Combining clues for lexical level aligning using the Null hypothesis approach

Olivier KRAIF
LIDILEM, Université Stendhal Grenoble3
Grenoble, France, 38000
Olivier.Kraif@u-grenoble3.fr

Boxing CHEN
LIDILEM, Université Stendhal Grenoble3
Grenoble, France, 38000
Boxingchen@yahoo.com

Abstract

Various informations can be used to align parallel texts at word level: co-occurrence frequencies, position difference, part-of-speech, graphic resemblance, etc. This paper proposes a simple method to combine these clues in an efficient way. The association score is computed from the probabilities of pairing two units under Null hypothesis, assuming that the association is fortuitous. This approach has been applied to a literary English-French parallel text with good results.

1 Introduction

From the early 1990's, much interest has been given to the research on bilingual parallel text aligning. Aligning corpora at lexical level proves to be very useful for many applications such as bilingual Lexicography or Terminography, Statistic Machine Translation, cross language information retrieval (Brown, 2000), Computer Assisted Language Learning (Nerbonne, 2000), or even Word Sense Disambiguation (Ng, 2003).

To verify the latter hypothesis, the CARMEL Project aims at gathering literary texts with translations in 4 languages (French, English, Spanish and Italian), and implementing Word Sense Disambiguation and Thematic Identification methods, taking advantage of the multilingual context of each unit. We assume than given a text, each translation makes explicit additional information about its semantic and referential content. After the relatively easy task of sentence aligning, we are now implementing lexical level aligning techniques to establish word correspondences between the 4 languages.

Though considerable progress has been made in this field (Dunning 1993, Dagan et al., 1993, Melamed 1998, Tufis 2002), this task remains difficult. The 75% accuracy of the best system for the translation spotting task, in the last Arcade campaign (Langlais and Véronis, 2000), showed that there was room for improvement.

In the latest three years, Jin-Xia et al. (2000) have investigated a linguistic-knowledge-based word similarity to compute the association score of the word pairs, between Chinese and Korean. Linguistic knowledge was acquired from linguistic comparison of all layers between two languages: word formation, part-of-speech, lexical internal structure and syntax. Lopes and Mexia (2001) used GenLocalMax algorithm to extract typical contiguous and non-contiguous sequences of characters as cognates, and then used these cognates to extract the word pairs. Tiedemann (2003) proposed an algorithm to combine several clues for word aligning. These clues were probabilities, computed from similarity measure or learned from a word-aligned training corpus.

The method we present is somehow similar, because we also combine various clues to take advantage of all the available indices. But it does not need any word-aligned training corpus. In section 2, we describe the principle of the Null hypothesis approach. Section 3 and 4 are devoted to the experiments and evaluation of the results.

2 The "Null hypothesis" approach

2.1 Aligning algorithm

The general framework of our aligning method is very simple. Given two aligned sentences, an association score is computed for every possible pairing between units, then the best pairs are selected iteratively.

Let Cand be the set of candidate pairs, and Sel the set of selected pairs. At initialization Sel←∅

0. $\forall (u_i,u'_j) \in Cand$ compute $Score(u_i,u'_j)$.

1. Sort *Cand* elements in descending order of the association score.

2. the best scoring pair (u_s,u'_t) is removed from *Cand* and recorded in *Sel* :

$Cand \leftarrow Cand / \{(u_s,u'_t)\}$.

$Sel \leftarrow Sel \cup \{(u_s,u'_t)\}$.

3. All the competing candidates are removed.

$\forall (u_s,u'_j) \in Cand$, $Cand \leftarrow Cand / \{(u_s,u'_j)\}$.

$\forall (u_i,u'_t) \in Cand$, $Cand \leftarrow Cand / \{(u_i,u'_t)\}$.

4. Return to 2, until $Cand = \emptyset$

As demonstrated by Melamed (1998), this algorithm approximately establishes the best scoring set of correspondences under the one-to-one assumption. Moreover, it allows to reduce the

effect of *indirect association*: when two units are strongly linked on the syntagmatic axis, they tend to be associated with the same unit in the translated part. Because of the one-to-one assumption, units compete with each other to find an association, and the best scoring ones come before.

