Evaluation of statistical tools for automatic extraction of lexical correspondences between parallel texts

Olivier Kraif

LILLA - Université de Nice Sophia Antipolis, Nice (France)
okraif@mageos.com

Keywords : Bilingual Alignment, Lexical Correspondence, Bi-text, Conditional Entropy, Log-likelihood Ratio, Machine Aided Translation, Example Based Machine Translation

Abstract

An interesting application of bilingual aligning is the automatic extraction of bilingual lexicons. At this prospect, this article proposes an evaluation of some statistical measures used for automatic extraction of lexical correspondences. After introducing classical measures like mutual information, t-score and log-likelihood ratio, we present another statistical measure based equally on distributions and cognateness, i.e. formal resemblance. In a simple algorithmic framework, we show how this combination can produce a slight improvement of the results. Then, in order to find adequate balance between precision and recall, we test three methods of filtering of the results. Finally, we try to find some correlation between an *a posteriori* evaluation (using a manually extracted gold standard) and an *a priori* evaluation based on formal characteristics of the correspondences.

Introduction

In the last few years, much interest has been given to the outcome of translation aligning : Isabelle (1992) proposed to use bilingual parallel texts, or *bi-texts*, i.e. segmented and aligned translation corpora, as a *Corporate Memory* for translators. In that kind of corpora, the linguistic and translational knowledge is stored implicitly in the recorded examples of translation.

An interesting application of bilingual aligning is the automatic extraction of bilingual lexicons. A lot of works (Dunning 1993, Dagan *et al.*, 1993, Gaussier & Langé, 1995, Melamed 1998) have shown how to use statistical filters to pair lexical units that have a similar distribution in each part of the bi-text. As a great proportion of these similar units are translational equivalents, they can be useful to establish bilingual (or multilingual) glossaries upon empirical observation.

Given the large variety of algorithms and techniques devoted to alignment, we are now entering an evaluation phase, and some large scale projects like Arcade (Langlais *et al.*, 1998) intend to give a coherent framework for definition and

evaluation of the aligning task. We propose here an evaluation of the some of these techniques.

Through a simple algorithm, we first compare the performance of several quantitative measures. In a second step, we show how these results can be correlated with formal characteristics of the set of correspondences: indeed this correlation indicates the possibility to evaluate the results without any recourse to the gold standard. Finally, we study different methods to filter out the erroneous pairs of lexical correspondences: these techniques are useful to find an appropriate balance between precision and recall.

Design of evaluation

The evaluation task consists of two steps : given a test corpus, we have to determine first a *gold standard*, i.e. a manually constructed set of correspondences that are considered to be exact. Then we have to implement some *metrics* in order to compare quantitatively any other set of lexical pairs with the standard.

The metrics used for this comparison are the classical measure of precision, recall and F-measure.

$$P = \frac{|C \cap C_{ref}|}{|C|} \quad R = \frac{|C \cap C_{ref}|}{|C_{ref}|} \quad et \quad F = \frac{2 \times (P \times R)}{(P + R)} \qquad (1)$$

where C represents the set of the evaluated correspondences, and C_{ref} the set of correspondences of the gold standard.

We manually identified multi-words units independently for each language, following semantic and syntactic criteria: non compositional compounds, frozen phrases, colloquial expressions have been clustered in single units. Then, to pair the units with each other, we followed a simple criteria : the translational equivalence at a general level, independently of the particular context of our corpus. For a discussion about problems raised by manual pairing, see Kraif (forthcoming). The corpus is composed of a sample of 700 pairs of sentences drawn from the French and English versions of the JOC corpus used in the Arcade Project. It is a record of written questions asked by members of the European Parliament, with the corresponding answer of the European Commission. These questions, published in 1993 in one section of the C Series of the Official Journal of the European Community, have been recorded within the MLCC-MULTEXT projects. They concern various matters regarding environment, economic policy, transport, agriculture, human rights, foreign policy, institutions, etc..

The statistics of co-occurrence were computed on the whole French and English versions of the JOC corpus, including 69 160 automatically aligned sentence pairs.

Statistical measures

We tested the following measures :

- MI: the mutual information which quantifies the amount of information brought by an event on another event (Shannon, 1949).

- TS: the t-score, designed to filter out insignificant mutual information values (Fung et al. 1994).

