A Word Alignment System based on a Translation Equivalence Extractor

Ana-Maria Barbu

Romanian Academy Center for Artificial Intelligence
abarbu@racai.ro

Abstract

This paper describes a word alignment system (TREQ-AL), which uses the lexicon extracted with a translation equivalence extractor (TREQ) from a training corpus, including the text of test. The improvement methods applied since the previous version TREQ-AL are one of the paper's focus, as well as the recalled types of the system.

1. Introduction

The work this paper relies on was roughly developed in the shared task organized as part of HLT/NAACL 2003 workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond" (Mihalcea & Peterson, 2003). The task consisted in finding correspondences between words and phrases in the parallel texts of a sentence-aligned Romanian-English corpus. At that moment our system had got good results (comparable with the results of the other participating systems) in terms of precision (P), recall (R) and f-measure (F), namely 81.29%, 60.26% and 69.21%, respectively (for non-null alignments) (Tufiş, Barbu & Ion, 2003). In the meantime, we have developed the component called TREQ-AL (i.e. TRanslation-EQuivalence ALigner) of our system and managed to significantly improve the results, which now amount to P=85.08%, R=65.36%, F=73.93%. The improvements are mainly obtained by linguistic methods and they prove that combining statistics with linguistics could lead to better results than a purely statistic approach (Dejean et al., 2003), or a purely linguistic one (Zao &Vogel, 2003).

The structure of the paper is the following. The first section describes the general system with its processing steps. Next, the improving methods brought to the system are emphasized, in order to explain the growth in performance from one version of the system to the other. The two last sections are devoted to general evaluations and conclusions.

2. TREQ-AL System

TREQ-AL has as input the lexicon extracted by the TREQ algorithm (Tufiş & Barbu, 2002) and the parallel text to be aligned at the word level. The text needs to be tokenized, sentence aligned and morpho-syntactically annotated. For the shared task, the translation equivalence extractor (TREQ) was previously applied to an one-million training parallel corpus, which contained novel and news-paper texts, including the text to be aligned.

The alignment is expressed in word-position terms, that is, the words are represented by their position in the translation unit (i.e. sentence-alignment unit), separately for each language. For instance, given the simplified parallel text below, expressed by lemmas and with words numbered, TREQ-AL should produce the indicated list of assignments (where '-1' is a null alignment and means that the corresponding word is not translated):

RO: 0>de_exemplu 1>, 2>ca 3>istoric 4>treptow 5>fi 6>american 7>.

EN: 0>that 1>historian 2>treptow 3>be 4>an 5>American 6>, 7>for_example 8>.

word alignment: (0 7), (1 6), (2 0), (3 1), (4 2), (5 3), (6 5), (7 8), (-1 4)

In order to get such results, TREQ-AL goes through the following processing steps reiterated with each translation unit.

2.1 Dictionary Looking-up

First, for each word in the source language (here Romanian), TREQ-AL looks for the appropriate translation equivalent(s) into the TREQ lexicon. For those words that are not found in the lexicon, the system searches cognates among not assigned target words. The looking-up is done regardless of the part-of-speech, in order to avoid tagging errors.

This step results in a list of (possibly non-consecutive) positions of those source words for which one or more translation equivalents were found.

2.2 Up-bottom alignment

The next step after dictionary looking-up is the up-bottom (or left-to-right) alignment, which processes the text in its normal reading sense. The target of this step is to do primary assignments and to coarsely solve the translation ambiguity.

Choosing a target word w_j from an ambiguity list depends on three factors: the cognate status (*cog*), the positional distance to the previous assignment (*pad*), and the relative distance to the source position (*spd*). So, a target position *j* wins if it gets the best (in general, minimal) value for one of these dimensions:

$$j \Leftrightarrow \min \{cog, pad, spd\}$$

In the cases of non-ambiguity, the unique translation equivalent represents the proper link.

Note that at the end of this step, a target position can be assigned to more than one source position if it satisfies the selection criteria.

