Some Technical Aspects about Aligning Near Languages

Lluís de Yzaguirre, Marta Ribas, Jordi Vivaldi, M. Teresa Cabré
Institute for Applied Linguistic. Universitat Pompeu Fabra
La Rambla, 30-32. 08002, Barcelona, Spain
de_yza@upf.es

Abstract

IULA at UPF has developed an aligner that benefits from corpus processing results to produce an accurate and robust alignment, even with noisy parallel corpora. It compares lemmata and part-of-speech tags of analysed texts but it has two main characteristics. First, apparently it only works for near languages and second it requires morphological taggers for the compared languages. These two characteristics prevent this technique from being used for any pair of languages. Whevener it its applicable, a high quality of results is achieved.

1 Introduction

This paper presents some technical aspects of an automatic text aligning system developed at the Institute for Applied Linguistics (IULA) at the Universitat Pompeu Fabra. This programme differs from previous ones in that it introduces a particular strategy consisting in the use of added linguistic information (lemmata and morphological tags). This means that the programme is highly dependent on the availability of linguistic mark-up tools for the languages involved in the alignment. However, it overcomes part of the noise produced by many parallel tasks, which is not solved by programmes based only on statistical criteria.

2 Programme overview

This programme comprises 3 levels: *a)* word level, *b)* sentence level, and *c)* overall text level. In the word level the similarity degree of two words is measured; in the sentence level results from the word level are extended to each pair of sentences. Finally in the text level a strategy is designed so as to decide which sentence from language A is to be compared with its counterpart from language B. This is done by taking into account classic models of sentence-length calculation and text location.

The aligner is started by reading two files: the source text (written in language A) and its corresponding translation (written in language B). These two files have been previously provided with both structural and linguistic mark-up. Alternatively, the programme can also read a dictionary of prealigned words as will be explained below.

Structural mark-up helps to show the text syntactic boundaries[1]. Thus, it avoids syntactic ambiguity that can be generated by marks like "." (full stop) since, depending on its position, it can lead to different functions such as final sentence mark or abbreviation boundary.

Users can make a revision of this dictionary of prealigned words, discard words that are not considered relevant and maintain those ones concerning the specialised domain involved (see 4 below).

The validated dictionary of prealigned words is useful to provide the aligner with feedback. This will yield to reach better results in further stages regarding the same texts as well as other texts from the same specialised domain.

3 Detailed functioning of the programme

3.1 Word level

3.1.1 Word Similarity Coefficient (WSC)

This coefficient is established according to the similarity of lemmata and morphological tags, together with a measure of word formal similarity (see GSC) and the distance measure between two words within the text.

Lemmata evaluation

```
if (both lemmata are equal) then
        {WSC is value1}
else if (SIMEX of lemma1 is equal to SIMEX of lemma2) then
        {WSC is value2}
else if (both lemmata are equal minus one letter) then
        {WSC is value3}
else
        {WSC is zero}
end if
```

Part-of-speech evaluation

```
if (part of speech of word1 is equal to part of speech of word2)
then
        {WSC is increased with value4}
else if (main morphological cathegory of both words is equal)
then
        {WSC is increased with value5}
end if
```

[1] See Vivaldi *et alii* (1996)

Other:

$WSC = WSC + GSC(L_1,L_2) - abs(p_1-p_2)$

where L_1 and L_2 are both lemmata and p_1 and p_2 are the position of each word in its sentence.

Therefore the more similar lemmata and grammatical categories of compared words are the higher this coefficient is. Both SIMEX comparison and GSC enforce the capacity of the compared lemmata.

3.1.2. Graphematic Similarity Coefficient (GSC)

This coefficient estimates the similarity for words without knowledge about them:

$$GSC(L_1,L_2) = a+b+c-abs(d-e)$$

where

a	number of identical letters in same position from left
b	number of identical letters in same position from right
c	number of identical letters in any position
d	length of $lemma_1$
e	length of $lemma_2$

3.1.3. The SIMEX technique

This technique aims at reducing random differences that are found among two words because of the orthographical peculiarities of each language, such as the English "ph" and the Spanish "f".

Contrary to the strategy of graphematic measure, which is language-independent, SIMEX is configured taking into account the specific features of each pair of languages that are to be aligned.

SIMEX shows word similarities through reductions which lead to the correction of orthographic differences (e.g. "*à*" --- "*á*") as well as morphemic differences (e.g. "*-idad*" --- "*-itat*"). These reductions are placed in an independent file which can be externally updated.