- LR: the log-likelihood ratio (Dunning, 1993), based on a binomial distribution model, more adapted for rare events.

- P0: the log-probability of the null hypothesis, i.e. the probability for two units (u_1,u_2) to co-occur only by chance. We computed this probability assuming a binomial distribution. Without simplification, this probability can be expressed by equation 2:

$$P_0(n_{12}/n,n_1,n_2)=\frac{\binom{n}{n_1}\cdot\binom{n_1}{n_{12}}\cdot\binom{n-n_1}{n_2-n_{12}}}{\binom{n}{n_1}\cdot\binom{n}{n_2}} \qquad (2)$$

where n is the number of sentence pairs, n_1 and n_2 are the respective numbers of occurrences of u_1 and u_2, and n_{12} is the number of times that u_1 and u_2 co-occur in the same sentence pairs. This probability is computed as the result of 3 independent draws, assuming that each unit occurs only once in the same sentence pair :

$\binom{n}{n_1}$ is the number of different possible draws for the n_1 occurrences of u_1.

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

$\binom{n-n_1}{n_2-n_{12}}$ 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.

The denominator $\binom{n}{n_1}\cdot\binom{n}{n_2}$ is the total number of possible draws without making any assumption on n_{12}.

- CO: the log probability of *cognateness*, (Simard et al. 1992) i.e. the probability to observe superficial resemblance between two compared strings, under null hypothesis. The event of cognateness is determined by counting the length of the common maximum sub-string, using techniques that we have previously developed for sentence aligning (Kraif, 1999). Two units are considered as potential cognates if the sub-string exceeds a certain proportion of the smallest unit. For instance, between *contrôle* (French) and *control* (English), there is a sub-string of length 6 : c-o-n-t-r-l, which represents 6/7 of *control*. We tested two different thresholds for this proportion: 2/3 and 1/2. Thus, we obtain two versions of CO, COa and COb, yielding different tunings between noise and silence in the identification of cognateness: COa, for which the threshold is 2/3, is less noisy and more silent than COb.

The probability of cognateness between two randomly drawn units has been computed from empirical observations (on another corpus).

- PC = P0 + CO: this metric cumulates two different kinds of information, co-occurrences and resemblance, assuming that they are independent. Given two units that co-occur n_{12} times and that are potential cognates, it estimates the unlikelihood that this event could happen only by chance.

Algorithm

We implemented all these statistics in a straightforward algorithm:

1. to create a set of candidate pairs, every unit of the source sentence is compared with every unit of the target, giving for each pair an association score. The scores are then ranked in descending order.

2. the best scoring pair (u_1,u_2) is recorded.

3. all the other candidate pairs that involve either u_1 or u_2 are removed.

Step 2 and 3 are reiterated until there is no more candidate pair.

In order to reduce the effect of indirect associations, step 3 implements the one-to-one assumption: each unit can be paired with only one unit in each sentence pair. As demonstrated by Melamed (1998), this algorithm approximately establishes the best scoring set of correspondences under the one-to-one assumption,

Results

As shown on figure 1, precision and recall are strongly linked, because each extraction yields roughly the same number of pairs. Thus, for precision, recall and F, we can rank equally the measures in ascending order in the following manner: COa, COb, MI, TS, P0, LR and PC. P0 and LR have a very close behaviour: their distributions are asymptotically the same. The best value of F is around 65%, with PC. The combination of CO and P0 improves slightly the results, showing that the two kinds of information are cumulative. For CO alone, we notice that COb is more efficient than COa: the extra noise brought by COb seems to be filtered out by the algorithm, because of the competition between different pairing. This fact indicates that, in the recourse of a bilingual dictionary to extract correspondences, the noise brought by polysemy can be reduced, and it may be more interesting to favour the completeness of the dictionary.

Finally, if we compute the co-occurrences after having lemmatised, to reduce morphological variations, the global results are slightly improved of about 1% (see figure 2).

