Hybrid Algorithm for Word-Level Alignment of Parallel Texts*

Eduardo Cendejas, Grettel Barceló, Alexander Gelbukh, and Grigori Sidorov

Center for Computing Research, National Polytechnic Institute, Mexico City, Mexico
{ecendejasa07,gbarceloa07}@sagitario.cic.ipn.mx
http://www.gelbukh.com, http://cic.ipn.mx/~sidorov

Abstract. Given a text in two languages, word alignment task consists of identifying in the two variants of the text specific word occurrences that are mutual translations. The majority of existing text alignment systems follow either a linguistic or a statistical approach. We argue for that both approaches are insufficient when used separately, and suggest a flexible algorithm that combines statistical and linguistic techniques.

1 Introduction

Given a bilingual corpus, the text alignment task establishes a correspondence between structures, in the two languages. There are two main alignment approaches: linguistic and statistical. Linguistic approaches use linguistic information, which implies its reliance on availability of resources. Statistical approaches are based on frequencies of occurrences, though they imply a lower computational cost but usually give lower precision.

Ideally, the units (words, sentences, paragraphs) of the two texts ought to be in direct one to one correspondence. However, the alignment task is complicated by many effects that break such an ideal model. There are many reasons for which word alignment is more difficult than alignment of others units, e.g.: the degree of inflectivity of the two languages, syntactic structure of the two languages, models employed for alignment and the available linguistic resources [1,2].

2 Alignment Algorithm

The proposed alignment algorithm combines statistical and linguistic approaches to reduce the disadvantages that both approaches present when used independently. The statistical stage of the proposed system relies on three different techniques: boolean modified K-Vec algorithm, modified K-Vec algorithm with frequencies and IBM model 2. The statistical processing includes:

1. Segmentation of the input texts. The original K-Vec algorithm allows the text to be divided into small pieces or segments. Our modification allows the pieces to be paragraphs, sentences, or a specific number of words (a window).

* Work done under partial support of Mexican Government (SIP-IPN 20091587 and 20090772, CONACYT 50206-H and 83270, SNI, PIFI-IPN).

H. Horacek et al. (Eds.): NLDB 2009, LNCS 5723, pp. 293–294, 2010.

2. Generating a list of words with associated vectors. A vector contains the occurrences (boolean values or the frequency) of the word in the segments.
3. Construction of a contingency table. The vector corresponding to each word in the source language is compared to all the vectors in the translation.
4. Calculation of similarity for each pair in each table. The similarity of words is determined by means of an association test (Pointwise Mutual Information, T-score, Log-likelihood ratio and Dice coefficient).
5. Selection of the word with the greatest level of association. The other candidates in the table are discarded as translations.

If the algorithm is used in a bidirectional way, the same process is carried out interchanging the languages. The linguistic processing incorporates: (1) dictionaries, to extract lexical information, (2) lexicons with morphological information, to compare lemmas and verify grammatical categories, (3) syntactic trees, for identification of its parts – subject, predicate, etc. – and to facilitate comparisons with its counterparts in the target language, (4) semantic domains, to establish semantic relations between the meanings of a word, (5) cognates, to align words that totally or partially coincide, and (6) learning: all alignment hypotheses that can be obtained will serve as a reference for future alignment tasks.

3 Preliminary Results

We used fragments (Spanish–English) from five novels. The following table shows the obtained results. We used a modified K-Vec procedure, and we applied the cognates during the linguistic processing to reinforce the alignments.

Aligner	precision (%)	generates a dictionary?	generates alignment file?
K-Vec++	35	Not for all words	No
Uplug	45	Not as a result	Yes
GIZA++	52	Yes	Yes
Proposed algorithm	**53**	Yes	Yes

4 Conclusions

Statistical techniques provide a good starting point for the alignment processes; however, incorporation of linguistic techniques improves the quality by involving intrinsic characteristics of the involved languages. The main disadvantage of linguistic processing is the need in the linguistic resources given their limited availability. In addition, this has an impact on the algorithm speed. Nonetheless, employing databases of optimization has proved to minimize this disadvantage.

References

1. Borin, L.: You'll take the high road and i'll take the low road: Using a third language to improve bilingual word alignment. In: ACL 2000, vol. 1, pp. 97–103 (2000)
2. Mihalca, R., Pedersen, T.: An evaluation exercise for word alignment. In: HLT-NAACL 2003 Workshop on Building and using parallel texts, vol. 3, pp. 1–10 (2003)