Automatic Extraction of Translation Equivalents From Parallel Corpora

Dan Tufiş, Ana-Maria Barbu
Romanian Academy RACAI, 13, "13 Septembrie", 74311 Bucharest, Romania,
(tufis,abarbu}@racai.ro

Abstract. This paper presents a simple and effective method for extraction of translation equivalents from parallel corpora. Experiments were conducted on Orwell's "1984" parallel corpus with translations available in six CEE languages, all of them being aligned to the English original. There were extracted six bilingual lexicons X-English (En), where X stands for one of Czech (Cz), Bulgarian (Bg), Estonian (Et), Hungarian (Hu), Romanian (Ro) or Slovene (Si) and a multilingual one En/Cz/Bg/Et/Hu/Ro/Si providing translation equivalents for English words in all other 6 languages. We provide the evaluation of the results for part of the language pairs involved in the experiment. The paper ends by drawing some conclusions and discussing further work.

Key words: alignment, , parallel corpora, POS tagging, translation equivalents

1. Introduction

Extracting bilingual dictionaries from corpora can be seen as a very fine-grained alignment process, were the aligned units are not paragraphs or sentences but words and phrases. There are many statistical approaches to build translation lexica from bilingual texts, roughly falling into two categories: *the hypotheses testing* approach such as (Gale and Church, 1991: 152:157), (Smadja et all, 1996: 1-38) etc. and *the estimating* approach (Brown et all, 1993:467-469), (Kupiec, 1993:19-22), (Hiemstra, 1997:21-26) etc. We opted for a hypotheses testing method by first generating a list of translation equivalent candidates (TECs) and then iteratively extracting the most likely translation-equivalence pairs (TEPs). The translation equivalents extraction process does not rely on a pre-existing bilingual lexicon for the considered languages. The candidate list is constructed from the translation/alignment units (TU). That is to say that the translation of an item in a source language sentence is looked for only in the alignment corresponding sentence(s) of the target language.

The underlying assumptions we rely on (as many others do) are the following:

- a lexical token in one half of the TU corresponds to at most one lexical unit in the other half of the TU;
- a lexical token is one word or a multiple word expression and the proper identification of the multi-word tokens is ascribed to the segmentation preliminary phase;

- a lexical token in one part of a TU can be aligned to a lexical token in the other part of the TU only if the two tokens have compatible types (part-of-speech); in most cases, compatibility reduces to the same POS, but it is also possible to define compatibility mappings (e.g. participles in one language mapping to adjectives or nouns in the other language).
- although the word order is not an invariant of translation, it is not random either; candidate translation pairs that contains words which are closer in relative position are preferred.

2. The baseline and the iterative algorithm

Based on the alignment, the first step is to compute a list of translation equivalent candidates (TECL). This list contains several sub-lists, one for each POS considered in the extraction procedure. Each POS-specific sub-list contains several pairs of tokens <source_language_token : target_language_token> of the corresponding POS that appeared in the same TUs. These pairs (translation equivalents candidates-TECs) are generated by a Cartesian product of the set of tokens (of the given POS) in one half of the TU with the set of tokens (of the same POS) in the other half. Each pair has attached the number of occurrences of the respective association throughout all the TUs.

The baseline algorithm is represented by a chi-square test applied to each translation equivalent candidate(TEC). By counting the number of TUs in which the current TEC <T_S T_T> appeared (n_{11}), the number of TUs in which the source language token appeared with any other token but T_T (n_{12}) , the number of TUs in which the target language token appeared with any other token but T_S (n_{21}), and the TU in which neither T_S nor T_T appeared, one could build a 2*2 contingency table as in figure 1, where:

	T_T	$\neg T_T$	
T_S	n_{11}	n_{12}	n_{1*}
$\neg T_S$	n_{21}	n_{22}	n_{2*}
	n_{*1}	n_{*2}	n_{**}

$n_{1*} = n_{11} + n_{12}$, $n_{2*} = n_{21} + n_{22}$

$n_{*1} = n_{11} + n_{21}$, $n_{*2} = n_{12} + n_{22}$

. $n_{**} = n_{1*} + n_{2*} = n_{*1} + n_{*2}$

Figure 1: Contingency table for a translation equivalent candidate <T_S T_T>

Chi-square coefficients, given by the formula $\chi^2 = \frac{n_{**}(n_{11} * n_{22} - n_{12} * n_{21})^2}{(n_{11} + n_{12}) * (n_{11} + n_{21}) * (n_{21} + n_{22}) * (n_{21} + n_{22})}$ may be used to select the most likely candidates as TEPs. For a 99.9% confidence level, the threshold condition for selection would be $\chi^2 > 10.83$. One could use also a minimal number of occurrences for <T_S T_T> (usually, 3). This baseline algorithm may be enhanced in many ways (using a dictionary of already extracted TEPs for

eliminating generation of spurious TECs, stop-word lists, considering token string similarity a.s.o.). An algorithm with such extensions (plus a few more) is described in (Gale, Church 1991:152-157). In spite of being extremely simple, this algorithm was reported to provide impressive results (Canadian Hansard, precision about 98% and recall about 50%). However the response time is not among its assets and it is not clear how or whether different translations of the same item (because for instance of different lexicalisations in the other language of the multiple meanings) are extracted.