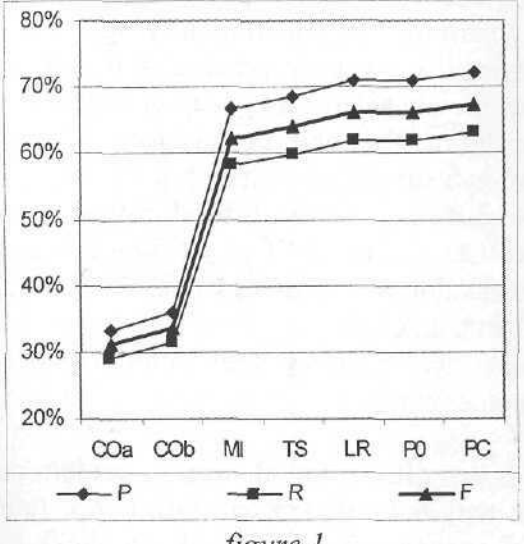

figure 1
Precision, recall and F for each metric

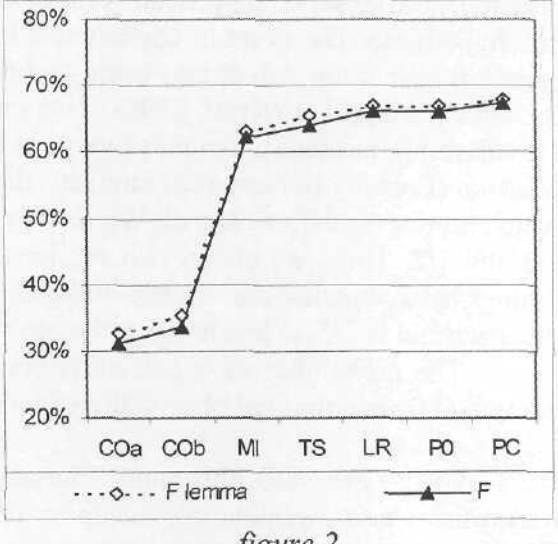

figure 2
Results with / without lemmatisation

Filtering methods

A filtering method has to fulfil two conditions: eliminating the most erroneous pairing while keeping the most correct pairs. For this task, we can use the calculated scores as a good indicator of the reliability of an association.

We tested three methods of filtering:

- absolute filtering: we filter out all the pairs which get a score below a certain threshold.

- relative filtering: for each aligned sentence, we keep a fixed proportion of the best scoring pairs.

- differential filtering: we can suppose that if different target units compete with each other to be associated with a same source unit, there is a greater uncertainty about the association. Thus, for each recorded pair, we compute the ratio between its score and the score of the second best competing pair. If the ratio is lower than a certain threshold they are both eliminated.

