A cheap and fast way to build useful translation lexicons

Dan TUFIS
Romanian Academy Centre for Artificial Intelligence
13, “13 Septembrie”
Bucharest, ROMANIA, RO-74311
tufis@racai.ro

Abstract

The paper presents a statistical approach to automatic building of translation lexicons from parallel corpora. We briefly describe the pre-processing steps, a baseline iterative method, and the actual algorithm. The evaluation for the two algorithms is presented in some detail in terms of precision, recall and processing time. We conclude by briefly presenting some of our applications of the multilingual lexicons extracted by the method described herein.

Introduction

The scientific and technological advancement in many domains is a constant source of new term coinage and therefore keeping up with multilingual lexicography in such areas is very difficult unless computational means are used. Translation lexicons, based on *translation equivalence* relation are lexical knowledge sources, which can be extracted from parallel texts (even from comparable texts), with very limited human resources. The translation lexicons appear to be quite different from the corresponding printed lexicons, meant for the human users. There are well known reasons for these differences and we will not discuss the issue here, but exactly these differences make them very useful (in spite of inherent noise content) in many computer-based applications. We will discuss some of our experiments based on automatically extracted multilingual lexicons.

Most modern approaches to automatic extraction of translation equivalents rely on statistical techniques and roughly fall into two categories. The *hypotheses-testing* methods such as Gale and Church (1991), Smadja et al. (1996), Tiedmann (1998), Ahrenberg (2000), Melamed (2001) etc. use a generative device that produces a list of translation equivalence candidates (TECs), extracted from corresponding segments of the parallel texts (translation units-TU), each of them being subject to an independence statistical test. The TECs that show an association measure higher than expected under the independence assumption are assumed to be translation-equivalence pairs (TEPs). The TEPs are extracted independently one of another and therefore the process might be characterized as a local maximization (greedy) one. The *estimating* approach such as Brown et al. (1993), Kay and Röscheisen (1993), Kupiec (1993), Hiemstra (1997) etc. is based on building from data a statistical bitext model, the parameters of which are to be estimated according to a given set of assumptions. The bitext model allows for global maximization of the translation equivalence relation, considering not individual translation equivalents but sets of translation equivalents (sometimes called *assignments*). There are pros and cons for each type of approach, some of them discussed in Hiemstra (1997).

Our translation equivalents extraction process may be characterized as a “*hypotheses testing*” approach and does not need a pre-existing bilingual lexicon for the considered languages. If such a lexicon exists it can be used to eliminate spurious candidate translation equivalence pairs and thus to speed up the process and increase its accuracy.

1 Assumptions, preprocessing and a baseline

There are several underlying assumptions one may consider in keeping the computational complexity of a translation lexicon extraction algorithm as low as possible. None of these hopotheses is true in general, but the situations where they are not observed are rare enough so that ignoring the exceptions would not produce a significant number of errors and would not lose too many useful translations. The assumptions we made were the following:

- a lexical token in one half of the translation unit (TU) corresponds to at most one non-empty lexical unit in the other half of the TU; this is the 1:1 mapping assumption which underlines the work of many other researchers (Ahrenberg et al (2000), Brew and McKelvie (1996), Hiemstra (1996), Kay and Röscheisen (1993), Tiedmann (1998), Melamed (2001) etc);
- a polysemous lexical token, if used several times in the same TU, is used with the same meaning; this assumption is explicitly used by Gale and Church (1991), Melamed (2001) and implicitly by all the previously mentioned authors;
- 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 other compatibility mappings (e.g. participles or gerunds in English are quite often translated as adjectives or nouns in Romanian and vice-versa);
- although the word order is not an invariant of translation, it is not random either (Ahrenberg et al (2000)); when two or more candidate translation pairs are equally scored, the one containing tokens which are closer in relative position are preferred.

