# Accurate automatic extraction of translation equivalents from parallel corpora

Dan Tufiş and Ana-Maria Barbu
RACAI-Romanian Academy Center for Artificial Intelligence
13, "13 Septembrie", RO-74311, Bucharest, 5, Romania
{tufis, abarbu}@racai.ro

## Abstract

The paper describes a simple but very effective approach to bilingual lexicons extraction from parallel corpora. After briefly describing the method we present the evaluation for six pairs of languages in terms of precision and recall and processing time. We conclude by discussing the merits and the drawbacks of out method in comparison with other works and comment on further developments.

## 1. Introduction

The vast and continuously growing amount of information available nowadays on the WEB poses numerous and challenging problems which strongly influenced current approaches to natural language processing. One of the most obvious tendencies in NLP and related technologies, is a distinct preference for *shallow processing*. The term usually implies not only a partial coverage of the difficult problems in computational linguistics but also, accepting a limited imprecision or inaccuracy in automatic decisions made in order to achieve a useful degree of language analysis or generation. The motivation for this trend, to a large extent imposed by the Internet industry and market, is given by the requirement to process in real time very large quantity of texts. It has been noticed that many useful tasks (document indexing, document classification, information retrieval, etc) may be achieved when using even a quite superficial processing of natural language texts. Increasing the amount of linguistic processing in more and more web-based applications is a constant preoccupation of many R&D professionals and companies. The development of large linguistic resources (thesauri, ontologies, translation memories, multilingual dictionaries, annotated corpora etc.) became a very active area both for academic research and industry. As one would expect, printed dictionaries and lexicons were primary sources for constructing lexical databases. However, relying on human-dedicated lexical sources was not as successful as expected for computer programs due to various reasons among which one could mention variation of lexical stock and lexical gaps, shifts of meaning, sense granularity to fine for automatic discrimination, etc. The costs involved in turning the human-oriented linguistic knowledge sources into useful computational resources are also high (copyright fees, man-power costs) so the idea of using computers to extract and organise linguistic information according to specific needs is natural and significant research is carried along these lines. Combining supervised and unsupervised linguistic knowledge acquisition is a trade-off between time and costs versus quality and completeness. As basic language resources (meant for training /learning) are cleaner and larger, the acquisition of more structured information and higher levels of linguistic knowledge is possible with less human supervision and with simpler computational means. We showed (Tufiş, 2000) for instance that the accuracy of tagging new texts can go higher than 99% when very clean training corpora are used and language modelling is adequate with respect to the underlying technology. For this example (based on HMMs), we argued (Tufiş et. all 2000) that preserving in the language model distinctions that cannot be reliably made by a specific distributional analysis model is source of noise and the cause of performance deterioration. The tagging technology is more or less the same in many projects but the accuracy of the results varies (even for the same language) and the explanations can be found both in the quality of the training data and (mainly) in the quality of the tagsets.

Aligning parallel texts is a very good example where simple techniques and limited linguistic knowledge (if any at all) can ensure surprisingly good results for the problem of interest. Since *charAlign* (Gale, Church 1993) was published, many variants refinements and improvements of this program were implemented, but the basic underlying ideas remained extremely simple and thus easy to implement and fast in operation.

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. Most approaches to this problem rely on statistical means to build translation lexica from bilingual texts, roughly falling into two categories: *the hypotheses testing* approach such as (Gale, Church 1991), (Smadja et all 1996) etc. and *the estimating* approach (Brown et all 1993), (Kupiec 1993), (Hiemstra 1997) etc. The first approach involves a generative device that produces a list of translation equivalence candidates (TECs), each of them being subject to an independence statistical test. The TECs that pass the test are assumed to be translation-equivalence pairs (TEPs). The second approach assumes building from data a statistical model the parameters of which are to be estimated according to a given set of assumptions. There are pros and cons for each type of approach, some of them discussed in  (Hiemstra 1997).