The iterative algorithm we propose is also very simple but significantly faster than the baseline algorithm and addresses the multiple translations of an item in a straightforward manner. It can be enhanced in many ways (including those discussed above). It has some similarities to the iterative algorithm presented in (Ahrenberg et all. 1998: 29-35) but unlike it, our algorithm avoids computing various probabilities (or better said probability estimates) and scores (t-score). The algorithm gets as input the aligned parallel corpus and the maximum number of iterations. At each iteration step, the pairs that pass the selection (see below) will be removed from TECL so that this list is shortened after each step and eventually may be emptied. Based on TECL, for each POS is constructed a contingency table (TBLk) as shown in Figure 2:

	T_{T1}	…	T_{Tn}	
T_{S1}	n_{11}	…	n_{1n}	n_{1*}
…	…	…	…	…
T_{Sm}	n_{m1}	…	n_{mn}	n_{m*}
	n_{*1}	…	n_{*n}	n_{**}

Figure 2: Contingency table with counts for TECs at step *K*

The rows of the table are indexed by the distinct source tokens and the columns are indexed by the distinct target tokens (of the same POS). Each cell (i,j) contains the number of occurrences in TECL of the $<T_{Si}, T_{Tj}>$ TEC:

$$n_{ij} = occ(T_{Si},T_{Tj}); \; n_{i*} = \sum_{j=1}^{n} n_{ij} \; ; \; n_{*j} = \sum_{i=1}^{m} n_{ij} \; ; \text{ and } n_{**} = \sum_{j=1}^{n} (\sum_{i=1}^{m} n_{ij})$$

The selection condition is expressed by the equation:

$$\text{EQ1:} \quad TP^{k} = \{< T_{Si}; T_{Tj} > \mid \forall p,q \; (n_{ij} \geq n_{iq}) \wedge (n_{ij} \geq n_{pj})\}$$

This is the key idea of the extraction algorithm and it expresses the requirement that in order to select a TEC $<T_{Si}, T_{Tj}>$ as a translation equivalence pair, the number of its occurrences must be a local frequency maximum, or put it otherwise, the number of associations of T_{Si} with T_{Tj} must be higher than (or at least equal to) any other T_{Tp} ($p \neq j$). The same holds for the other way around. If T_{Si} is translated in more than one way (either because of

having multiple meanings that are lexicalised in the second language by different words, or because use in the target language of various synonyms for T_{Tj}) the rest of translations will be found in subsequent steps (if frequent enough). The most used translation of a token T_{Si} will be found first. The basic algorithm is sketched below:

```
procedure lex-extract(Al_Par_Corpus, step) is:
 k=0;
 TECL(k)=build-cand(Al_Par_Corpus)
 for each POS in TECL with k=k+1 do TBLk=build_the_TEC_table(TECL(k-1));
                                    TP(k)= select(TBLk);  ## EQ1 ##
                                    TECL(k)=delete (TP(k), TECL(k-1));
  until {(TECL(k) is empty) or(TP(k) is empty) or (k > step)}
end
```

3. Experiments and results

We conducted experiments on the "1984" multilingual corpus (Dimitrova et.all, 1998: 315-319) containing 6 translations of the English original. This corpus was developed within the Multext-East project, published on a CD-ROM (Erjavec et all. 1998) and recently improved within the CONCEDE project (to be soon publicly available at CONCEDE's homepage: www.itri.brighton.ac.uk/projects/concede/). Each monolingual part of the corpus (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene) was tokenised, lemmatised, tagged and sentence aligned to the English hub. The table in Figure 3 provides quantitative data about each language pair (XX-En) from the "1984" parallel corpus. In Figure 4 there are shown the number of lemmas in each monolingual part of the multilingual corpus as well as the number of lemmas that occurred more than twice.