Another important task fulfilled at this step is the detecting of *alignment chains*, that is, sequences of at least four consecutive words in the source part associated with consecutive or close to each other words in the target part. These chains are of great confidence in the further word alignment process.

2.3 Bottom-up alignment

This step tries to refine and correct the primary assignation. It achieves a bottom-up (or right-to-left)

alignment, i.e. in contrary sense to the reading one, and takes into account more information than the previous step. The alignment criterion is a function depending on the following data:

- the distance to the lower assignment;
- the distances to the upper two assignments;
- the distance between source positions (especially relevant in cases of gaps);
- the alignment chains;
- the precedence constraint (presented later).

The result is a strict one-to-one word mapping, which can reflect modifications or even deletions of the links in the previous step, if no translation equivalent satisfies the alignment criterion. Note that this criterion affects both the ambiguous and non-ambiguous positions.

The next two steps use general linguistic knowledge for aligning the words that remain unaligned because there is no translation equivalent for them or the existing one(s) missed the alignment criterion.

2.4 Alignment zones

The system delimitates and count off, in each part of the translation unit, contiguous pieces of text that begin with a conjunction, a preposition or a punctuation mark and end with the token preceding the next conjunction, preposition, punctuation or end of the sentence. These are used as alignment zones in that they are mapped from one language to the other via the links assigned in the previous steps. That helps to filter out the links that exhibit aberrant zone mapping, for instance if the source words in zone 1 are aligned with target words in zones 2, 7 and again 2, then the link inducing the mapping with the zone 7 is deleted. It should be said that it is possible to get some unmapped zones, namely those which contain no aligned words.

2.5 The final word-alignment

Now, the algorithm looks for aligning un-linked words inside the zones mapped at the previous step. First, the words of the same part-of-speech are aligned and then the system tries to do cross-part-of-speech or multiple alignments according to some general or language-specific rules.

For an unmapped zone, the search space for new alignments is that between the closest links on the sides of that zone.

Any word in each language that has not been aligned after these processing steps is automatically assigned a null link.

An example for the alignment process is given in *Table 1*. Each word in the texts is expressed by its lemma and its position in the sentence. The example was especially chosen for illustrating how the process can face translation "discontinuities" and what the kind of the recalled alignments is. Translation "discontinuities" refer to the fact that different constituents in the two languages do not follow the same order inside their sentences. For instance, on the gold standard alignment one can see that the first Romanian words are aligned with English words in the middle of the sentence and so on. Even if our algorithm missed the start alignments, it could recover the other constituent-order differences. This proves that it can manage such differences pretty well. On the other hand, there is, in this example, a discontinuous colloquial Romanian idiom: *i-o fi apucat ... dragul de* translated with the English phrase: *they grow fond of*. Note that the algorithm aligned the two last words of the idioms: (*dragul fond*) and (*de of*) but missed the main verbs (*apucat grow*) and the pronouns (*i- they*). However, we think it is important that for these verbs the system assigned null-alignments, because this leaves the possibility of further recovering.

By evaluating the alignment results of this example, we get P=61.53% and R=57.14%. However, if we take into account that the non-null alignments are the most informative, then the precision and the recall for such

RO: 0>el, 1>el 2>fi 3>apuca 4>pe 5>politist 6>si 7>pe 8>procuror 9>drag 10>de 11>treptow 12>de 13>avea 14>adopta 15>un 16>asemenea 17>atitudine 18>?

EN: 0>could 1>it 2>be 3>that 4>the 5>police 6>and 7>the 8>prosecutor 9>adopt 10>that 11>attitude 12>as 13>they 14>grow 15>fond 16>of 17>treptow 18>?

