TREQ-AL: A word alignment system with limited language resources

Dan Tufiş, Ana-Maria Barbu, Radu Ion
Romanian Academy Institute for Artificial Intelligence
13, "13 Septembrie", 74311, Bucharest 5, Romania
{tufis,abarbu,radu}@racai.ro

Abstract

We provide a rather informal presentation of a prototype system for word alignment based on our previous translation equivalence approach, discuss the problems encountered in the shared-task on word-aligning of a parallel Romanian-English text, present new evaluation results and suggest further ways of improving the alignment accuracy.

1 Introduction

In (Tufiş and Barbu, 2002; Tufiş, 2002) we largely described our extractor of translation equivalents, called TREQ. It was aimed at building translation dictionaries from parallel corpora. We described in (Ide et al. 2002) how this program is used in word clustering and in checking out the validity of the cross-lingual links between the monolingual wordnets of the multilingual Balkanet lexical ontology (Stamou et al. 2002). In this paper we describe the TREQ-AL system, which builds on TREQ and aims at generating a word-alignment map for a parallel text (a bitext). TREQ-AL was built in for the Shared Task proposed by the organizers of the workshop on "Building and Using Parallel Texts:Data Driven Machine Translation and Beyond" at the HLT-NAACL 2003[1] conference. TREQ-AL has no need for an a priori bilingual dictionary, as this will be automatically extracted by TREQ. However, if such a dictionary is available, both TREQ and TREQ-AL know to make best use of it. We provide evaluation results for the proposed task in both experimental settings: without a startup bilingual lexicon and with an initial (medium size) bilingual lexicon.

In a dictionary extraction task one translation pair is considered correct, if there is at least one context in which it has been rightly observed. A multiply occurring pair would count only once for the final dictionary. This is in sharp contrast with the alignment task where each occurrence of the same pair equally counts.

Another differentiating feature between the two tasks is the status of functional word links. In extracting translation equivalents one is usually interested only in the major categories (open classes). In our case (because of the WordNet centered approach of our current projects) we were especially interested in POS-preserving translation equivalents. However, since in EuroWordNet and Balkanet one can define cross-POS links, the different POS translation equivalents became of interest (provided these categories are major ones).

The word alignment task requires each word (irrespective of its POS) or punctuation mark in both parts of the bitext be assigned a translation in the other part (or the null translation if the case).

Finally, the evaluations of the two tasks, even if both use the same measures as precision or recall, have to be differently judged. The null alignments in a dictionary extraction task have no significance, while in a word alignment task they play an important role (in the Romanian-English gold standard the null alignments represent 13.35% of the total number of links).

2 The preliminary data processing

The TREQ system requires sentence aligned parallel text, tokenized, tagged and lemmatized. The first problem we had with the training and test data was related to the tokenization. In the training data there were several occurrences of glued words (probably due to a problem in text export of the initial data files) plus an unprintable character (hexadecimal code A0) that generated several tagging errors due to guesser imperfect performance (about 70% accurate). To remedy these inconveniences we wrote a script that automatically split the glued words and eliminated the unprintable characters occurring in the training data.

The text tokenization, as considered by the evaluation protocol, was the simplest possible one, with white spaces and punctuation marks taken as separators. The hyphen ('-') was always considered a separator and consequently taken to be always a token by itself. However, in Romanian, the hyphen is more frequently used as an elision marker (as in "intr-o"= "intru o"/*in a*), a clitics separator (as in "da-mi-l"="da –mi –l"="da mie l"/*give to me it/him*) or as a compound marker (as in "terchea-berchea" /(*approx.) loafer*) than as a separator. In such cases the hyphen cannot be considered a token.

[1] http://www.cs.unt.edu/~rada/wpt/index.html#shared

A similar problem appeared in English with respect to the special *quote* character, which was dealt with in three different ways: it was sometimes split as a distinct token (we'll = we + ' + ll), sometimes was adjoined to the string (a contracted positive form or a genitival) immediately following it (I'm=I+'m,you've = you+'ve, man's = man+'s etc.) and systematically left untouched in the negative contracted forms (couldn't, wasn't, etc).

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 the test data model (and consequently re-index) all the linking pairs.

For tagging the Romanian side of the training bitext we used the tiered-tagging approach (Tufiş, 1999) but we had to construct a new language model since our standard model was created from texts containing diacritics. As the Romanian training data did not contain diacritical characters, this was by no means a trivial task in the short period of time at our disposal (actually it took most of the training time). The lack of diacritics in the training data and the test data induced spurious ambiguities that degraded the tagging accuracy with at least 1%. This is to say that we estimate that on a normal Romanian text (containing the diacritical characters) the performance of our system would have been better. The English training data was tagged by ***Eric Gaussier, warmly acknowledged here***. As the tagsets used for the two languages in the parallel training corpus were very different (Multext-East for Romanian, PennTreeBank for English) we defined a tagset mapping and translated the tagging of the English part into a tagging closer to the Romanian one. Based on the training data (both Romanian and English texts), tagged with similar tagsets, we built the language models used for the test data alignment.