Our method hardly fit into one of these two categories, but is closer in spirit to hypotheses testing approach, by first generating a list of translation equivalent candidates and then iteratively extracting the most likely translation-equivalence pairs. 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.

## 2. Format of the input data

The translation equivalents extraction process does not assume a pre-existing bilingual lexicon for the considered languages. If such a lexicon exists, it might be used for the validation purposes, but also for some kind of “land-marking” for increasing the certainty in building the TECs.

Three files physically represent the input data:

- the source language file, containing one half of the aligned bi-text $T_S$
- the target language file, containing the other half of the aligned bi-text $T_T$
- the alignment index for the sentences in $T_S$ and $T_L$.

Each part of the bi-text is represented in the same tabular format with one tagged and lemmatised token per line and sentence mark-up explicitly shown. Table 1 shows excerpts from the contents of the three files.

The first column in the Source/Target file represents the type of the item (lexical token, left/right split, punctuation etc), the second column contains the word-form while the last column contains the lemma, the morpho-syntactic description of the word-form and the POS only tag used in the alignment (separated by a “\”).The SGML alignment file is a cesAlign type of document and specifies the sentence pairing in the two languages by means of sentence identifiers in the first two files. An alignment unit defines a translation unit (TU) as composed by the sentences identified by the IDs.

| **Source file** | **Target file** | **Alignment file** |
|---|---|---|
| <S FROM=Oen.1.1.1.1> | <S FROM=“Oro.1.2.2.1”> | <!doctype cesAlign PUBLIC”-//CES//DTD cesAlign//EN”[]> |
| TOK It it\Pp3ns\P | LSPLIT Într- =\Spsay\S | <linkList id=“oroen”> |
| TOK was be\Vmis3s\AUX | TOK o un\Tifsr\T | <linkGrp id=“oroen.1” type=“body” targtype=“s” domains=“oro oen”> |
| TOK a a\Di\D | TOK zi =\Ncfsrn\N | <link xtargets=“ Oro.1.2.2.1 ; Oen.1.1.1.1 Oen.1.1.1.2 “> |
| TOK bright bright\Af\A | TOK senină senin\Afpfsrn\A | <link xtargets=“ Oro.1.2.3.1 ; Oen.1.1.2.1 “> |
| … | … | … |

**Table 1: The format of the input data**

For instance, in Table 1, the first TU is made of the sentence identified by the id “Oro1.2.2.1” (*Într-o zi senină…* ) and the two sentences identified by the ID “Oen.1.1.1.1”  (*It was a bright…*) and “Oen.1.1.1.2” (*Winston Smith, his chin…*).

The format of the input data is conformant with different conventions adopted within the MULTEXT-EAST (MTE) project (which developed the “1984” multilingual corpus we worked with). Various pre-processing steps (tokenisation, alignment and tagging) were initially achieved by using tools developed within the MULTEXT project, for which MTE was a follow up. Because of interest in this corpus, it was extended with new languages within the TELRI project and further cleaned up within the CONCEDE project. We provide more details on the corpus in the “Experiments and results” section.

## 3. 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. Obviously if the parallel text contains alignment and tagging errors, several real translation equivalents would not be found because they will not be member of the corresponding $TEC^{pos}$. Each POS-specific sub-list contains several pairs of tokens $<token_{LANG1}:token_{LANG2}>$ 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 retaining from the list of TECs only those pairs which cannot be considered to occur just by chance. This hypothesis can be tested by different statistical tests (we used S. Banerjee’s and T. Pedersen’s Bigram Statistics Package, which includes chi-square, dice, mutual information and log-likelihood statistical measures). For instance, when chi-square test is used, the coefficients (computed for each TEC) 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 rejecting the null hypothesis ($T_S$ and $T_T$ co-occurred by chance) would be $\lambda^2 > 10.83$. One could use also a minimal number of occurrences for $<T_S\ T_T>$ (usually this is 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 etc.). An algorithm with such extensions (plus a few more) is described in (Gale, Church 1991). 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 are extracted.