The following words account for the SIMEX behaviour. The first column from each row comprises the source Catalan word followed by the SIMEX reduction; the second column from each row comprises the Spanish word followed by the SIMEX reduction. It can be seen that words are different but SIMEX are equal.

rei->rey	rey->rey
autoritat->autoritat	autoridad->autoritat
improcedent->improcedent	improcedente>improcedent
recurs->recurs	recurso->recurs
revisió->rebisio	revisión->rebisio
processal->procesal	procesal->procesal
absolució->absolucio	absolución->absolucio
sentència->sentencia	sentencia->sentencia
demandat->demandat	demandado->demandat
dia->dia	día->dia
recurrent->recurrent	recurrente->recurrent
sol·licitar->solicitar	solicitar->solicitar

3.2. Sentence level

3.2.1. Word comparison within the sentence

The analysis of the similarity degree between two sentences comprises the following steps:

a) Each sentence from language A and language B is splitted word by word and both lists of words (including orthographic marks) are ordered in accordance with their appearance order throughout the text. Grammatical words (e.g. prepositions) are deleted from the list.

b) Each element from list A is compared with all elements from list B at a word level (see **3.1**). Then a contrastive table is created with its WSC. Here it should be noted that words are only compared with words, and the same holds for punctuation marks.

c) Starting with all figures found in the constrastive table regarding each word from language A with each word from language B, the highest figure of each column and each row is selected and the remainig figures are zeroed.

d) Finally, it is calculated the median for the figure of lexical pairs that have been related each other.

3.2.2. Contrastive analysis

The analysis of the robustness of similarities according to the robustness of adjunct lexical pairs is carried out through the analysis of the three preceding and following lexical pairs.

The WSC for each pair is increased for each of the three preceding and following lexical pairs considered robust in the established coefficient (see **3.1.** above).

So, when a word similarity figure is of 0 in the strict lexical comparison and it is preceded or followed by three more lexical pairs with high similarity figures, then this word is given a coefficient that is roughly the same as having both the SIMEX lemmata and similar lexical categories.

Together with point addition on the grounds of contigous reliable lexical pairs, points are also added to each lexical pair already documented in the dictionary of prealigned words, provided that the dictionary is being processed. Here the main goal is to foster the relevance of all lexical pairs that were already documented in the dictionary.

3.2.3. Overall sentence figures

Once the similarity figure for each word has been determined, it is calculated *a)* the coefficient median of each lexical pair and *b)* the robustness of each sentence pair.

Thus once the coefficient median for each lexical pair has been obtained, we measure the number of lexical pairs whose figure is higher than a pre-established figure at which the robustness of a lexical pair is estimated.

If when making the lexical comparison an equivalency dictionary is being used, then all lexical pairs encountered in the dictionary are automatically considered reliable.

The programme tuning has yielded to make small changes in the sentence overall coefficients, according to the evaluation of the programme results. In the future we will have to decide whether this tuning will have to be specifically adapted to each language pair that is compared.

3.3. Text level

In order to decide which sentence from language A is to be compared to its counterpart from language B the programme follows the steps below:

a) Establishment of three figures for each sentence:

- number of sentence characters
- number of sentence words
- absolute character distance from the beginning of the text

b) Establishment of a tolerance boundary for each figure. Tolerance boundaries will be restrictively applied to the first tests so as to compare the longest sentences and the most similar ones regarding length and position within text. In successive stages of the process the programme's tolerance boundary is broadened and at this stage all sentences devoid of counterpart are compared, even those that are in very different positions in their corresponding sentences.

c) Whenever two sentences have their figures within the tolerance boundary, they are compared at a sentence level (**3.2**). If the global sentence figures (**3.2.3**) are satisfactory, the two sentences are paired.

d) When a sentence from language A is compared at two different times with two different sentences from language B, the programme will maintain the pairing with the highest overall sentence figures.

e) Once a number of processes have been carried out with a progressive lessening of their tolerance boundaries the programme will attempt to sort out all cases that remain unpaired by grouping them into the preceding or following sentences. In other words, it will compare at a sentence level two sentences from language A with one sentence from language B, or the other way round.

f) Finally, the programme will left unpaired sentences that will not have been solved over the preceding stages.

4. The building of the dictionary of prealigned words

4.1. An entirely automatic processing

Once a number of texts have been processed without making use of any dictionary of prealigned words, a file containing lexical pairs is obtained. Then, all files from the same specialised domain are compiled in a database.

This database contain several information fields: *a)* language A lemma, *b)* language B lemma, *c)* part-of-speech tag of lemma A, *d)* part-of-speech tag of lemma B, *e)* similarity coefficient for the lexical pair, and *f)* sentence coefficient in which the lexical pair has been found.

Then the whole file is revised by looking for those lexical pairs that are repeated in a row and all repetitions of this lexical pair are counted. At this point its coefficient mean is calculated for each occurrence. Finally, lexical pairs repeated at least X times are exported to a new database (which will constitute the dictionary of prealigned words). It also holds for lexical pairs whose coefficient mean is higher than Y (both X and Y can be tuned depending on the domain or on quality constraints). Alternatively, a combination of these two possibilities is also allowed.