POS-preserving translation equivalence is a too restrictive condition for the present task and we defined a meta-tagset, common for both languages that considered frequent POS alternations. For instance, the verb, noun and adjective tags, in both languages were prefixed with a common symbol, given that verb-adjective, noun-verb, noun-adjective and the other combinations are typical for Romanian-English translation equivalents that do not preserve the POS. Such symbols, used as meta-categories, prefixed the proper tags so this procedure produced no loss of information. With these prefixes, the initial algorithm for extracting POS-preserving translation equivalents could be used without any further modifications. The net effect of the meta-categories was that the search space for translation equivalence was much enlarged, allowing for taking into account statistical evidence for pairs of words not considered before because of un-matching POS. Using the tag-prefixes seems to be a good idea not only for legitimate POS-alternating translations, but also for overcoming some typical tagging errors, such as participles versus adjectives.

The last preprocessing phase is encoding the corpus in a XCES-Align-ana format as used in the MULTEXT-EAST corpus (see http://nl.ijs.si/ME/V2/) which is the standard input for the TREQ translation equivalents extraction program. Since the description of TREQ is extensively given elsewhere, we will not go into further details, except of saying that the resulted translation dictionary extracted from the training data contains 49283 entries (lemma-form). The filtering of the translation equivalents candidates (Tufiş and Barbu, 2002) was based on the log-likelihood and the cognate scores with a threshold value set to 15 and 0,43 respectively.

3 The TREQ-AL linking program

This program takes as input the dictionary created by TREQ and the parallel text to be word-aligned. The alignment procedure is a greedy one and considers the aligned translation units independent of the other translation units in the parallel corpus. It has 4 steps: left-to-right pre-alignment; right-to-left adjustment of the pre-alignment; determining alignment zones and filtering them out; the word-alignment inside the alignment zones.

3.1 The left-to-right pre-alignment

For each sentence-alignment unit, this step scans the words from the first to the last in the source-language part (Romanian). The considered word is initially linked to all the words in the target-language part (English) of the current sentence-alignment unit, which are found in the translation dictionary as potential translations. If for the source word no translations are identified in the target part of the translation unit, the control advances to the next source word. The cognate score and the relative distance are decision criteria to choose among the possible links. When consecutive words in the source part are associated with consecutive or close to each other words in the target part, these are taken as forming an "*alignment chain*" and, out of the possible links, are considered those that correspond to the densest grouping of words in each language. High cognate scores in an alignment chain reinforce the alignment. One should note that at the end of this step it is possible to have 1-to-many association links if multiple translations of one or more source words are found in the target part of the current translation unit (and, obviously, they satisfy the selection criteria). The next phase, using a competitive linking approach (Melamed, 2002) aims at selecting the most likely links for this very case.

3.2 The right-to-left adjustment of the pre-alignment

This step tries to correct the pre-alignment errors (when possible) and makes a 1-1 choice in case of the 1-m links generated before. The alignment chains (found in the previous step) are given the highest priority in alignment disambiguation. That is, if for one word in the source language there are several alignment possibilities, the one that belongs to an alignment chain is always selected. Then, if among the competing alignments one has a cognate score higher than the others then this is the preferred one (this heuristics is particularly useful in case of several proper names occurring in the same translation unit). Finally, the relative position of words in the competing links is taken into account to minimize the distance between the surrounding already aligned words.

The first two phases result in a 1-1 word mapping. The next two steps use general linguistic knowledge trying to align the words that remain unaligned (either due to no translation equivalents or because of failure to meet the alignment criteria) after the previous steps. This could result in n-m word alignments, but also in unlinking two previously linked words since a wrong translation pair existing in the extracted dictionary might license a wrong link.

3.3 Alignment zones and filtering suspicious links out

An alignment zone (in our approach) is a piece of text that begins with a conjunction, a preposition, or a punctuation mark and ends with the token preceding the next conjunction, preposition, punctuation or end of sentence. A source-language alignment zone is mapped to one or more target-language alignment zones via the links assigned in the previous steps (based on the translation equivalents). One has to note that the mapping of the alignment zones is not symmetric. An alignment zone that contains no link is called a virgin zone.

In most of the cases the words in the source alignment zone (starting zone) are linked to words in the target alignment zone/s (ending zone/s). The links with either side outside the alignment zones are suspicious and they are deleted. This filtering proved to be almost 100% correct in case the outlier resides in a zone non-adjacent to the starting or ending zones. The failures of this filtering were in the majority of cases due to a wrong use of punctuation in one or the other part of the translation unit (such as omitted comma, a comma between the subject and predicate).

3.4 The word-alignment inside the alignment zones

For each un-linked word in the starting zone the algorithm looks for a word in the ending zone/s of the same category (not meta-category). If such a mapping was not possible, the algorithm tries to link the source word to a target word of the same meta-category, thus resulting in a cross-POS alignment. The possible meta-category mappings are specified by the user in an external mapping file. Any word in the source or target languages that is not assigned a link after the four processing steps described above is automatically assigned a null link.