The iterative algorithm we propose is also very simple but significantly faster than the baseline. It can be enhanced in many ways (including those discussed above). 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 Table 2:

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

**Table 2: TBLk - 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: $n_{ij} = occ(T_{Si},T_{Tj})$; $n_{i*} = \sum_{j=1}^{n} n_{ij}$ ; $n_{*j} = \sum_{i=1}^{m} n_{ij}$ ; and $n_{**} = \sum_{j=1}^{n}(\sum_{i=1}^{m} n_{ij})$ . The selection condition is expressed by the equation: (1) $TP^k = \{< T_{Si}; T_{Tj} > | \forall p, q\ (n_{ij} \geq n_{iq}) \wedge (n_{ij} \geq n_{pj})\}$.

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, at step *k*, the number of associations of $T_{Si}$ with $T_{Tj}$ must be higher than (or at least equal to) any other $T_{Tp}$ ($p \neq j$) that are represented in the TBLk. The same holds for the other way around. If $T_{Si}$ is translated in more than one way, 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.

## 4. Experiments and results

We conducted experiments on one of the few publicly available multilingual aligned corpora, namely the "1984" multilingual corpus (Dimitrova et al 1998) containing 6 translations of the English original. This corpus was developed within the Multext-East project, published on a CD-ROM (Erjavec et al 1998) and recently improved within the CONCEDE project (to be soon released to the research community 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.

| 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 lemmas* | 3350 | 3329 | 2916 | 2876 | 3294 | 3052 | 3189 |

**Table 3:The lemmatised monolingual "1984" overview**

* the number of lemmas does not include interjections, particles, residuals)

The number of lemmas in each monolingual part of the multilingual corpus as well as the number of lemmas that occurred more than twice are shown in Table 3.

For validation purposes we set the step limit of the algorithm to 4. 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).

The evaluators judged a translation pair as correct, partially correct or incorrect. Pairs that were judged as partially correct (all cases appeared as expected in Hungarian and Estonian) corresponded to English multiword equivalents of the source words. For instance the Estonian word "armastusministeerium" means "ministry of love", but our algorithm found as translation equivalents (armastusministeerium = ministry) and (armastusministeerium = love). For all such cases we used the BSP package to test whether the correct multiword equivalent pass a collocation test and if so, we included the partially correct TEP into the class of correct pairs. Such a decision might be motivated on the grounds that a preliminary collocations analysis and recognition would have raised this issue.

The figures in Table 4 show the results and the evaluation. In the following the notion of correctness of a pair is taken in the above-mentioned interpretation.

The extracted bilingual lexicons are available at http://www.racai.ro/bi-lex/. The precision (Prec) was computed as the number of correct TEPs divided by the total number of extracted TEPs. 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.

As one can see from the figures in Table 4, 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 significantly improved: 86,89%.

We should mention that Rec*, as we compute it, is slightly different from the usual recall, because Rec* is reporting the percentage of the number of correct translations found for the lemmas (occurring more than the threshold) in the source language. Let us assume that in the source language there are N different lemmas occurring more than the preset threshold and the program found M correct translation equivalents. Further, let us assume that each lemma is used, on average, with S different senses each one occurring more than the set threshold. Then, our Rec* will be M/N when it should return M/N*S. As we specified before, different translations of the same lemma are found, usually, in subsequent steps. Since we set the number of iteration steps to 4, only for a few words (those very frequent) there would be found multiple valid translations. That is to say, that Rec ≈ Rec*. However, when the number of iteration steps is increased, Rec* becomes an overestimation of Rec.

In an initial version of this algorithm we used a chi-square test (as in the baseline algorithm) to check the selected TEPs. Since the selection condition (EQ1) is very powerful, the vast majority of the selected TEPs passed the chi-square test while many pairs that used to pass the chi-square threshold did not pass the condition (EQ1) and therefore we eliminated the supplementary and time consuming statistical tests. This is certainly one of the reasons for the speed (see next section) of our extraction algorithm.