Lang. Pair	# Alignment Units (AUs)	1:1 AUs (%)	1:2 AUs (%)	2:1 AUs (%)	Other AUs (%)
Bg-En	6699	98.42	1.08	0.34	0,16
Cz-En	6656	96.75	1.22	1.54	0.49
Et-En	6607	97.42	1.51	0.9	0.15
Hu-En	6669	97.15	0.58	1.52	0,75
Ro-En	6340	94.04	4.02	1.32	0.62
Si-En	6680	98.38	0.79	0.71	0.12

Figure 3:The "1984" parallel corpus overview

Language	Bulgarian	Czech	English	Estonian	Hungarian	Romanian	Slovene
No. of wordforms	15093	17659	9192	16811	19250	14023	16402
No. of lemmas	8225	8677	6871	8403	9729	6626	7157
No.of >2-occ lemma*	3350	3329	2916	2876	3294	3052	3189

Figure 4:The lemmatised monolingual "1984" overview
* the number of lemmas does not include interjections, particles, residuals)

For validation purposes we set the step limit of the algorithm to 4. The table in Figure 5 shows the results and their evaluation for those languages where we found voluntary native speakers with good command of English. The extracted bilingual lexicons are available at http://www.racai.ro/bi-lex/. The precision (Prec) was computed as usual, i.e. the number of correct TEPs divided by the total number of extracted TEPs while the recall (Rec*) was computed as the number of correct TEPs divided by the number of lemmas in the source language with more than 3 occurrences. When the (usual) threshold of minimal 3 occurrences is considered, the algorithm provides a high precision and a good recall. The evaluation was fully done for Estonian, Hungarian and Romanian and partially for Slovene (the first step was fully evaluated while from the rest were evaluated randomly selected pairs). As one can see from the figures in Table 5, the precision is higher than 98% for Romanian and Slovene almost 97% for Hungarian and more than 96% for Estonian. The recall (our defined Rec*) ranges from 50.92% (Slovene) to 63.90% (Estonian). We run the extractor for the Ro-En bitext without imposing a step limit. The program stopped after 25 steps with a number of 2765 extracted pairs, out of which 113 were wrong. The precision decreased to 95,91%, but the recall (Rec*) significantly improved: 86,89%.

Language	Bg-En Prec/Rec*	Cz-En Prec/Rec*	Et-En Prec/Rec*	Hu-En Prec/Rec*	Ro-En Prec/Rec*	Sl-En Prec/Rec*
Step 1	1336 NA/NA	1399 NA/NA	1216 99.50/42.07	1299 98.61/38.88	1394 99.71/42.74	1177 99.91/36.87
Step 2	1741 NA/NA	1886 NA/NA	1617 97.89/55.04	1737 97.63/51.48	1867 99.30/52.23	1489 99.52/46.47
Step 3	1896 NA/NA	2085 NA/NA	1807 96.63/60.84	1863 96.99/54.85	2067 99.03/54.84	1589 99.06/49.63
Step 4	1986 NA/NA	2188 NA/NA	1911 96.18/63.90	1935 96.89/56.92	2182 98.57/56.36	1646 98.66/50.92

Figure 5: The results after 4 iteration steps and partial evaluation

In an initial version of this algorithm we used a chi-square test (as in the baseline algorithm) to check the selected TEPs. However, as the selection condition (EQ1) is very powerful, the vast majority of the selected TEPs passed the chi-square test and therefore we eliminated it. This is certainly one of the reasons for the speed of our extraction algorithm.

From the 6 bilingual lexicons we also derived a 7-language lexicon (2862 entries), with English as a search hub. As more than half of the English words had equivalents only in 2 or three languages, we considered only those entries for which our algorithm found translations in all but at most one of the other 6 languages. This filtered multilingual lexicon contains 1237 entries and can be found at the same site as the bilingual lexicons. A typical

entry in this multilingual lexicon is given below (in Figure 6 the multiword dictionary entry is exemplified by using each language character set; in the actual file there are used SGML entities).

En	Bg	Cs	Et	Hu	Ro	Sl
cold	студен	studený/chladný	külm	hideg	rece/friguros	mrzel/hladen

Figure 6: An entry from the extracted multilingual lexicon

One interesting aspect of our algorithm is that the words that are found as translations of one word *in the same iteration step* are very likely to be a multiword translation of the respective single language word (such as the Estonian "armastusministeerium" = ministry & love). The additional condition for identifying such a particular case of multiword translation (both collocates sharing the same part of speech) is that the candidate words must co-occur in the same translation units. If this condition does not hold, then it simply happened that two different translations of the same word were equally used. Translations of the same word that are found in different steps are lexicalisations of different senses or synonyms (mare=big; mare=large; mare=vast; mare=important) or homographs (mare=sea) of the source word. The order in which are discovered the different translations of the same word is related to the frequency one target word is used in the current corpus as a translation for a source word and not necessary to the prevalence of a given sense over the other (although this might be also the case). On the other hand, if a word is (justifiably) translated by N distinct words, in general the algorithm would need at least N iteration steps in order to find all the N TEPs. However, if a specific word w_S appears in the corpus more than the specified threshold value, it is not necessary that our algorithm will find a translation for it. This happen when w_S is translated in different sentences by different words and none of the pairing is frequent enough to meet the threshold frequency condition. For instance, when processing the RO-EN bitext of "1984" parallel corpus, there were extracted 10 correct TEPs for "mare" (big, great, large, vast, sea, long, main, thick, general, important) but none of them would have been found unless each pair appeared in TECL more than twice.
From the results shown in the table in Figure 5 one can notice that most part of bilingual lexicons is extracted in the first step (between 63% and 71%). It is intuitive to see that if one would like to consider even rare occurrences of translation-pairs, the lowering of the frequency threshold should be done in the last step(s) and not from the very beginning: this way, what can be quite safely extracted would not interfere with the noise introduced by a much larger search space; also, the correct rare TEP will survive (or at least most of them) to the continuous shortening of the TECL.