4 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. Re-tokenizing the text meant splitting compounds and contracted future forms and gluing together the previously split negative contracted forms (*do+n't=don't*). Although the re-tokenization was a post-processing phase, transparent for the task itself, it was a source of missing some links for the negative contracted forms. In our linking the English "n't" was always linked to the Romanian negation and the English auxiliary/modal plus the main verb were linked to the Romanian translation equivalent found for the main verb. Some multi-word expressions recognized by the tokenizer as one token, such as dates (*25 Ianuarie, 2001*), compound prepositions (*de la*, *pina la*), conjunctions (*pentru ca*, *de cind*, *pina cind*) or adverbs (*de jur imprejur*, *in fata*) 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. If the English word was also a compound the number of links generated for one aligned multiword expression was equal to the N*M, where N represented the number of words in the source compound and M the number of words in the target compound.

5 Evaluation

The results of the evaluation of TREQ-AL performance are shown in Table 1 and Table 2. For the evaluation we used two experimental settings. In the first one, TREQ started with an empty lexicon and extracted the translation equivalent from the training data, while in the second experiment TREQ started with a medium size lexicon extracted from our incipient (11,000 synsets) Romanian Wordnet linked to the Princeton Wordnet. This initial lexicon contains about 40,000 entries. The figures in the first and second columns of the Table 1 and Table 2 are those considered by the official evaluation. The last column contains the evaluation of the result that was our main target. Since

TREQ-AL produces only "sure" links, AER (alignment error rate - see the Shared Task web-page for further details) reduces to 1 - *F-measure*.

Surprisingly enough the performances of TREQ-AL are not significantly better (as we expected) when TREQ starts with a seed lexicon. The present version of the system (which is still under development) works quite well on the non-null assignments. Moreover if we consider only the content words (main categories: noun, verbs, adjectives and general adverbs), which are the most relevant with respect to our immediate goals (multilingual wordnets interlinking and word sense disambiguation), TREQ-AL performances are even better than shown in Table 1 and Table 2 (where we considered all the words). The slight reduction of the TREQ-AL lexicon performances might be surprising: the explanation is that the initial bilingual lexicon contained most of the words that TREQ would discover anyway and that the rest of the additionally extracted entries were rare events with lower probability to be found otherwise.

	Non-null links only	**Null links included**	**Dictionary entries**
Precision	84.43%	65.58%	86.68%
Recall	64.34%	66.08%	81.96%
F-measure	73.03%	65.83%	84.26%
AER	26.97%	34.17%	

Table 1. Evaluation results WITHOUT a seed lexicon

	Non-null links only	**Null links included**	**Dictionary entries**
Precision	84.72%	66.07%	86.56%
Recall	64.73%	66.43%	81.85%
F-measure	73.39%	66.25%	84.14%
AER	26.61%	33.75%	

Table 2. Evaluation results WITH a seed lexicon

6 Conclusions and further work

TREQ-AL was developed in a short period of time and is not completely tested and debugged. At the time of writing we already noticed two errors that were responsible for several wrong or missed links. There are also some conceptual limitations which, when removed, are likely to further improve the performance. For instance all the words in virgin alignment zones are automatically given null links but the algorithm could be modified to assign all the links in the Cartesian product of the words in the corresponding virgin zones. The typical example for such a case is represented by the idiomatic expressions (*tanda pe manda = the list that sum up*). A bilingual dictionary of idioms as an external resource certainly would significantly improve the results. Also, with an additional preprocessing phase, for collocation recognition, many missing links could be recovered. At present only those collocations that represent 1-2 or 2-1 alignments are recovered.

A major improvement will be to make the algorithm symmetric. There are many cases when reversing the source and target languages new links can be established. This can be explained by different polysemy degrees of the translation equivalent words and the way we associate alignment zones.

The word order in Romanian and English to some extent is similar, but in the present version of TREQ-AL this is not explicitly used. One obvious and easy improvement of TREQ-AL performance would be to take advantage of the similarity in word order and map the virgin zones and afterwards, the words in the virgin zones.

Finally, we noticed in the gold standard some wrong alignments. One example is the following:
"… a XI – a …" = "… eleventh…"
Our program aligned all the 4 tokens in Romanian (a, XI, –, a) to the English token (eleventh), while the gold standard assigned only "XI" to "eleventh" and the other three Romanian tokens were given a null link. We also noticed some very hard to achieve alignments (anaphoric links).

7 References

Tufiş, D. Barbu, A.M.: „Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing", in *International Journal of Speech Technology*. Kluwer Academic Publishers, no.5, pp.199-209, 2002.

Tufiş, D. "A cheap and fast way to build useful translation lexicons" in *Proceedings of the 19th International Conference on Computational Linguistics, COLING2002,* Taipei, 25-30 August, 2002, pp. 1030-1036p.

Ide, N., Erjavec, T., Tufis, D.: „Sense Discrimination with Parallel Corpora" in Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. ACL2002, July Philadelphia 2002, pp. 56-60.

Stamou, S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M.. "BALKANET A Multilingual Semantic Network for the Balkan Languages", in Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002.

Tufiş, D. "Tiered Tagging and Combined Classifiers" In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999, pp. 28-33.