignments produced by the basic counting method, so that the result is of a high enough quality to be valuable to users such as EU translators?". In this context, high quality means either that the percentage of incorrect alignments that remain should be low enough for further inspection and correction by humans to be feasible or (probably more likely) that the error rate for a "translator's assistant" system that displays candidate translations by accessing the stored alignments is not large enough for a translator to regard it as a nuisance.

This paper reports briefly on our experiences with the basic method, and then continues with an account of our work in response to that last question.

Initial Experience

The first focus of the dynamic-programming methods in past work by other groups has been alignment at the paragraph level. The initial objective of TRANSLEARN was thus to ensure acceptable alignment at the paragraph level. As this proved to be easier than expected to achieve, we moved quickly to the sentence level. That appears to be what has happened in previous work as well. Alignment at a finer scale than this is still a matter of research, but fortunately it is of much less interest to translators whom we have contacted than sentences and paragraphs, which are the most comfortable units of data (e.g. those that fit naturally into windows of manageable size and legibility on a bit-mapped display) from their point of view.

The basic test set for TRANSLEARN is composed of parallel corpora in four languages. Each corpus consists of 49 documents recording official acts, agreements and regulations of the EU. The documents are all first "cleaned", which involves removal of defects such as errors in typing and departures from the standard conventions that the EU uses for drafting of its documents. In the final phase of cleaning, marks to indicate ends of sentences and paragraphs are added.

TRANSLEARN has also applied taggers in each of its four languages to the relevant texts, to produce explicit syntactic information. This information is used in support of the phase of matching between different fragments of text in the same language, as referred to in the Introduction. The matching process also makes use of a counting approach with some similarities to the counting method that underlies automated alignment, but it has been developed separately and is outside the area of immediate interest considered in this paper. We have not supplemented our counting method for alignment with any reference to tags or syntactic information. However, as we state below, such information becomes useful when we are investigating how to automate an improvement of the alignments obtained via the counting method alone.

In our experiments with the EU corpus, we have found that slightly more than 98% of the alignments based on counting and dynamic programming were correct. For the material taken from manuals and textbooks, the quality often exceeded 99%, and in the worst case was about 96.5%. (These percentages are for sentence-level alignments; paragraph alignments have been 100% correct except for isolated cases where the TRANSLEARN conventions for text marking have produced paragraphs that are short single sentences and occasionally even single words).

We are continuing to collect data on misalignments so that we can create a taxonomy of problem cases and their relative frequencies, but it has been evident from the earliest days of this activity of collection that there are two significant sources of errors. First, there are situations where m sentences in one language have been translated by n sentences in the other, and where m = n does not hold. Second, we have seen that there is some tendency (not large in itself, but significant if we are considering the absolute number of misalignments in a very large

volume of text) for alignments to stray from what is correct if there is not much variation between the lengths of successive sentences over any extended passages of material. Of course, when the two conditions occur in the same place, the tendency to error is increased noticeably. Conversely, and fortunately, the dynamic programming that supports the counting method is robust in restoring correct alignment when a region of text with a running misalignment undergoes a noticeable change in the number of characters per sentence in either language. From the point where the change occurs, correct alignment resumes.

The remainder of the paper discusses techniques that we have seen to be useful in detecting possible misalignments.

Special Words

Some words, by themselves, are reliable indicators of position and are therefore good anchors for alignments. Usually they have some specialised meaning in the relevant documents, and (a necessary condition for reliability in our application) translations that are unique in practice in the context of the documents where they are used: a candidate word W in language X should be represented by one word in another language Y, and the unique representative of that word of Y, in X, should be W. On the one hand, they should not be so rare that they occur only once or twice in a document, because the chance that they would ever occur in a misaligned portion of the text would then be small. On the other hand, the text should not use them so frequently that they are highly likely to occur in most of the sentences near a point where a misalignment starts. We refer to such words as "special words".

The software that we have built for TRANSLEARN contains data structures to store information about special words and their detected positions in texts. At present the user is invited to nominate the special words in the languages of the translation to be considered. However, it is easy to add automated facilities, e.g. to search for nouns that occur in a text within an appropriate range of frequencies and to verify in a dictionary that they have the unique- translation property. There is an informal argument, for example, that a good special word should occur fairly regularly throughout a text on the basis of one or more occurrences in 1 out of e sentences, followed by no occurrences in the next e-1 sentences (where e is the base of natural logarithms, i.e. about 2.7).

If a text T is modified to ensure that a designated special word occurs according to this rule of thumb, our experiments when T is the material that we have drawn from manuals and textbooks show that about 30% of the misalignments are detected. More important, detection of the special word in a correct alignment does not disturb that alignment. The rate of detection is almost constant over different pairs of languages in our test sets. We preferred those sets to the EU corpora because of the greater variety of languages.

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 this short 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 further experiments we have observed that cognates are not as effective as special words in highlighting possible misalignments, but that their effect is significant: typically they give such a cue in more than 10% of the actual misalignments, but not as much as 15%.