The proper extraction of translation equivalents requires special pre-processing:

- *sentence alignment*; we used a slightly modified version of CharAlign described by Gale and Church (1993) .
- *tokenization*; the segmenter we used (MtSeg, developed by P. di Cristo for the MULTEXT project: http://www.lpl.univ-aix.fr/projects/multext/MtSeg/), may process multiword expressions as single lexical tokens. The segmenter comes with tokenization resources for several Western European languages, further enhanced in the MULTEXT-EAST project (Dimitrova et al (1998), Erjavec et al (1998), Tufis et al (1998)) with corresponding resources for Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene.
- *tagging and lemmatization*; we used a tiered tagging with combined language models approach (Tufis (1999, 2000)) based on a Brants's TnT tagger.

After the sentence alignment, tagging and lemmatization, the first step is to compute a list of translation equivalence 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 <token$_S$ token$_T$> of the corresponding POS that appeared in the same TUs. TECL contains a lot of noise and many TECs are very improbable. In order to eliminate much of this noise, the most unlikely candidates are filtered out of TECL. The filtering is based on scoring the association between the tokens in a TEC.

For the ranking of the TECs and their filtering we experimented with 4 scoring functions: MI (*pointwise* mutual information), DICE, LL (loglikelihood), and χ^2 (chi-square). After empirical tests we decided to use LL test with the threshold value set to 9.

Our baseline algorithm, BASE, is a very simple iterative algorithm, very fast and can be enhanced in many ways. It has some similarities to the iterative algorithm presented in Ahrenberg et al (2000) but unlike it, our algorithm avoids computing various probabilities (or better said probability estimates) and scores (t-score). 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 a S_m*T_n contingency table is constructed, with S_m the number of token types in the first part of the bitext (call it source) and T_n the number of token types in the other part of the bitext (call it target). Source token types index the rows of the table and the target token types (of the same POS) index the columns. Each cell (i,j) contains the number of occurrences in TECL of the <T_{Si}, T_{Tj}> TEC. Equations below express the selection condition:

$$(1)\ TP^k = \{< T_{Si}\ T_{Tj} > | \forall p,q\ (n_{ij} \geq n_{iq}) \wedge (n_{ij} \geq n_{pj})\} \&$$

$$(2)\qquad n_{ij} \geq 3$$

This is the key idea of the iterative extraction algorithm: it expresses the requirement that in order to select a TEC <T_{Si}, T_{Tj}> as a translation equivalence pair, 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. All the pairs selected in TP^k are removed (the respective counts are zeroed). 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. The equation (2) represents a frequency relevance threshold, necessary in order to diminish the influence of data sparseness.

2 An improved algorithm (BETA)

One of the main deficiencies of the BASE algorithm is that it is sensitive to what Melamed (2001) calls indirect associations. If <T_{Si}, T_{Tj}> has a high association score and T_{Tj} collocates with T_{Tk}, it might very well happen that <T_{Si}, T_{Tk}> gets also a high association score. Although, as observed by Melamed (2001), in general, the indirect associations have lower scores than the direct (correct) associations, they could receive higher scores than many correct pairs and this will not only generate wrong translation equivalents, but will eliminate from further considerations several correct pairs, deteriorating the procedure’s recall. To weaken this sensitivity, the BASE algorithm had to impose that the number of occurrences of a TEC be at least 3, thus filtering out more than 50% of all the possible TECs. Still, because of the indirect association effect, in spite of a very good precision (more than 98%) out of the considered pairs another approximately 50% correct pairs were missed. The BASE algorithm has this deficiency because it looks on the association scores globally, and does not check within the TUs if the tokens making the indirect association are still there.