2.2 Null Hypothesis

The results of such a simple algorithm strongly depend on the association score. As we lack word-aligned training corpus, we cannot easily compute empirical distributions for all the interesting clues, in order to estimate the probability for two units to be translational equivalent. Thus, we just propose to evaluate the probability for two units to be non-equivalent.

We make the following assumption, namely the *Null hypothesis*: the co-occurrence of two units that are not translational equivalent is a *fortuitous* event (i.e. bearing no linguistic determination).

Of course, this assumption does not hold strictly, because it does not take into account the syntagmatic associations between words inside each language. For instance, between the two following sentences (extracted from Flaubert's *Madame Bovary*):

EN: The night was covered with stars, a warm wind blowing in the distance; the dogs were barking.

FR: le ciel était couvert d'étoiles, un vent chaud passait, au loin des chiens aboyaient.

it appears that the Null hypothesis is verified at various degrees: the co-occurrence of (*stars*, *aboyaient*) is fortuitous (from a linguistic point of view), but not the one of (*stars*, *étoiles*). The case of (*stars*, *ciel*) is in-between.

2.3 Association score computing

The probability to observe k independent clues C_1, $C_2 \ldots C_k$ under the null hypothesis at the same time is given by: $P_0 = \prod_{i=1}^{k} P_0(Ci)$

where $P_0(C_i)$ is the probability to observe the clue C_i under the null hypothesis.

The smaller this probability, the more unlikely the null hypothesis, and the more probable the assumption that units are mutual translation. Thus, the association score can be built as:

$$Score(u_1, u_2) = \sum_{i=1}^{k} -\log P_0(Ci) = \sum_{i=1}^{k} Si$$

We chose to use the following clues: word distributions, graphic resemblance, word positions, and word parts-of-speech. To compute an efficient association score, one needs to focus on features that allow to discriminate between fortuitous and non fortuitous correspondence. For each clue, the computing of probabilities is designed for the best discrimination:

2.3.1 Word distributions across text

The first association score (S_d) is based on word co-occurrence. Given two units (u_1,u_2), given their frequencies n_1 and n_2, it is possible to estimate the probability that they globally co-occur n_{12} times among n segment pairs[1], only by chance. We computed this probability assuming a binomial distribution. Without simplification, this probability can be expressed by:

$$P_0(n_{12} / n, n_1, n_2) = \frac{C_n^{n_1} C_{n_1}^{n_{12}} C_{n-n_1}^{n_2 - n_{12}}}{C_n^{n_1} C_n^{n_2}}$$

This probability is computed as the result of three independent draws, assuming that each unit occurs only once in the same segment pair:

$C_n^{n_1}$ is the number of different possible draws for the n_1 occurrences of u_1.

$C_{n_1}^{n_{12}}$ is the number of different possible draws for the n_{12} occurrences of u_2 that co-occur with u_1.

$C_{n-n_1}^{n_2 - n_{12}}$ is the number of different possible draws for the n_2-n_{12} occurrences of u_2 that don't co-occur with u_1.

$C_n^{n_1} C_n^{n_2}$ is the total number of possible draws without making any assumption on n_{12}.

2.3.2 Graphic resemblance

The association score based on cognate (S_{cog}) is the log-probability to observe superficial resemblance between two randomly drawn words inside an aligned segment pair. The event of cognateness is determined by counting the length of the Longest Common Sub-string (LCS). Two units are considered as potential cognates if the sub-string exceeds a certain proportion (here, 2/3) of the longest unit. The probability of cognateness P_{cog+} between two randomly drawn units has been computed from empirical observations on a automatically sentence aligned corpus The score is expressed by the following equation:

$$S_{cog}(u_1, u_2) = \begin{cases} -\log P_{cog+} & \text{if l(LCS)} \geq 2/3.\max(\text{l}(u_1,u_2)) \\ -\log P_{cog-} & \text{if l(LCS)} < 2/3.\max(\text{l}(u_1,u_2)) \end{cases}$$

[1] we call *segment* a group of aligned sentences

2.3.3 Word position

The association score based on word position (S_{posi}) is the log-probability to observe a small position difference between two randomly drawn words inside an aligned segment pair. The position difference is computed by:

$$D_{posi}(u_1, u_2) = \left| i - j \cdot \frac{l_s}{l_t} \right|$$

where i is the position of the source word, j is the position of the target word, l_s is the length of the source sentence, l_t is the length of the target sentence. Three cases are taken into account:

$$S_{posi}(u_1,u_2) = \begin{cases} -\log P_{posi1} \text{ if } D_{posi} <= 3 \\ -\log P_{posi} \text{ if } 3 < D_{posi} <= 5 \\ -\log P_{posi3} \text{ if } 5 < D_{posi} \end{cases}$$

These probabilities can be roughly estimated according to L the average length of the aligned segments (a segment is a group of aligned sentences).