Syntactic Fragments

As we have mentioned above, we can make one use of tagged corpora when these are available. In the EU corpora misaligned sentences quite often follow from the fact that translators do not always treat periods "." as invariant. A period in a text in X may become a colon, semicolon or dash in Y, and vice versa. When this happens, and also for some other effects that provoke misalignment, there will usually be some mismatch between the syntactic structures of the sentences that are misaligned. A common phenomenon in the EU material is a mismatch between the numbers of subjects, objects and/or verbs in the two items - which is computationally easy to detect. We are continually collecting more examples of syntactic mismatches; it is too early to report further firm conclusions. In the EU case, the mismatch that we have just mentioned identifies as many as 38% of the misalignments in some texts, though there is considerable variability.

If we take the idea of computations on fragments of syntax much further, what constitutes a match or mismatch shades off into semantic issues. We have therefore also considered a line of research that starts from semantics as such, in the next section.

Lexical versus Functional Items

We can motivate the semantic line of research by saying that we should look for a property not dependent on character counts and that remains relatively stable despite inter-language differences.

One such property could be a measurement of the semantic weight which is mostly carried by the lexical items as distinct from function items in the input sentence. Function items are words that can be attributed to closed classes whoses members can be listed exhaustively. They are typically personal pronouns, determiners, conjunctions, conjunctive adverbs, prepositions, auxiliaries, and non-"ly" adverbs. All the other items are generally classified as lexical items, such as nouns, adjectives, and verbs. Consider the following pair of sentences:

[a] Slavish imitation of models is nowhere implied.

[b] It is not implied anywhere that there are models which should be slavishly imitated.

It is apparent that though different in many aspects, the two sentences are still the same in terms of their linguistic meanings, which are mostly conveyed by the lexical items:

imitation implied models slavish

Function items, on the other hand, mainly contribute to the different sentence structures and lengths:

anywhere are be is it not
nowhere of should that there which

That the sentence comprises lexical and grammatical aspects 'forms a persistent part of grammatical theory' and has been discussed extensively by grammarians since the time of the Stoics (cf. [1]). Ure [2] used this characteristic in her study of registers in English. She introduced the notion of lexical density, defined as the ratio of the number of lexical items and the number of graphic words, and reported that this ratio demonstrates a constant difference between, for instance, speech and writing. Lexical density has also been extended to areas such as information retrieval and authorship attribution [3].

We assume that any sentence pair expressing the same meaning has the same amount of information load. Just as the actual entropy and redundancy combine to convey information, lexical words and function items also combine to express the linguistic meaning of a sentence. Our hypothesis is that the difference between sentence pairs in terms of lexical content is smaller than the difference between sentence pairs in terms of number of characters and words; sentence-length variations between sentence pairs are mainly due to the different number of grammatical items. To test whether this intuition could be confirmed through real data, we have conducted two experiments. The first was aimed at confirming our hypothesis while the latter was intended to determine the degree of difference.

In the first experiment, we examined mainly four properties, namely graphic word, lexical item, grammatical item, and noun counts. Only one aligned text was used, as we had to mark all the noun information manually, which was rather time-consuming. This one text comprised 95 sentence pairs, with 3476 words for English and 3696 words for French, consistent with the ratio we have seen in the larger English and French EU corpora.

Variable	Mean	Std Dev	Minimum	Maximum	Sum	N
English	36.59	25.14	2	102	3476.00	95
French	38.91	26.17	2	123	3696.00	95

Table 1: Mean English and French sentence lengths

The distribution of lexical items, however, shows a much smaller range of variation:

Variable	Mean	Std Dev	Minimum	Maximum	N
English	19.67	13.27	2	65	95
French	20.86	13.89	2	69	95

Table 2: Mean English and French sentence lengths in terms of lexical items

The absolute differences between English and French in terms of the four properties were subsequently calculated:

Variable	Mean	Std Dev	Minimum	Maximum	N
Lexical Items	1.93	1.82	.0000	8.0000	95
Nouns	1.95	1.79	.0000	7.0000	95
Function Items	3.17	3.26	.0000	17.0000	95
Graphic Words	4.19	4.44	.0000	21.0000	95

Table 3: Mean differences between English and French of the four properties

Among the four properties examined, graphic words had the greatest difference (mean = 4.19) while lexical items revealed the smallest (mean = 1.93). Standard deviations were calculated to show the degree of consistency in difference. Differences in nouns had the smallest deviation (1.79), suggesting that though greater in terms of mean, the difference in nouns is less liable to change than that of lexical items (1.82). In comparison, lexical items revealed a smaller mean but a greater deviation. But on a general scale, it may be concluded that nouns and lexical items have the same range of variation. Graphic words and grammatical items, on the other hand, both demonstrate a substantially greater deviation.

These results confirm our hypothesis that the difference between sentence pairs in terms of lexical content is smaller than the difference between sentence pairs in terms of number of characters and words; sentence-length variations between sentence pairs are mainly due to the different number of grammatical items.