To diminish the influence of the indirect associations and consequently removing the frequency threshold, we modified the BASE algorithm so that the maximum score is not considered globally but within each of the TUs. This brings BETA closer to the competitive linking algorithm described in Melamed (2001). The competing pairs are only the TECs generated from the current TU and the one with the best score is the first selected. Based on the 1:1 mapping hypothesis, any TEC containing the tokens in the winning pair are discarded. Then, the next best scored TEC in the current TU is selected and again the remaining pairs that include one of the two tokens in the selected pair are discarded. The multiple-step control in BASE, where each TU was scanned several times (equal to the number of iteration steps) is not necessary anymore. The BETA algorithm will see each TU unit only once but the TU is processed until no further TEPs can be reliably extracted or TU is emptied. This modification improves both the precision and recall in comparison with the BASE algorithm. In accordance with the 1:1 mapping hypothesis, when two or more TEC pairs of the same TU share the same token and they are equally scored, the algorithm has to make a decision and choose only one of them. If there exists a seed lexicon and one of the competitors is in this lexicon it will be the winner. Otherwise, decision is made based on two heuristics: string similarity scoring and relative distance.

The similarity measure we used, COGN(T_S, T_T), is very similar to the **XXDICE** score described in Brew and McKevie (1996).

The threshold for the COGN(T_S, T_T) test was empirically set to 0.42. This value depends on the pair of languages in the considered bitext. The actual implementation of the COGN test considers a language dependent normalization step, which strips some suffixes, discards the diacritics and reduces some consonant doubling etc. This normalization step was hand written, but, based on available lists of cognates, it could be automatically induced.

The second filtering condition, DIST(T_S, T_T) is based on the difference between the relative positions in the TU of the T_S and T_T respectively. The threshold for the DIST(T_S, T_T) was set to 2.

The COGN(T_S, T_T) filter is stronger than DIST(T_S, T_T), so that the TEC with the highest similarity score is the preferred one. If the similarity score is irrelevant, the weaker filter DIST(T_S, T_T) gives priority to the pairs with the smallest relative distance between the constituent tokens.

3 BASE and BETA Evaluations

We conducted experiments on 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. The newer version is distributed by TRACTOR (www.tractor.de).

Each monolingual part of the corpus (Bulgarian-Bg, Czech-Cz, Estonian-Et, Hungarian-Hu, Romanian-Ro and Slovene-Si) was tokenized, lemmatized, tagged and sentence aligned to the English hub.

The evaluation protocol specified that all the translation pairs be judged in context, so that if one pair is found to be correct in at least one context, then it should be judged as correct. The evaluation was done for both BASE and BETA algorithms but on different scales. The BASE algorithm was run on all the 6 bitexts with the English hub and native speakers of the second language in the bitexts (with good command of English) validated 4 of the 6 bilingual lexicons. The lexicons contained all parts of speech defined in the MULTEXT-EAST lexicon specifications (Erjavec et al (1998)) except for interjections, particles and residuals.

The BETA algorithm was ran on the Romanian-English bitext, but at the time of this writing the evaluation was finalized only for the nominal translation pairs.

3.1 BASE Evaluation

For validation purposes we limited the number of iteration steps to 4. The extracted dictionaries contain adjectives (A), conjunctions (C), determiners (D), numerals (M), nouns (N), pronouns (P), adverbs (R), prepositions (S) and verbs (V). The precision (**Prec**) was computed as the number of correct TEPs divided by the total number of extracted TEPs. The recall (considered for the non-English language in the bitext) was computed two ways: the first one, ***Rec******,** which took into account only the tokens processed by the algorithm (those that appeared at least three times). The second one, **Rec**, took into account all the tokens irrespective of their frequency counts. ***Rec**** is defined as the number of source lemma types in the correct TEPs divided by the number of lemma types in the source language with at least 3 occurrences. **Rec** is defined as the number of source lemma types in the correct TEPs divided by the number of lemma types in the source language.

The rationale for showing ***Rec**** is to estimate the proportion of the missed considered tokens. This might be of interest when precision is of utmost importance. When the threshold of minimal 3 occurrences is considered, the algorithm provides a high precision and a good recall (***Rec****). The evaluation was fully done for Et, Hu and Ro and partially for Si (the first step was fully evaluated while the rest were evaluated from randomly selected pairs).

