Combined word alignments

Dan Tufiş, Radu Ion, Alexandru Ceauşu, Dan Ştefănescu
Romanian Academy Institute for Artificial Intelligence
13, "13 Septembrie", 74311, Bucharest 5, Romania
{tufis, radu, alceusu, danstef}@racai.ro

Abstract

We briefly describe a word alignment system that combines two different methods in bitext correspondences identification. The first one is a hypotheses testing approach (Gale and Church, 1991; Melamed, 2001; Tufiş 2002) while the second one is closer to a model estimating approach (Brown et al., 1993; Och and Ney, 2000). We show that combining the two aligners the results are significantly improved as compared to each individual aligner.

Introduction

In (Tufiş, 2002) we described a translation equivalence extraction program called TREQ the development of which was twofold motivated: to help enriching the synsets of the Romanian wordnet (Tufiş et al. 2004a) with new literals based on bilingual corpora evidence and to check the interlingual alignment of our wordnet against the Princeton Wordnet. The translation equivalence extractor has been also incorporated into a WSD system (Tufiş et al., 2004b) part of a semantic web annotation platform. It also constituted the backbone of our **TREQ-AL** word aligner which successfully participated in the previous HLT-NAACL 2003 Shared Task[1] on word alignment for Romanian-English parallel texts. A detailed description of TREQ&TREQ-AL is given in (Tufiş et al. 2003b) and it will be very shortly overviewed.

A quite different approach from our hypotheses testing implemented in the TREQ-AL aligner is taken by the model-estimating aligners, most of them relying on the IBM models (1 to 5) described in the (Brown et al. 1993) seminal paper. The first wide-spread and publicly available implementation of the IBM models was the GIZA program, which itself was part of the SMT toolkit EGYPT (Al-Onaizan et al., 1999). GIZA has been superseded by its recent extension GIZA++ (Och and Ney, 2000, 2003) publicly available[2]. We used the translation probabilities generated by GIZA++ for implementing a second aligner, **MEBA**, described in a little more details in a subsequent section. The alignments produced by MEBA were compared to the ones produced by TREQ-AL. We used for comparison the Gold Standard[3] annotation from the HLT-NAACL 2003 Shared Task. In order to combine the two aligners we had to check whether their accuracy was comparable and that when they are wrong the set of mistakes made by one aligner is not a proper set of the errors made by the second one. The first check was performed by using McNamer's test (Dieterich, 1998) and for the second we used Brill &Wu test (Brill, Wu, 1998). Both tests confirmed that the conditions for combining were ensured so, we built the combiner.

The Combined Word Aligner, **COWAL**, is a wrapper of the two aligners (TREQ-AL and MEBA) ensuring the pre- and post-processing. It is complemented by a graphical user interface that allows for the visualisation of the alignments (intermediary and the final ones) as well as for their editing. We should note that the corrections made by the user are stored by COWAL as positive and negative examples for word dependencies (in the monolingual context) and translation equivalencies (in the bilingual context). In the current version the editorial logs are used by the human developers but we plan to further extend COWAL for automatic learning from this extremely valuable kind of data.

The bitext processing

The two base aligners and their combination use the same format for the input data and provide the alignments in the same format. The input format is obtained from two raw texts which represent reciprocal translations. If not already sentence aligned, the two texts are aligned. In the shared task this step was not necessary since both the training data and evaluation data were provided in the sentence aligned format.

The texts in each language are then tokenized with the MULTEXT multilingual tokenizer[4]. The tokenizer is a finite state automaton using language specific

[1] http://www.cs.unt.edu/~rada/wpt/index.html#shared
[2] http://www.fjoch.com/GIZA++.2003-09-30.tar.gz

[3] We noticed in the Gold Standard two sentences where alignments were wrongly shifted by one position (due to an unprintable character) and we corrected them.
[4] http://aune.lpl.univ-aix.fr:16080/projects/multext/MtSeg/

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 107–110,
Ann Arbor, June 2005.

resources. It recognizes several compounds (phrasal verbs, idioms, dates) and split contrasted or cliticized constructions. This tokenization considerably differs from the one prescribed by the Shared Task where a token is any character string delimited by a blank or a punctuation sign (which itself is considered a token).