Gold-standard alignments: (0 13)(1 14)(2 14)(3 14)(4 4)(4 5)(5 4)(5 5)(6 6)(7 7)(7 8)(8 7)(8 8)(9 15)(10 16)(11 17)(12 -1)(13 9)(14 9)(15 11)(16 10)(17 11)(18 18)(-1 0)(-1 1)(-1 2)(-1 3)(-1 12)

The alignment-line structure: ROposition ENposition ROword /ENword1-ENposition1/... (* marks cognates)

0 1	el	/it-1/they-13
1 1	el	/it-1/they-13
2 2	fi	/be-2
4 16	pe	/of-16
5 5*	politist	/police-5*
6 6	si	/that-3/and-6/that-10
7 16	pe	/of-16
8 8*	procuror	/prosecutor-8*
9 15	drag	/fond-15
10 16	de	/of-16
11 17*	treptow	/treptow-17*
12 -1	de	/of-16
13 2	avea	/be-2
14 9*	adopta	/adopt-9*
15 10	un	/that-3/that-10
16 11	asemenea	/could-0/attitude-11
17 11*	atitudine	/adopt-9/attitude-11*
18 18*	?	/?-18*

Dictionary looking-up & Up-bottom alignment

0 -1	el	/it-1/they-13
1 1	el	/it-1/they-13
2 2	fi	/be-2
4 -1	pe	/of-16
5 5*	politist	/police-5*
6 6	si	/that-3/and-6/that-10
7 -1	pe	/of-16
8 8	procuror	/prosecutor-8*
9 15	drag	/fond-15
10 16	de	/of-16
11 17*	treptow	/treptow-17*
12 -1	de	/of-16
13 -1	avea	/be-2
14 9*	adopta	/adopt-9*
15 10	un	/that-3/that-10
16 -1	asemenea	/could-0/attitude-11
17 11*	atitudine	/adopt-9/attitude-11*
18 18*	?	/?-18*

Bottom-up alignment

(0 –1)	el,pp	??
(1 1)	el,pp	it,p
(2 2)	fi,va	be,vm
(3 –1)	apuca,vm	??
(4 –1)	pe,s	??
(5 5)(5 4)	politist,n	police,n
(6 6)	si,c	and,c
(7 –1)	pe,s	??
(8 8)(8 7)	procuror,n	prosecutor,n
(9 15)	drag,n	fond,a
(10 16)	de,s	of,s
(11 17)	treptow,n	treptow,n
(12 –1)	de,s	??
(13 9)	avea,va	adopt,vm
(14 9)	adopta,vm	adopt,vm
(15 10)	un,ti	that,di
(16 –1)	asemenea,a	??
(17 11)	atitudine,n	attitude,n
(18 18)	?,b	?,b

Final word-alignment

Table 1: Alignment Example

alignments become P=86.6% and R=59%. At this point it is worth discussing the nature of the recall. The way the gold standard has been built is controversial especially because of its cartesian products applied to so-called multiword translations. For instance, if one considers that the Romanian phrase '*pe procurori*' (where *pe* is a case marker preposition and the noun is unarticled) is translated with the English phrase '*the prosecutors*', the gold standard records the following alignments: (*pe the*) (*pe prosecutors*) (*procurori the*) (*procurori prosecutors*). It is obvious that only the alignment (*procurori prosecutors*) is really informative (e.g. for bilingual lexicon or even terminology extraction) but it represents only 25% from the corresponding recall. Therefore the recall of 59% obtained for nonnull-alignments in our example is not at all relevant for our algorithm's pretty high power of extracting informative pairs. Nevertheless we do not deny that for machine translation the phrase alignment could be more important than the word-to-word one and we assume the obtained figures as such and the challenge they imply, as well.

3. Improvement Methods

In this section we describe the ways we have got the f-measure growth from 69.21% to 73.81% for the TREQ-AL algorithm. In principal, the improvements are of linguistic order and they refer to cognates, precedence constraints, pair assignments and language-specific rules.