The results after 4 iteration steps are shown in the table below for the Et-En, Hu-En, Ro-En and Si-En lexicons. From the 6 bilingual lexicons we also derived a 7-language lexicon (2862 entries), with English as a search hub (see http://www.racai.ro/~tufis/BilingualLexicons/AutomaticallyExtractedBilingualLexicons.html).

	Et-En	Hu-En	Ro-En	Si-En
Entries	1911	1935	2227	1646
Prec/Rec*	96.2/57.9	96.9/56.9	98.4/58.8	98.7/57.9
Rec	*18.8*	*19.3*	*25.2*	*22.7*

Table 1: BASE evaluation for all POS and 4 iteration steps

To facilitate the comparison with the evaluation of the BETA algorithm we ran the BASE algorithm for extracting the noun translation pairs from the Romanian-English bitext. The noun extraction had the second worst accuracy (the worst was the adverb), and therefore we considered that an in-depth evaluation of this case would be more informative than a global evaluation. We set no limit for the number of steps and lowered the occurrence threshold to 2. The program stopped after 10 steps with a number of 1900 extracted translation pairs, out of which 126 were wrong. Compared with the 4 steps run the precision decreased to 93.4%, but both ***Rec**** (70.1%) and **Rec** (39.8%) improved.

In the 10-step run of the BASE algorithm, the extracted noun pairs covered 85.83% of the nouns in the Romanian part of the bitext.

We should mention that in spite of the general practice in computing recall for bilingual lexicon extraction task (be it ***Rec****** or **Rec**) this is only an approximation of the real recall. The reason for this approximation is that in order to compute the real recall one should have a gold standard with all the words aligned by human evaluators. In general such a gold standard bitext is not available and the recall is either approximated as above, or is evaluated on a small sample and the result is taken to be more or less true for all the bitext.

3.2 BETA Evaluation

The BETA algorithm preserves the simplicity of the BASE algorithm but it significantly improves its recall (**Rec**) at the expense of some loss in precision (**Prec**). Its evaluation was done for the Romanian-English bitext, without a seed lexicon and only with respect to the lexicon of nouns. The filtering condition in case of ties was the following:

$$\max(\mathrm{COGN}(T^j{}_S,T^j{}_T)\geq 0.4)\vee\min(\mathrm{DIST}(T^j{}_S,T^j{}_T)\leq 2)$$

The results show that the **Rec** (72.7%) almost doubled compared with the best **Rec** obtained by the BASE algorithm for nouns (39.9%). However, the price for these significant improvements was a serious deterioration of the **Prec** (78.3% versus 93.4%).

Noun types in text	3435
No. entries	4023
Correct entries	3149
Types in correct entries	2496
Prec/Rec	78.3/72.7

Table2: BETA evaluation for the Ro-EN lexicon of nouns; both COGN and DIST filters used

The analysis of the wrong translation pairs revealed that most of them were hapax pairs (pairs appearing only once) and they were selected because the DIST measure enabled them, so we considered that this filter is not discriminative enough for hapaxes. On the other hand for the non-hapax pairs the DIST condition was successful in more than 85% of the cases. Therefore, we decided that the additional DIST filtering condition be preserved for non-hapax competitors only.

Although 166 erroneous TEPs were removed, 144 good TEP were lost. **Prec** improved (81.0% versus 78.3%) but **Rec** depreciated (69.0% versus 72.7%).

The BASE algorithm allows for trading off between Prec and Rec by means of the number of iteration steps.

Noun types in text	3435
No. entries	3713
Correct entries	3007
Types in correct entries	2371
Prec/Rec	81.0/69.0

Table3: BETA evaluation for the Ro-EN lexicon of nouns; only COGN filter used

The BETA algorithm allows for similar trading off between Prec and Rec by means of the COGN and DIST thresholds and obviously by means of an occurrence threshold. For instance when BETA was set to ignore the hapax pairs, its Prec was 96.1% (better then the BASE precision 93.4%) Rec* was 96.4% (BASE with 10 iterations had a Rec* of 70.1%) and Rec was 60.0% (BASE with 10 iterations had a Rec of 39.8%).