Since our processing tools (especially the tokeniser) were built with a different segmentation strategy in mind, we generated the alignments based on our own tokenization and, at the end, we "re-tokenised" the text according to original evaluation data (and consequently re-index) all the linking pairs. After tokenization, both texts are tagged and lemmatized. We used in-house language models and lemmatizers and the Brants's TnT tagger[5]. For both English and Romanian we used MULTEXT-EAST[6] compliant tagsets. With different tags, a tagset mapping table becomes an obligatory external resource. Although, more often than not, the translation equivalents have the same part-of speech, relying on such a restriction would seriously affect the alignment recall. However, when the translation equivalents have different parts of speech, this difference is not arbitrary. During the training phase we estimated bilingual *POS affinities:* $\{p(POS_m^{RO}| POS_n^{EN})\}$ and $\{p(POS_n^{EN}|POS_m^{RO})\}$. POS affinities were used as one of the information sources in dealing with competitive alignments.

The next preprocessing step is represented by a rather primitive form of sentence chunking in both languages. They roughly correspond to (non-recursive) noun phrases, adjectival phrases, prepositional phrases and verb complexes (analytical realization of tense, aspect mood and diathesis and phrasal verbs). The "chunks" are recognized by a set of regular expressions defined over the tagsets. Finally, the bitext is assembled as an XML document (XCES-Align-ana format), as used in the MULTEXT-EAST corpus, which is the standard input for most of our tools, including COWAL alignment platform.

The three aligners

TREQ-AL generates translation equivalence hypotheses for the pairs of words (one for each language in the parallel corpus) which have been observed occurring in aligned sentences more than expected by chance. The hypotheses are filtered by a loglikelihood score threshold. Several heuristics (string similarity-cognates, POS affinities and alignments locality[7]) are used in a competitive linking manner (Melamed, 2001) to make the final decision on the most likely translation equivalents. Given that, initially, this program was designed for extracting translation equivalents for the alignment of the Romanian wordnet to the Princeton wordnet, it deals only with one to one mappings. To cope with the many to many mappings (especially for functional words alignment), the earlier version of the translation equivalence extractor encoded some general rules assumed to be valid over a large set of natural languages such as: auxiliaries and verbal particles (infinitive, subjunctive, aspectual and temporal) are related to the closest main verb, determiners (articles, pronominal adjectives, quantifiers) are related to the closest nominal category (noun or pronoun). Currently this part of the TREQ-AL code became redundant because the chunking module mentioned before does the same job in a more general and flexible way.

MEBA is an iterative algorithm which uses the translation probabilities, distorsions and POS-affinities generated by GIZA++ and takes advantage of all preprocessing phases mentioned in the previous section. In each step are aligned different categories of tokens (content words, named entities, functional words) in decreasing order of statistical evidence. The score of a link is computed by a linear function of 7 parameters' scores: translation probability, POS affinity, string similarity, *alignments locality* (both strict and weaker versions) distortions and the entropy of the translation equivalents. For all these parameters, in each processing step, we empirically set minimal thresholds and various weights. The tokens considered for the computing translation probabilities are the lemmas trailed by the grammatical categories (eg. plane_N, plane_V plane_A). This way we aimed at avoiding data sparseness and filtering noisy data. For highly inflectional languages (as Romanian is) the use of lemmas instead of word occurrences contributes significantly to the data sparseness reduction. For languages with weak inflectional character (as English is) the POS trailing contributes especially to the filtering the search space. Each processing step is controlled by above mentioned parameters, the weights and thresholds of which vary from step to step (even the order of the processing steps is one of the possible parameters).

The first alignment step builds only links with a high level of certainty (that is cognates, pairs of high translation probability and high POS affinity). The grammatical categories which are considered in this step are user controlled (usually nouns, adjectives or non-auxiliary verbs and which have the fewest competitive translations). The next processing steps try to align

[5] http://acl.ldc.upenn.edu/A/A00/A00-1031.pdf

[6] http://nl.ijs.si/ME/V2/

[7] The *alignments locality* heuristics exploits the observation made by several researchers that adjacent words of a text in the source language tend to align to adjacent words in the target language. A more strict alignment locality constraint requires that all alignment links starting from a chunk, in the one language end in a chunk in the other language. This restricted form of locality is relevant for related languages.

content words (open class categories) as confidently as possible, following the alignments in previous steps as anchor points. In all steps the candidates are considered if and only if they meet the minimal threshold restrictions. If the input bitext is chunked, the strict alignment locality heuristics is very effective to determine the correct alignment even for unseen pairs of words (or for which the translation equivalence probability is below the considered threshold). When the pre-chunking of the parallel texts is not available, MEBA uses the weaker form of the locality heuristics by analyzing the alignments already existing in a window of N tokens centered on the focused token. The window size is variable, proportional to the sentence length. For all alignments in the window, an average displacement is computed and, among the competing alignments, preference will be given to the links with displacement values closer to the average one.