$P_{case1} \approx 7/L \quad P_{case2} \approx 11/L \quad P_{case3} \approx 1-P_{case1}-P_{case2}$

where L is supposed to be higher then 11.

2.3.4 Word part-of-speech

The association score based on word part-of-speech (S_{pos}) is the log-probability to observe the same part-of-speech between two randomly drawn words inside an aligned segment pair. This probability P_{pos+} can be computed from empirical observations on any sentence aligned corpus.

$$S_{pos}(u_1,u_2) = \begin{cases} -\log P_{pos+} \text{ if POS are identical} \\ -\log P_{pos-} \text{ if POS are different} \end{cases}$$

2.3.5 Score combination

The distribution score S_d has a different meaning than S_{cog}, S_{posi} and S_{pos}, because it is not the result of a random draw of two units inside an aligned segment. So we propose to combine these scores with different weight:

$S(u_1,u_2) = S_d + k \cdot (S_{cog}+S_{posi}+S_{pos})$

3 Experiment

We implemented this method on Flaubert's novel *Madame Bovary* and its English translation. The corpus has been tokenized, lemmatized and POS tagged using XeLDA[2]. The parameters P_{cog+}, P_{posi1}, P_{posi2}, P_{posi3} and P_{pos+} have been directly computed from the aligned segments of BOVARY. S_d has been computed from the distributions inside a bigger corpus including other texts of the CARMEL corpus (see table 1).

Lang.	Word Occ.	Word Types	Sentences	Segments
BOVARY corpus				
English	139,030	9,387	8,873	6,663
French	139,968	8,373	6,879	6,663
Extended corpus				
English	389,000	18,168	17,312	13,705
French	382,102	16,456	14,052	13,705

Table 1: corpora description

3.1 Results

For evaluation, we created a small gold standard consisting of 149 English and French segment pairs, extracted from the first chapter of the BOVARY. The manual aligning yielded 1,156 content-word pairs. Results have been evaluated using a fine-grained metrics for precision and recall (Ahrenberg et al., 2000), and a balanced F-measure. For the competitive linking algorithm, only content words have been taken into account. Figure 1 displays the results for various values of k.

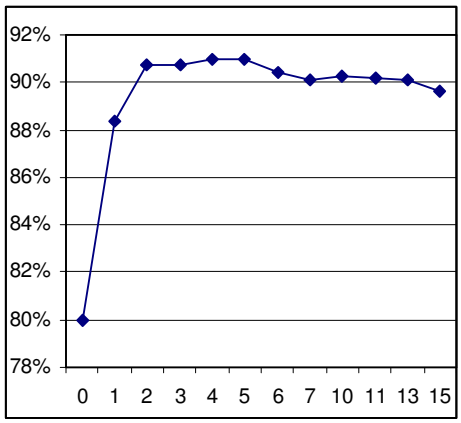

Figure 1: evolution of F according to k

Of course the best k depends on the size of the corpus from which S_d is computed. In the present case, the best results are reached for $k = 4$: P= 91.66% and R=90.31%. But even without tuning, using $k = 1$, the results as still good: P=89.10% R=87.71%

To highlight the respective role and efficiency of each clue, we have extracted the lexical correspondences for various combinations (see table 2).