4.2. Semiautomatic processing

This processing is identical to the previous one except that the last stage of direct export is omitted. Instead, all registers that are likely to be exported are marked. Then, there is a human revision of all lexical pairs that, if applicable, can be modified. Finally, all selected pairs are exported to the dictionary of prealigned words. Both in automatic and semiautomatic processing each lexical pair is exported once.

5. Preliminary assessment of the aligner

11 aligned texts from the domain of economy taken from the IULA's corpus were analysed so as to obtain a preliminary assessment of the aligner. Both Catalan and Spanish versions were considered, amounting to 40.000 words each version. Of the 1.604 alignments carried out, 1.551 turned out to be correct, 51 incorrect and 2 were not done. Thus, the results show a precision rate of 96,81% and a recall rate of 96,69%.

The aligner appears to be successful in those cases where 2 sentences of the same version correspond to a single sentence of its counterpart version. Similarly, it also succeeds in matching sequences whose structures are very different one from each other.

As for the incorrect alignments, errors can be divided as follows:

(i) *Errors due to differences in the two versions.* Differences in the writing are found, that is, changes in the sentence contents and deletions or adding of sentences in one of the versions. Besides structural differences such as different paragraph distribution and punctuation variation are also detected. It yields to use this programme to validate translations.

(ii) *Errors due to problems in previous stages.* Here errors can be put down to the scanner, the preprocess or the erroneous attribution of lemmata and grammatical categories.

(iii) *Alignment errors.* Sequences that are not parallel are aligned or clusters that are interchanged.

6. Future work

In the near future we will deal with the following topics:

a) Enlarging precision and recall assessment to other areas
b) Programme performance with regard to excerpts from other languages
c) Testing then system with other languages (French and English)
d) Comparing texts that are not fully disambiguated (here the programme would contribute to the disambiguation task)
e) Allowing for the use of prealigned phrases such as terminology from the domain, phraseology, idioms, etc.

As for web samples there is a web site [http://traductica.upf.es/alinea/] in which the functioning of the aligner is shown with some examples. Further information about the whole project will be available through this page.

7. References

BAKER, M. (1995) «Corpora in Translation Studies: An Overview and Some Suggestions for the Future Research» *Target* 7(2), 223-43.

BROWN, P.F.; LAI, J.C.; MERCER, R.L. (1991) «Aligning Sentences in Parallel Corpora» *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. University of California. Morriston, NJ.

CHANG, J.S.; CHEN, M.H. (1997) «An Alignment Method for Noisy Parallel Corpora based on Image Processing Techniques» *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (Madrid, UNED)*. San Francisco.

DE YZAGUIRRE, L.; MATAMALA, A; CABRÉ, T. (2000) «El lematitzador "PALIC" del IULA (UPF)» *Accepted paper at the XVII Congreso AESLA, section on Corpus Linguistics and Computational Linguistics*.

DE YZAGUIRRE, L.; MATAMALA, A.; BACH, C.; CASTILLO, N.; USTRELL, E. (2000) «AMBILIC, el desambiguador lingüístico del corpus del IULA (UPF)», *Accepted paper at the XVII Congreso AESLA, section on Corpus Linguistics and Computational Linguistics*.

DE YZAGUIRRE, LL.; MATAMALA, A. (in press) «El paquet LIC de lematització, categorització i desambiguació». Barcelona: Universitat Pompeu Fabra. *Institut Universitari de Lingüística Aplicada. Papers de l'IULA*. Sèrie Informes.

GALE, W.A.; CHURCH, K.W. (1991) «A Program for Aligning Sentences in Bilingual Corpora» *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. University of California. Morriston, NJ.

JOHANSSON, S.; OKSEFJELL, S. (eds.) (1998) *Corpora and Cross-Linguistic Research. Theory, Method and Case Studies*. Amsterdam / Atlanta (GA): Rodopi.

KENNY, D. (1998) Corpora in translation studies. In M. Baker & K. Malmkjaer (Eds). Routledge Encyclopedia of Translation Studies. London: Routledge.

MELAMED, D. (1997) «A Portable Algorithm for Mapping Bitext Correspondence» *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (Madrid, UNED)*. San Francisco.

Morel, J.; Torner, S.; Vivaldi, J.; De Yzaguirre, Ll.; Cabré, M.T. (1998). «El corpus de l'IULA: Etiquetaris». *Papers de l'IULA. Sèrie Informes,18*. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada., [Second edition, revised and amended].

VIVALDI, J.; DE YZAGUIRRE, LL.; SOLÉ, X.; CABRÉ, M.T. (1996) «Marcatge estructural i morfosintàctic del corpus tècnic amb l'estàndard SGML». *Papers de l'IULA, Sèrie Informes, 1*. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.