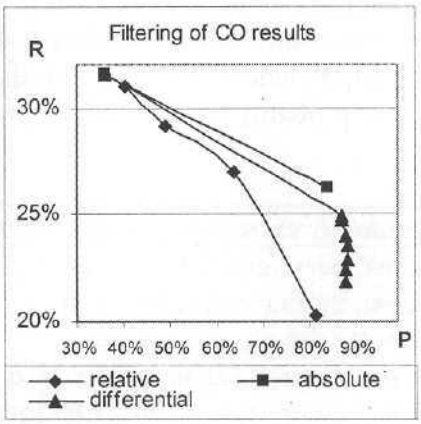

figure 3

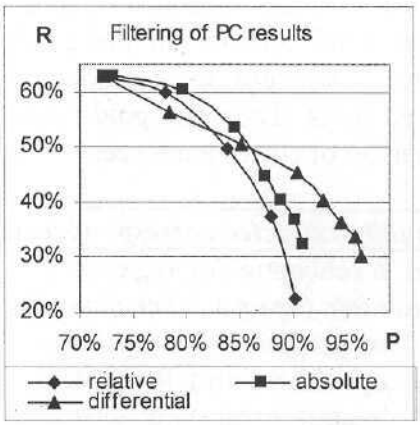

figure 4

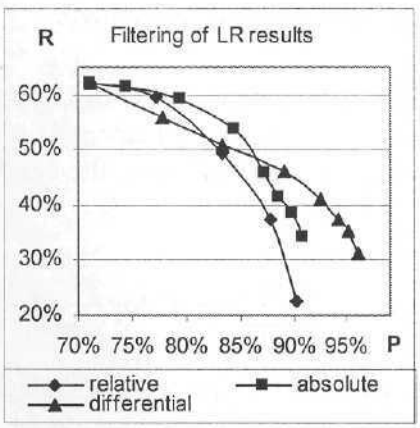

figure 5

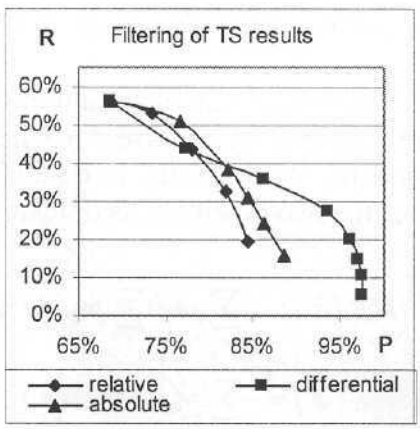

figure 6

For each method, we computed precision and recall of correspondences with different values of the threshold. Figures 3, 4,5 and 6 display these filtered results for COb, LR, PC and TS. They clearly show that it is possible to increase precision to very high levels by sacrificing recall: for instance, with PC, we can get a 96% precision with a recall of about 35%.

For the cognate-based measure, the differential filtering allows a 90% precision for a 25% recall, demonstrating that the important noise brought by n-gram comparison can easily be reduced by a simple algorithmic framework.

We notice that these methods are suitable for different tasks: if one needs to emphasise recall, the absolute filtering is adapted; conversely, for a high precision, differential filtering yields better results.

Conditional entropy of a set of correspondences

If we compare the gold standard with a set of randomly drawn correspondences, we notice some differences at a formal level. As expected, the correspondences are far more regular in the case of the gold standard: a source lexical unit is often paired with the same target units. Of course, in this case, paired units are strongly linked by a same semantic content. When units are randomly paired, without any constraint, the correspondences are unsystematic. For instance, for the 10 occurrences of 'against' in the gold standard, we count only 3 different French translations, whereas in a random set of correspondences we get 10 different associated units. Thus, the gold standard contains probably more "order" than any erroneous set of correspondences.

Manually extracted correspondences	*Randomly extracted correspondences*
(against, à l'encontre de), (against, à l'encontre de), (against, à l'encontre de), (against, au détriment de), (against, contre), (against, contre), (against, contre), (against, contre), (against, contre)	(against, par), (against, procédure), (against, moratoire), (against, à l'encontre de), (against, dont) (against, contre), (against, effectivement) (against, charges), (against, Etat membre) (against, qui)

table 1: manually extracted correspondences contains less entropy

This indicates an other kind of evaluation, based on the following hypothesis: the more regular a set of correspondences is, the closer to the gold standard it should be. To quantify the regularity of a set of pairs, we propose to calculate the conditional entropy of the two distributions of lexical units :

$$H(F/E) = -\sum_e p(e) \sum_f p(f/e) \log p(f/e) = -\sum_e \sum_f p(e,f) \log \frac{p(e,f)}{p(e)} \quad (3)$$

$$H(E/F) = -\sum_f p(f) \sum_e p(e/f) \log p(e/f) = -\sum_f \sum_e p(e,f) \log \frac{p(e,f)}{p(f)} \quad (4)$$

where e and f are referring to lexical units of the English and French texts.

To observe the possible correlation between conditional entropy and the correctness of an extraction of correspondences, we need to get different sets of correspondences, with various results for precision and recall. Using the previous algorithm (called *Algo* 2), we developed a measure combining PC and a random draw, in different proportions : we obtained seven sets with F-measure from 6% to 65%.

In order to have more generality we implemented several other extractions using CO, IM, TS, LR, P0 and PC with another simpler algorithm (called *Algo* 1),

where each source unit is paired with the best-scoring target unit. The results of this algorithm are inferior and have different formal characteristics: the pairing between the units of two aligned sentences are not one-to-one, but sometimes many-to-one.

Then, we filtered the results of Algo 1 and Algo 2 (using differential filtering). We finally obtained 31 sets of correspondences. For each of these sets, we computed H(e/f) and H(f/e).

As shown in figure 7, we observe a strong correlation between the precision P and the value of max(H(e/f),H(f/e)). The linear correlation coefficient between P and max(H(e/f),H(f/e)) is about -0,95.

Notice that recall (as well as F) can be deduced from precision, taking into account the number of proposed pairs, but it is not *directly* linked to the conditional entropy.