4 Partial translations

As the alignment model used by the translation equivalence extraction is based on the 1:1 mapping hypothesis, inherently it will find partial translations for those cases where one or more words in one language must be translated by two or more words in the other language. Although we used a tokenizer aware of compounds in the two languages, its resources were obviously partial. In the extracted noun lexicon, the evaluators found 116 partial translations (3.86%). In this section we will discuss one way to recover the correct translations for the partial ones, discovered by our 1:1 mapping-based extraction program.

First, from each part of the bitext a set of possible collocations was extracted by a simple method called “repeated segments” analysis. Any sequence of two or more tokens that appears more than once is retained. Additionally, the tags attached to the words occurring in a repeated segment must observe the syntactic patterns characterizing most of the real collocations. For the noun lexicon we considered only forms of <head-noun (functional_word) modifier> as Romanian patterns and <modifier (functional_word) head-noun> as English patterns. If all the content words contained in a repeated segment have translation equivalents, then the repeated segment is discarded as not being relevant for a partial translation. Otherwise, the repeated segment is stored in the lexicon as a translation for the translation of its head-noun. This simple procedure managed to recover 62 partial translations and improve other 12 (still partial, but better).

5 Implementation

The extraction programs, both BASE and BETA, are written in Perl and run under practically any platform (Perl implementations exist not only for UNIX/LINUX but also for Windows, and MACOS). Although, as one reviewer rightfully noticed, the speed is not really relevant for such an algorithm, evaluation of the current speed shows that, the approach being computationally very cheap, there is room for adding more sophisticated "association functions" without too much concern for the overall response time. Table 4 shows the BASE running time for each bitext in the "1984" parallel corpus (4 steps, all POS considered).

Bitext	Bg-En	Cz-En	Et-En	Hu-En	Ro-En		Si-En
					4 steps	28 steps	
Extraction time (sec)	181	148	139	220	183	415	157

Table 4:BASE extraction time for each of the bilingual lexicons (all POS)

The running time for extraction of the noun Romanian-English lexicon (Cygwin UNIX emulator for Windows on a PII/233Mhz with 96 MB RAM) for BASE was 103 seconds while for BETA was 234 seconds.

A quite similar approach to our BASE algorithm (also implemented in Perl) is presented in Ahrenberg et al (2000) and 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. They used a frequency threshold of 3 and the best results reported are 92.5% precision and 54.6% recall (our ***Rec****). For a computer manual containing about 45% more tokens than our corpus, their algorithm needed 4.5 hours with the best results being 74.94% precision and 67,3% recall (***Rec****).

The BETA algorithm is closer to Melamed's extractor, although our program is greedier and never returns to a visited translation unit.

6 Applications and further work

We used the multilingual lexicon, mentioned before, for a sense discrimination exercise described in Erjavec et al (2001) where the criterion for sense clustering was the way the different occurrences of an English word in the "1984" parallel corpus were translated in the other 6 languages. The experiment carried on involved 91 highly ambiguous English nouns and was extremely encouraging; new results are described in Erjavec&el all (2002).

Another application of the translation lexicons was in the BALKANET project aimed at developing wordnets for Balkan languages, Romanian included. The translation lexicons were used both in building from scratch, but in a harmonized way, the synsets for the base concepts and also for cross-lingual validation on running text (this was again the "1984" novel) of the interlingual index (ILI) mapping of these basic concepts. Considering that 4 languages in the BALKANET are represented in the "1984" parallel corpus we plan to take advantage of the ILI mapping for further refinement of the word-sense discrimination method mentioned above and add cluster labeling. The obvious language independent labeling is based on ILI-record numbers.