The functional words and punctuation are processed in the last step and their alignments are guided by the POS-affinities and alignment locality heuristics. If none of the alignment clues or their combination (Tiedemann, 2003) is strong enough, the functional words are automatically aligned with the word(s) their governor is aligned to. The governor is chunk-based defined: it is the content word of a chunk (if there are more content words in a chunk, then the governor is the grammatical head). If the chunking is not available, the closest content word is selected as the governor. Proximity is checked to the left or to the right according to the frequencies of the POS-ngram containing the current functional word.

We should mention that the probabilities computed during the training phase are not re-estimated for each run-time processing step. At run-time only the weights and thresholds change from step to step.

***COWAL*,** the combined aligner takes advantage of the alignments independently provided by TREQ-AL and MEBA. The simplest combination method consists in computing either the union (high recall, low precision), or the intersection (lower recall, higher precision) of the independent alignments. We evaluated both these simple methods of combination and found that the best F-measure was provided by the union-based combination. Although for the shared task we submitted the union-based combined alignment (*Baseline COWAL*, see Table 1), there are various ways to improve it. We discuss three cases where improvement is possible (C1, C2 and C3, see below) and which were evaluated after the submission deadline. The results of this (unofficial) evaluation are summarized in Table 1 by the *f-COWAL* line. These cases refer to competing links that appeared after the union of the independent alignments. The conflicts resolution is based on the (weak) locality and distortion heuristics discussed before. The currently identified competing links are only those for which the following conditions apply:

C1) if one aligner found for a word W a non-null alignment and the other aligner generated for the same word W a null link, then the baseline alignment contains an impossible situation: the token W is recorded both as translated and not-translated in the other language. The translation probabilities, POS affinity and the relative displacement of the tokens in the non-null candidates were the strongest decision criteria. We found that in about 60% of the cases the null alignments were mistaken. So, for the time being, we simply eliminated the null competing alignments (this should be addressed in a more principled way by the future version of the combiner).

C2) long distant competing links; this case appears when one aligner found for the word Ws the link to the target word Wt_m, the other aligner found for Ws the target Wt_n, and the distance between Wt_m and Wt_n, is more than 3 words (in a future version this maximum distance will be a dynamic parameter, depending on the sentence length and the POS of Ws).

C3) competing links to the same target(s) of a word occurring several times in the same sentence; consider, for example, the Romanian fragment:
"…la_1 Neptun, la_2 Orastie si la_3 Afumati, …
which in English is translated by the next segment:
"…in Neptun, Orastie and Afumati…
In spite of the gold standard considering that all three occurrences of the preposition "la" in Romanian (la_1, la_2 ,la_3) are aligned to the same word in English ("in"), the filtering, in this case, licensed only the alignment "la_1 <-> in". We consider that this filtered alignment is correct, since omitting "la_2" and "la_3" does not alter the syntactic correctness of the Romanian text, and also because the insertion in the English fragment of the preposition "in" before "Orastie" and before "Afumati" wouldn't alter the grammaticality of the English fragment. Since both repetitions and omissions are optional, we consider that only the first occurrence of the preposition ("la_1") is translated in English, while the others are omitted.

Another possible improvement (not implemented yet) was revealed by observing that the final result contained several incomplete n-m (phrasal) alignments. It is likely that even an elementary n-gram analysis (both sides of the bitext) would bring valuable evidence for improving the phrasal alignments.

Post-processing

As said in the second section, our tokenization was different from the tokenization in the training and test data. To comply with the evaluation protocol, we had to re-tokenize the aligned text and re-compute the indexes

of the links. Some multi-word expressions recognized by the tokenizer as one token, such as dates (*25 ianuarie, 2001*), compound prepositions (*de la*, *până la*), conjunctions (*pentru ca*, *de când*, *până când*) or adverbs (*de jur împrejur*, *în față*) as well as the hyphen separated nominal compounds (*mass-media*, *prim-ministru*) were split, their positions were re-indexed and the initial one link of a split compound was replaced with the set obtained by adding one link for each constituent of the compound to the target English word. The same hold for the other way around. Therefore if two multiword expressions were initially found to be translation equivalents (one alignment link) after the post-processing number of generated links became N*M, where N represented the number of words in the first language compound and M the number of words in the second language compound.