3.1 Cognates

Cognates are words with similar phonetic body and the same meaning in different languages. At first glance, there are cognates only in related languages. Nevertheless there are many factors determining the existence of cognates in unrelated ones. For instance, there are historical factors (see Latin words in English) and economical ones, which induce terminology migration between languages. But, first of all, translations use cognates for proper nouns, in completely unrelated languages, for which, if they use different alphabets, one can use transliteration mechanisms, as Melamed (2000) suggests in his approach. Our lexicon extractor, TREQ, looks for cognates, but not all of them pass the statistical score in order to be extracted as translation equivalences. That is why the word-to-word alignment algorithm, TREQ-AL, applies, on its turn, the cognate detection, by calculating the LCS score (Hunt & Szymansky, 1977) for each pair not found in the lexicon, except for the functional ones (prepositions, conjunctions etc.). A problem we had to solve was to set up the minimal limit for declaring words to be cognates. That has to be done so that the alignments gaps left by TREQ be filled with as many as possible correct equivalences. A too high threshold leaves many gaps unsolved, while a too low one induces alignment errors. From our experiments, we got the optimal cognate limit of 0.65. Without doubt, this limit depends on the language pair and the text nature. For illustration, in Table 2, we give some Romanian-English pairs of cognates and their corresponding scores.

RO-word	EN-word	Cognate-score
organizatie	organisation	0.90
patetism	pathetic	0.80
nesofisticat	unsophisticated	0.76
dezinforma	misinform	0.66
insuportabil	insufferable	0.60

Table2: Examples of cognates

3.2 Precedence constraints

As the example above shows, the most ambiguous words with respect to their translation equivalences are the functional ones, such as prepositions, conjunctions, determiners. These are also the most frequent. Therefore, choosing the appropriate translation of a functional word from its ambiguity list is an important step, because an error at this level triggers errors in other points of the sentence. In order to help the disambiguation process we set on the precedence constraint, relying on the general linguistic fact that certain parts-of-speech always precede others. For instance, prepositions, articles, determiners, even conjunctions precede nouns, adjectives, adverbs and sometimes verbs (especially participles).

At the first reading of the bilingual text, for each functional word of this kind, the position(s) of the subsequent noun, adjective, adverb or verb is (are) memorized. There results positional sequences of syntactically related words. That simulates somehow a chunker task for prepositional and noun phrases and exploits the fact that conjunctions (either coordinating or subordinating ones) always precede conjuncts. Afterwards, at the level of bottom-up inspection, functional-words positions are with priority aligned if at least one pair in their corresponding sequences was aligned

3.3 Pair assignments

This linguistic assumption applies at the level of finding alignments which translation equivalence dictionary says nothing about, that is, at final alignment stage. It consists in taking pairs of consecutive parts-of-speech depending, in some extent, on each other. For instance, given the Romanian POS-structure: *s a n* and the English one: *s* n_1 n_2, we assume that if there is the alignment (*n* n_2), then (*a* n_1) holds, as well. Let the following parallel text be:

RO: … 8>intru 9>un 10>desant 11>judiciar 12>, …
EN: … 5>in 6>a 7>judiciary 8>raid …

The algorithm has already aligned the adjectives (11 7) (that is, *judiciar* with *judiciary*). By applying the pair assignment assumption it also aligns the nouns (10 8) (that is, *desant* with *raid*).

Our experiments show that there are groups of parts-of-speech with high degree of cohesion. So for example, prepositions, articles, nouns, and adjectives form such a group, but also adverbs, prepositions and conjunctions, or verbs, particles and conjunctions.

3.4 Language-specific rules

Besides these assumptions, we have applied some language-specific rules concerning Romanian versus English syntax particularities or cross-linguistic differences in part-of-speech mapping. The latter refers, for instance, to nouns, adjectives and verbs translated with

each other, to mood/tense verbal particles or auxiliaries differing from one language to the other, to articles and determiners. Among syntax particularities we can mention that English structures '*noun*$_1$ *noun*$_2$' are often translated into Romanian as '*noun*$_2$ de *noun*$_1$'.

The table below illustrates the contribution of each such improvement. The figures show how the considered measures vary if we disable the respective method in turn.