	Log-like	S_d	$S_{posi}+S_{cog}+S_{pos}$	$S_d+4.(S_{posi}+S_{cog}+S_{pos})$
P	0.8145	0.8080	0.6352	0.9166
R	0.7976	0.7915	0.6176	0.9031
F	0.8060	0.7998	0.6263	0.9098

[2] See http://www.xrce.xerox.com/

	S_d+4.S_{posi}	S_d+4.S_{cog}	S_d+4.S_{pos}
P	0.8951	0.8214	0.8464
R	0.8780	0.8036	0.8296
F	0.8865	0.8124	0.8379

Table 2: Results for various clue combination

3.2 Discussion

The results displayed on table 3 shows that the distributional clue is preponderant. It gives more or less the same results than *log-like* (Dunning, 1993). The combination of all other clues gives poor results and shows what can be expected on a small corpus (S_d needs to be computed on a large set of segment pairs). It is noticeable that all the clues are *cumulative*: the more the clues we use, the better the results we get. The efficiency of each clue can be ranked as follows: $S_{pos}<S_{cog}<S_{posi}<S_d$

It appears that the part-of-speech clue gives not very interesting information. The results for a non-tagged corpus would be almost the same.

To give a benchmark, we have also implemented the Melamed (1998) method, which bears some similarity with ours, within an iterative framework inspired by EM-algorithm. For method A, stability was reached after 3 iterations. For method B, the λ+ was set as 0.86 and λ- as 0.095. Parameters were stable after 4 iterations. For comparison's sake, our results are computed using the same data (i.e. S_d has been computed on BOVARY only). Precision and recall displayed on table 3 shows that, even without tuning, the Null hypothesis approach using four clues is more efficient.

	method A	method B	*S* with *k*=1	*S* with *k*=4
P	0.7774	0.7744	0.8379	0.8784
R	0.7552	0.7604	0.8183	0.8625
F	0.7661	0.7674	0.8280	0.8704

Table 3: Results for method A, B and *S*

4 Conclusion

This experiment shows that it is possible to get good results, with precision and recall around 90%, for bilingual correspondences extraction between content words. The originality and interest of the *Null hypothesis* approach is that no training set is required. In forthcoming experiments, we plan to study the effect of a semantic clue, based on the EuroWordNet interlingual index.

Acknowledgements

Thanks to our partners: Marc El-Bèze and Grégoire Moreau de Montcheuil (LIA), Claude Richard and Régis Meyer (ACCE), SINEQUA, and RIAM which supports the CARMEL Project.

References

L. Ahrenberg, M. Merkel, A.S. Hein and J. Tiedemann 2000. Evaluation of word alignment systems. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC-2000.* European Language Resources Association.

R.D. Brown, J.G. Carbonell, Y. Yimin 2000. Automatic dictionary extraction for cross-la,guage information retrieval. In *Parallel Text Processing*, J. Véronis, ed., 275-297. Dordrecht, Nederlands: Kluwer Academic Publishers.

Ido Dagan, K.W. Church and W. Gale. 1993 Robust Bilingual Word Alignment for Machine Aided Translation. In *Proceedings of the Workshop on Very Large Corpora*, Academic and Industrial Perspectives. pp. 1–8.

Ted Dunning 1993. Accurate Methods for the Statistics of surprise and Coincidence. *Computational Linguistics*. Vol 19, 1, pp. 61–74.

Huang Jin-Xia and Choi Key-Sun 2000. Chinese-Korean Word Alignment Based on Linguistic Comparison. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL-2000*, pp.392-399

Gabriel Lopes and João Mexia 2001. Cognates alignment. In *Proceedings of MT Summit VIII*.

Dan Melamed 1998. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora (WVLC3)*, Boston, MA.

J. Nerbonne 2000. Parallel texts in computer-assisted language learning, In *Parallel Text Processing*, J. Véronis, ed., pages 299-311. Dordrecht, Kluwer Academic Publishers.

Hwee Tou Ng, Wang Bin and Chan Yee Seng 2003. Exploiting Parallel Texts for Word Sense Disambiguation: In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL 2003*.

Jörg Tiedemann 2003. Combining Clues for Word Alignment. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL03*), Budapest, Hungary, April 12-17, 2003.

Dan Tufis. 2002. A cheap and fast way to build useful translation lexicons. In *Proceedings of the 19th International Conference on Computational Linguistics, COLING-2002.*

J. Véronis and P. Langlais. 2000. Evaluation of parallel text alignment systems – The ARCADE project. In *Parallel Text Processing*, J. Véronis, ed., 49-68. Dordrecht, Nederlands: Kluwer Academic Publishers.