4. Implementation

The extraction program is written in Perl and runs under practically any platform (Perl implementations exists not only for UNIX/LINUX but also for Windows, and MACOS). The table in Figure 6 shows the running time for each bitext in the "1984" parallel corpus. The program was run under LINUX on a Pentium III/600Mhz with 96 MB RAM.

A quite similar approach to ours (also implemented in Perl) is presented in (Ahrenberg et all., 1998:29-35) and (Ahrenberg et all., 2000:97-116) and for a novel of about half length of Orwell's "1984" their algorithm needed 55 minutes on a Ultrasparc1 Workstation with 320 MB RAM and the best results reported are 96.7% precision and 54.6% recall. For a computer manual containing about 45% more token than our corpus, their algorithm needed 4,5 hours with the best results being 85,6% precision and 67,1% recall.

Language	Bg-En	Cz-En	Et-En	Hu-En	Ro-En		Si-En
					7 steps	25 steps	
Extraction time (sec)	181	148	139	220	183	415	157

Figure 7:Extraction time for each of the bilingual lexicons

5. Conclusions and further work

We presented a simple but very effective algorithm for extracting bilingual lexicons, based on a 1:1 mapping hypothesis. We showed that in case a language specific tokenizer able to recognize and "pack" the compounds is responsible for preprocessing the input to the extractor the 1:1 mapping approach is not a limitation anymore. If the compounds cannot be dealt with in the segmentation pre-processing phase one may consider either extending the bilingual lexicon extractor's model to an N:M paradigm or consider using a monolingual tool as a pre-processor for recognizing the compounds. We are currently considering both options. For the first one we started the implementation of a collocation extraction based on incorporating S. Banerjee's and T.Pedersen's, BSP (Bigram Statistics Package). For the second option, we are carrying out some preliminary experiments with a slightly modified version of the program presented in this paper. Conceptually, the modified version of the program can be seen as receiving the same text as source and target input file with all the sentence alignments being 1:1. Two additional modifications are:

- the TECL must not include pairs made of identical strings.; this condition is necessary for limiting the search space to the only potential collocations

- the POS condition is removed; this restriction is not necessary anymore since most sequences of words that should be translated as one unit are not characterized by the same POS.

A new version (C++) of the algorithm described in this paper, incorporating, BSP, is under construction.

Acknowledgements. Special thanks are due to Heiki Haalep, Csaba Oravecs, and Tomaz Erjavec for the validation of the Et-En, Hu-En and Si-En extracted dictionaries.

References

Ahrenberg, L., M. Andersson, M. Merkel. 1998. "A simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts", in *Proceedings of COLING'98, Montreal*, 29-35

Ahrenberg, L., M. Andersson, M. Merkel. 2000. "A knowledge-lite approach to word alignment", in J.Véronis (ed.) *Parallel Text Processing*. Text, Speech and Language Technology Series, Kluwer Academic Publishers, Vol. 13, 2000 2000: 97-116)

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

Dimitrova, L, T. Erjavec, N. Ide, H. Kaalep, V. Petkevic and D. Tufiş. 1998. " Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and East European Languages" in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics(COLING'98)*, Montreal, Canada, 315-319.

Gale, W.A. and K.W. Church. 1991. "Identifying word correspondences in parallel texts". In Fourth *DARPA Workshop on Speech and Natural Language*, pp. 152-157

Erjavec T., Lawson A., Romary L.. 1998. *East Meet West: A Compendium of Multilingual Resources.*TELRI-MULTEXT EAST CD-ROM, 1998, ISBN: 3-922641-46-6.

Hiemstra, D. 1997. "Deriving a bilingual lexicon for cross language information retrieval". In *Proceedings of Gronics* 21-26

Kupiec, J. 1993. "An algorithm for finding noun phrase correspondences in bilingual corpora". In *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, 17:22

Smadja, F., K.R. McKeown, and V. Hatzivassiloglou. 1996. "Translating collocations for bilingual lexicons: A statistical approach". *Computational Linguistics*, 22(1):1:38