Linguistic method	Precision [%]	Recall [%]	F-m. [%]
Cognates	-0.42	-0.64	-0.57
Precedence constraint	-1.92	-0.43	-1.00
Pair alignments	-0.02	-0.49	-0.32
Lang.-spec. rules	+1.08	-3.04	-1.60

Table 3: Methods contributions

As one can see, the language-specific rules bring the most important contribution to the f-measure value. By deactivating this module the precision grows indeed but the recall decreases dramatically and the f-measure, too. It turns out that this is a necessary module and, as a general conclusion, that a language-oriented algorithm could be better than a general one.

4. General Evaluation

After previously giving partial evaluation on a single translation unit, this section offers a general hint about the recalled types with respect to the used gold standard (GS) and the whole text. The test parallel corpus contains texts from novels and newspapers, consisting of 248 sentences expressed in 10816 words and punctuation marks.

Alignments	GS	TREQ-AL	
	Nr. of links	Prec. [%]	Recall [%]
Total	7149	**66.89**	**67.06**
Null	954	30.67	78.09
Same POS	3959	87.95	81.15
Cross-POS	2236	75.36	49.01
Non-null	6195	**85.08**	**65.36**

Table 4: TREQ-AL's general evaluation

Table 4 shows the high power of the algorithm as to alignments of the same part-of-speech, while there are poorer results regarding the cross-part-of-speech ones. However, remember that many of such alignments come from multi-word expressions and they are not at all relevant for dictionary building if they are taken by themselves.

The figures in Table 4 could be more eloquent if it is taken into account that null-alignments made by TREQ-AL represent 33.83% from total, while those in GS only 13.34%. This mainly happens because of low power of the algorithm for detecting multi-word expressions. However, it paves the way for further methods of finding them.

5. Conclusions and Further Work

The system fully exploits the performances of the translation equivalents extractor, which basically uses a statistic approach. Then, the system applies its positional searches for accomplishing the word-to-word alignment. On the other hand, the linguistic methods presented here prove that linguistics can help statistics in improving the alignment results. More over, it turns out that a language-oriented algorithm can be better than purely statistic or too general ones.

Our algorithm is actually not finished. Recovering multi-word expressions is still a challenging task to be done. However, at this stage, the system is a valuable resource for building bilingual lexica and, to some extent, for terminology extraction.

References

Hunt, J.W. & Szymanski, T.G. (1977). A Fast Algorithm for Computing Longest Common Subsequences. *Comunications of the ACM*, 20(5), 350-353.

Melamed, D. (2000). Pattern recognition for mapping bitext correspondence. In J. Véronis (ed.) Parallel Text Processing. Alignment and Use of Translation Corpora (pp. 25-47). Kluwer Academic Publishers.

Mihalcea, R. & Pedersen, T. (2003). An Evaluation Exercise for Word Alignment. In Proceedings of the HLT-NAACL 2003 Workshop on *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, (pp. 1-10), Edmonton, Canada.

Tufiş, D. & Barbu, A.M. (2002). Revealing Translators' Knowledge: Statistical Methods in Constructing Practical Translation Lexicons for Language and Speech Processing. *International Journal of Speech Technology*, 5, 199-209.

Tufis, D., Barbu, A.M., and Ion, R. (2003). TREQ-AL: A word alignement system with limited language resources. In Proceedings of the HLT-NAACL 2003 Workshop on *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, (pp. 36-39), Edmonton, Canada.

Zhao, B., and Vogel, S. (2003). "Word Alignment Based on Bilingual Bracketing", in Proceedings of the HLT-NAACL 2003 Workshop on *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, (pp. 15-18), Edmonton, Canada.

Dejean, H., Gaussier, E., Goutte, C., and Yamanda, K. (2003). "Reducing Parameter Space for Word Alignment" in Proceedings of the HLT-NAACL 2003 Workshop on *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, (pp. 23-26), Edmonton, Canada.