Evaluation and conclusions

Neither TREQ-AL nor MEBA needs an a priori bilingual dictionary, as this will be automatically extracted by the TREQ or GIZA++. We made evaluation of the individual alignments in both experimental settings: without a startup bilingual lexicon and with an initial mid-sized bilingual lexicon. Surprisingly enough, we found that while the performance of TREQ-AL increases a little bit (approx. 1% increase of the F-measure) MEBA is doing better without an additional lexicon. So, in the evaluation below MEBA uses only the training data vocabulary.

Aligner	Precision	Recall	F-meas.	AER
TREQ-AL	81.71	60.57	69.57	30.43
MEBA	**82.85**	60.41	69.87	30.13
Baseline (union)COWAL	70.84	**76.67**	**73.64**	26.36
f-COWAL (H1+H2+H3)	**87.17**	70.25	**77.80**	**22.20**

Table 1. Evaluation results against the official GS

After the release of the official Gold Standard we noticed and corrected some obvious errors and also removed the controversial links of the type c) discussed in the previous section. The evaluations against this new "Gold Standard" showed, on average, 3.5% better figures (precision, recall, F-measure and AER) for the individual aligners, while for the combined classifiers, the performance scores were about 4% better.

MEBA is very sensitive to the values of the parameters which control its behavior. Currently they are set according to the developers' intuition and after the analysis of the results from several trials. Since this activity is pretty time consuming (human analysis plus re-training might take a couple of hours) we plan to extend MEBA with a supervised learning module, which would automatically determine the "optimal" parameters (thresholds and weights) values.

References

Al-Onaizan, Y., Curin, J., Jahr, M., Knight K., Lafferty, J., Melamed, D., Och, F. J., Purdy, D., Smith, N.A., Yarowsky, D. (1999) : Statistical Machine Translation, Final Report, JHU Workshop, 42 pages

Brill, E., and Wu, J. (1998). "Classifier Combination for Improved Lexical Disambiguation" *In Proceedings of COLING-ACL'98* Montreal, Canada, 191-195

Brown, P. F., Della Pietra, S.A., Della Pietra, V. J., Mercer, R. L.(1993) "The mathematics of statistical machine translation: Parameter estimation". *Computational Linguistics*, 19(2) pp. 263–311.

Dietterich, T. G., (1998). "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". *Neural Computation, 10* (7) 1895-1924.

Gale, W.A. and Church, K.W. (1991). „Identifying word correspondences in parallel texts". Proceedings of the Fourth DARPA Workshop on Speech and Natural Language. Asilomar, CA, pp. 152–157.

Melamed, D. (2001). *Empirical Methods for Exploiting Parallel Texts*. Cambridge, MA: MIT Press.

Och, F.J., Ney, H. (2003) "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, 29(1), pp. 19-51

Och, F.J., Ney, H.(2000) "Improved Statistical Alignment Models". Proceedings of the 38th ACL, Hongkong, pp. 440-447

Tiedemann, J. (2003) "Combining clues for word alignment". In Proceedings of the 10th EACL, Budapest, pp. 339–346

Tufiş, D.(2002) "A cheap and fast way to build useful translation lexicons". Proceedings of COLING2002, Taipei, pp. 1030-1036.

Tufiş, D., Barbu, A.M., Ion R (2003).: „TREQ-AL: A word-alignment system with limited language resources", Proceedings of the NAACL 2003 Workshop on Building and Using Parallel Texts; Romanian-English Shared Task, Edmonton, pp. 36-39

Tufiş, D., Ion, R., Ide, N.(2004a): Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. Proceedings of COLING2004, Geneva, pp. 1312-1318

Tufis, D., Barbu, E., Mititelu, V., Ion, R., Bozianu, L.(2004b): „The Romanian Wordnet". In Romanian Journal on Information Science and Technology, Dan Tufiş (ed.) Special Issue on BalkaNet, Romanian Academy, 7(2-3), pp. 105-122.