We plotted a dot for the gold standard, for which the conditional entropy is low but not minimal. This is due to the normal variations induced by the process of translation. If some extractions yield lower entropy, it can be explained by a very low recall.

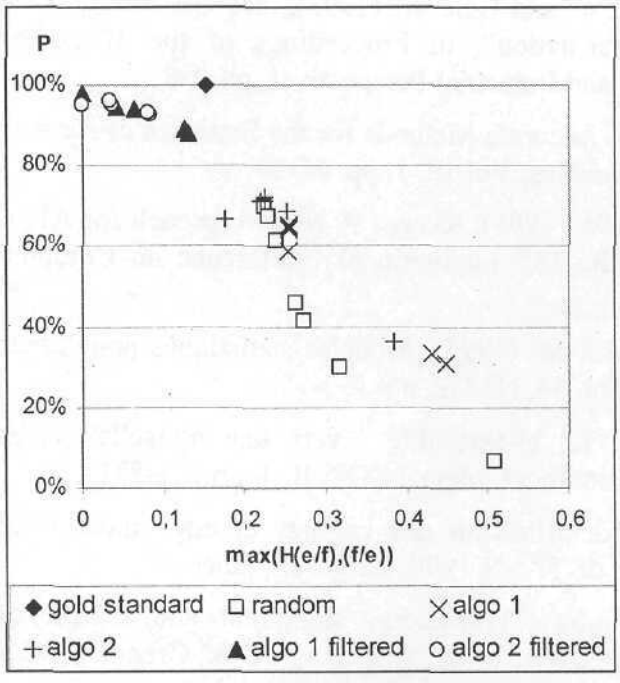

figure 7
Correlation between Conditional Entropy and Precision for different sets of correspondences : the Gold Standard, a random set, and various extractions obtained with different metrics and algorithms

Conclusion

We showed that the implementation of statistical measures in a simple framework can yield interesting results in the task of correspondence extraction. Statistics based on binomial distribution, the log-likelihood and the log-probability of null hypothesis, seem to behave very well. The latter is very similar to the former, but as a probability it has not the same meaning and can be combined with other

probability in a more coherent way (by modelling a draw process). Indeed, the combination with cognateness gave encouraging results. In further works, it could be interesting to study the same kind of combination, with information extracted from a bilingual dictionary.

In addition, we notice that it is possible to increase significantly the precision of results with simple filtering techniques. The absolute and differential filtering each have their advantage, depending on which balance between recall and precision is required.

Finally, we showed how to give an approximate evaluation of a set of correspondences even when the gold standard is not available: by calculating conditional entropy for the distributions of the paired lexical units, precision of different extractions can be compared and roughly estimated.

References

Dagan I., Church K.W. and Gale W. (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.

Dunning, T. (1993), Accurate Methods for the Statistics of surprise and Coincidence, Computational Linguistics, Vol 19, 1, pp. 61-74

Fung P., Church K.W. (1994), K-vec : A New Approach for Aligning Parallel Texts, in Proceedings of the 15th International Conference on Computational Linguistics, Kyoto

Gaussier, E., Langé J.-M. (1995), Modèles statistiques pour l'extraction de lexiques bilingues, T.A.L., Vol. 36, N° 1-2, pp. 133-155

Isabelle P. (1992), La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie, Meta, XXXVII, 4, pp.721-731

Kraif O. (1999), Identification des cognats et alignement bi-textuel : une étude empirique, In Actes de TALN 1999, Cargèse, France

Kraif O. (forthcoming), Translation alignment and lexical correspondences : a methodological reflection. In B. Altenberg & S. Granger, Ed., Lexis in contrast. Studies in Corpus Linguistics. John Benjamins

Langlais P., Simard M., Veronis J. et al, (1998), ARCADE : A Cooperative Research Project on Parallel Text Alignment Evaluation, available at : http://www.lpl.univ-aix.fr/projects/arcade

Melamed, D. (1998), Word-to-Word Models of Translational Equivalence, Institute for Research in Cognitive Science. Technical Report #98-06, University of Pennsylvania, available at http://www.cis.upenn.edu/~melamed/home.html

Shannon C. (1949), A mathematical theory of communication, Univ. of Illinois Press.

Simard M., Foster G., Isabelle P. (1992), Using cognates to align sentences, in Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, Montréal, Canada, pp. 67-81.