If one source word has different translations in the target language (either lexicalisations of different senses of a polysemous source word or different synonyms for the target word), in general they are found, if frequent enough, in different iteration steps. 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.

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

**Table 4: The results after 4 iteration steps and partial evaluation**

From the results shown in the Table 4 one can notice that most part of bilingual lexicons is extracted in the first step (between 63% and 71%). We tried to extract translation equivalents even for words that appeared in the source language only twice. Because of validation reasons we did this experiment only for the Ro-En bitext. The experiment considered an extra-iteration step, after the last one, with the occurrence threshold set to two. The extracted number of pairs (1311) was more than half of the pairs extracted in the previous steps altogether, but also the number of erroneous pairs (297) was almost 10 times higher than the previous number of errors (31). However, when lowering the occurrence threshold from the first step, the number of correct pairs was almost the same (3174 versus 3165) but the number of errors was significantly higher (417 versus 328) with a global error rate increase from 9,38% to 11,61%.

## 5. 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 5 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.

| Language | Bg-En | Cz-En | Et-En | Hu-En | Ro-En | | Si-En |
|---|---|---|---|---|---|---|---|
| | | | | | 4 steps | 25 steps | |
| Time (sec) | 181 | 148 | 139 | 220 | 183 | 415 | 157 |

**Table 5 :Extraction time (in seconds) for each of the bilingual lexicons**

A quite similar approach to ours (also implemented in Perl) is presented in (Ahrenberg et all 1998) and (Ahrenberg et all 2000). The languages considered by their experiments are English and Swedish. For a novel of about half the 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. Unlike us, they don’t rely on tagging and lemmatisation, although they use a “morphology” module that achieves some kind of stemming and grouping of inflectional variants. This strategy makes it possible to link low-frequency source expressions belonging to the same suffix paradigm. The obvious advantage of not using POS categorisation is that their approach would be able to identify TEPs where the POS of the source token is different from the POS of the target token.

An explanation of the much better response time in our case, besides not using statistical tests (they use t-test), is that the search space in our case is probably several orders of magnitude smaller.

## 6. 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 tokeniser able to recognise and “pack” the compounds is responsible for pre-processing the input to the extractor the 1:1 mapping approach is not a limitation anymore. The MULTEXT tokeniser, for instance allows for recognition of generic multiword expressions (dates, literally expressed numbers) or specific ones based on external resources containing lists of compounds, proper names, idiomatic expressions. 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 recognising the compounds. We are currently considering both options. For the first one we started the implementation of a new tokeniser including collocation recognition. As most of the multiword tokens (found by the collocation extraction module) are not expected to be in the lexicon, a post-tagging module will check (based on very simple chunking grammars) whether the assigned tag is compatible with the tags recorded in the lexicon for the constituent (known) words. 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 characterised by the same POS.

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

## Acknowledgements

The research reported here started as an AUPELF/UREF co-operation project, coordinated by Patrick Paroubeck of LIMSI/CNRS (CADBFR).

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

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

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

Dimitrova L, Erjavec T, IdeN, Kaalep H, Petkevic V, Tufiş D 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 ACL and 17th COLING International Conference*, Montreal, pp 315-319.

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

Gale W, Church K 1993 A program for aligning sentences in bilingual corpus. *Computational Linguistics* 19(1): 75-102.

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

Tufiş D 2000 Using a Large Set of Eagles-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. In *Proceedings of the Second Conference on Language Resources and Evaluation*, Athens, pp.1105-1112.

Tufiş D, Dienes P, Oravecz C, Váradi T 2000 Principled Hidden Tagset Design for Tiered Tagging of Hungarian. In *Proceedings of the Second Conference on Language Resources and Evaluation*, Athens, 1421-1426.