The experiments reported here were evaluated on European language. A new experiment has been preliminarily evaluated for an extract of 500 sentences Chinese-English form a parallel corpus of juridical texts. The experiment was focused on noun translations extraction, used an LL-score threshold set to 9 and no conflict resolution method for the competitive translations. We had two result sets:

RS1: contains translations which haven't competitors (that is whenever there were competing translations for the same word none of them was selected)

RS2: differs from DS1 by the inclusion in the output lexicon of all the competing translations.

It is obvious that if 1:1 mapping hypothesis is true, for any competing translations included in RS2 only 1 is correct and all the others are errors. Therefore the precision for RS2 is much less than for RS1.

The results of this experiment are shown in Table 5 and they show that without making a decision on the competing translations we either loose many good translations (RS1) or include a lot of noise (RS2).

Result set	# extr. pairs	precision	recall
RS1	187	93.04%	33.6%
RS2	545	49.9%	98.1%

Table 5: BETA results for CN-EN experiment

Further work will address the issue of defining adequate heuristics for filtering out competing candidates.

Acknowledgements

Our thanks go to Chang Baobao who evaluated the Chinese-English results as well as to Wolfgang Teubert who launched the CN-EN experiment in the framework of the European concerted action TELRI.

References

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

Brew, C., McKelvie, D. (1996) *Word-pair extraction for lexicography* http:///www.ltg.ed.ac.uk/~chrisbr/ papers/nemplap96

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, pp. 263-311.

Dimitrova, L, T. Erjavec, N. Ide, H. Kaalep, V. Petkevic, D. Tufis (1998) *Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and East European Languages.* In Proceedings of COLING, Montreal, Canada, pp. 315-319.

Gale, W., K.W. Church (1991) *Identifying word correspondences in parallel texts*. In Proceedings of the 4th DARPA Workshop on Speech and Natural Language, pp. 152-157.

Gale, W.A., K.W. Church (1993) *A Program for Aligning Sentences in Bilingual Corpora* in Computational Linguistics, 19/1, pp. 75-102

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

Erjavec T., Ide N., Tufis, D.(2001) *Automatic Sense Tagging Using Parallel Corpora*. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, pp. 212-219

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

Kay, M., Röscheisen M. (1993) *Text-Translation Alignment.* In Computational Linguistics, 19/1, pp. 121-142

Kupiec, J.(1993) *An algorithm for finding noun phrase correspondences in bilingual corpora*. In Proceedings of the 31st Annual Meeting of the ACL, pp. 17-22

Melamed, D (2001) *Empitical ethods for Exploiting Parallel Texts.* The MIT Press. Cambridge, Massachusetts, London, England, 195 p.

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

Tiedemann, J. (1998) *Extraction of Translation Equivalents from Parallel Corpora* http://stp.ling.uu.se /~joerg

Tufis, D.(1999) *Tiered Tagging and Combined Classifiers.* In *Text, Speech and Dialogue*, F. Jelinek, E. Nöth (eds), Lecture Notes in Artificial Intelligence 1692, Springer, pp. 29-33

Tufis, D., Ide, N. Erjavec, T (1998) *Standardized Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages*. In Proceedings of LREC, Granada, Spain, pp. 233-240

Tufis, D.(2000) *Using a Large Set of Eagles-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging*. In Proceedings of the LREC, Athens, Greece, pp. 1105 -1112

Tufis, D., Barbu, A. (2001) *Extracting multilingual lexicons from parallel corpora*. In Proceedings of the ACH/ALLC, New York University, pp.122-124

Tufis D., Barbu A.(2001) *Computational Bilingual Lexicography: Automatic Extraction of Translation Dictionaries*. In International Journal on Science and Technology of Information, Romanian Academy, ISSN 1453-8245, 4/3-4, pp.325-352.

Erjavec T., Ide N., Tufis D. (2002) *Sense discrimination with Parallel Corpora*. Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. ACL2002, July, Philadelphia