The main point of the second experiment is to determine the range within which differences in content items between sentence pairs may vary. This experiment involved the use of 49 aligned texts with 3027 sentence pairs, which represented 62945 words for English and 66181 for French. Results further confirmed that lexical items are less liable to variation than both function items and graphic words:

Variable	Mean	Std Dev	Minimum	Maximum	N
Lexical items	1.16	1.49	0	24	3027
Function items	2.05	2.48	0	28	3027
Graphic words	2.54	3.20	0	52	3027

Table 4: Mean differences between English and French in terms of the three properties

A frequency profile of the distribution of the differences in lexical items between sentence pairs revealed a surprisingly small number of differences; out of 3027 sentence pairs, there were only 13 modes of difference:

Value	Frequency	Percent	Percent
0	1275	42.1	42.1
1	842	27.8	69.9
2	483	16.0	85.9
3	219	7.2	93.1
4	106	3.5	96.6
5	50	1.7	98.3
6	29	1.0	99.2
7	11	.4	99.6
8	6	.2	99.8
9	3	.1	99.9
13	1	.0	99.9
14	1	.0	100.0
24	1	.0	100.0
Total	3027	100.0	100.0

Table 5: A distribution profile of the lexical content difference between English and French

According to the distribution profile, 40.2% of the sentence pairs did not have any difference in terms of the number of content items. More relevant to our purpose, it can be observed that more than 99% of the sentence pairs had a difference smaller than 6 content words. It is thus possible to adjust the threshold for the judgement of a misaligned sentence pair according to a sliding scale. Such a scale, as represented by Table 5, is nevertheless rather crude. We subsequently classified the 3027 sentences according to the English sentence length in terms of graphic words and then calculated the mean difference between English and French for their lexical content. The results are summarised in Table 6:

Sentence length	Number	Mode	Mean	SD
1 - 10	1082	5	.330	.650
11 - 20	636	9	1.181	1.152
21 - 30	544	8	1.496	1.350
31 - 40	368	7	1.666	1.331
41 - 50	211	9	2.194	1.871
51 - 60	107	10	2.729	2.877
61 - 70	26	7	1.923	1.695
71 - 80	29	9	3.241	3.043
81 - 90	11	6	2.727	1.679
91 - 100	7	4	2.571	3.101
101 - 110	5	4	6.200	4.658
111 >	1	1	3.000	-

Table 6: A summary of groups of sentence lengths

From Table 6, we learn that the majority of the sentences are fewer than 10 words in length (1082), that for sentence pairs of up to ten words, the mean difference in terms of lexical content between English and French is less than 1 (.330), and that there are only five different types of differences, with a standard deviation (SD) of only .650. We found it significant that sentences of up to 70 words in length has a mean lexical- content difference of fewer than 3 words. Frequency-distribution profiles for each group of sentence lengths were also generated (see Appendix A) and set variable thresholds of 3 for sentences of up to 10 words, then 5 up to 20 words, then 6 up to 40, and finally 8 for more than 40 words.

After this preparation, we used the 49 texts as test data and flagged the 3027 sentence pairs according to the thresholds. Our program came up with 19 pairs that were misaligned. By comparison with the individual techniques stated in previous sections, this result is not as good as for detection of special words in texts where they are favourably distributed (which is in any case not likely to happen for all types of text), but otherwise it is the best technique that we have examined.

Conclusion and Acknowledgements

The experiments that we report here are intended mainly to establish the feasibility and likely relative value of four techniques of automated detection of misalignments in multilingual texts aligned by a dynamic-programming method that takes account of only the crudest surface features of natural language. While a substantial further programme of experiments to compare them more thoroughly and to suggest how best to combine them in efficient software is desirable, the generally positive character of the results above imply that it is reasonable to aim for the goal of automated alignment that is useful as a component of computer-based support for translators in large-scale practical applications.

We are grateful for the collaboration of the partner organisations in the TRANSLEARN project (ILSP, Athens; Knowledge S.A., Patras; SITE, Paris; ILTEC, Lisbon), and particularly Mr. Stelios Piperidis. We are pleased to thank Dr. Niladri Chatterjee, Dr. Mauro Manela and Mr. Neil Morgenstern for their help and comments at UCL, and we acknowledge the value of the financial support from the European Union through its LRE programme.

References

[1] G Carlson & M Tanenhaus. Lexical Meanings, Structural Meanings and Concepts. In D Testen, V Mishra & J Drogo. *Papers from the Parasession on Lexical Semantics*, 39-52. Chicago: Chicago Linguistic Society, 1984.

[2] J Ure. Lexical Density and Register Differentiation. In G E Perren & J L M Trim. *Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics, Cambridge 1969*, 443-452. Cambridge: Cambridge University Press, 1971.

[3] M Dras. Automatic Identification of Support Verbs: a Step Towards a Definition of Semantic Weight. In *Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence*, 451-458. Singapore: World Scientific, 1995.