Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, July 2002, pp. 54-60. Association for Computational Linguistics.

Sense Discrimination with Parallel Corpora

Nancy Ide
Dept. of Computer Science
Vassar College
Poughkeepsie,
New York 12604-0520
USA
ide@cs.vassar.edu

Tomaz Erjavec
Dept. of Intelligent Systems
Institute "Jozef Stefan"
Jamova 39,
SI-1000 Ljubljana
SLOVENIA
tomaz.erjavec@ijs.si

Dan Tufis
RACAI
Romanian Academy
Casa Academiei,
Calea 13 Septembrie 13,
Bucharest 74311, ROMANIA
tufis@racai.ro

Abstract

This paper describes an experiment that uses translation equivalents derived from parallel corpora to determine sense distinctions that can be used for automatic sense-tagging and other disambiguation tasks. Our results show that sense distinctions derived from cross-lingual information are at least as reliable as those made by human annotators. Because our approach is fully automated through all its steps, it could provide means to obtain large samples of "sense-tagged" data without the high cost of human annotation.

1 Introduction

It is well known that the most nagging issue for word sense disambiguation (WSD) is the definition of just what a word sense is. At its base, the problem is a philosophical and linguistic one that is far from being resolved. However, work in automated language processing has led to efforts to find practical means to distinguish word senses, at least to the degree that they are useful for natural language processing tasks such as summarization, document retrieval, and machine translation.

Resnik and Yarowsky (1997) suggest that for the purposes of WSD, the different senses of a word could be determined by considering only sense distinctions that are lexicalized cross-linguistically. In particular, they propose that some set of target languages be identified, and that the sense distinctions to be considered for language processing applications and evaluation be restricted to those that are realized lexically in some minimum subset of those languages. This idea would seem to provide an answer, at least in part, to the problem of determining different senses of a word: intuitively, one assumes that if another language lexicalizes a word in two or more ways, there must be a conceptual motivation. If we look at enough languages, we would be likely to find the significant lexical differences that delimit different senses of a word.

Several studies have used parallel texts for WSD (e.g., Gale *et al.,* 1993; Dagan *et al.,* 1991; Dagan and Itai, 1994) as well as to define semantic properties of and relations among lexemes (Dyvik, 1998). More recently, two studies have examined the use of cross-lingual lexicalization as a criterion for validating sense distinctions: Ide (1999) used translation equivalents derived from aligned versions of Orwell's *Nineteen Eighty-Four* among five languages from four different languages families, while Resnik and Yarowsky (2000) used translations generated by native speakers presented with isolated sentences in English. In both of these studies, translation information was used to validate sense distinctions provided in lexicons such as WordNet (Miller *et al*., 1990). Although the results are promising, especially for coarse-grained sense distinctions, they rest on the acceptance of a previously established set of senses. Given the substantial divergences among sense distinctions in dictionaries and lexicons, together with the ongoing debate within the WSD community concerning which sense distinctions, if any, are appropriate for language processing applications, fitting cross-linguistic information to pre-established sense inventories may not be the optimal approach.

This paper builds on previously reported work (Ide *et al.*, 2001) that uses translation equivalents derived from a parallel corpus to determine sense distinctions that can be used to automatically sense-tag the data. Our results show that sense distinctions derived from cross-lingual information are at least as reliable as those made by human annotators. Our approach therefore provides a promising means to automatically identify sense distinctions.

2 Methodology

We conducted a study using parallel, aligned versions of George Orwell's *Nineteen Eighty-Four* (Erjavec and Ide, 1998) in seven languages: English, Romanian, Slovene, Czech, Bulgarian, Estonian, and Hungarian. The study involves languages from four language families (Germanic, Romance, Slavic, and Finno-Ugric), three languages from the same family (Czech, Slovene and Bulgarian), as well as two non-Indo-European languages (Estonian and Hungarian). Although *Nineteen Eighty-Four,* (ca. 100,000 words), is a work of fiction, Orwell's prose is not highly stylized and, as such, it provides a reasonable sample of modern, ordinary language that is not tied to a given topic or sub-domain (which is the case for newspapers, technical reports, etc.). Furthermore, the translations of the text seem to be relatively faithful to the original: over 95% of the sentence alignments in the full parallel corpus of seven languages are one-to-one (Priest-Dorman, *et al.,* 1997).

2.1 Preliminary Experiment

We constructed a multilingual lexicon based on the Orwell corpus, using a method outlined in Tufis and Barbu (2001, 2002). The complete English Orwell contains 7,069 different lemmas, while the computed lexicon comprises 1,233 entries, out of which 845 have (possibly multiple) translation equivalents in all languages. We then conducted a preliminary study using a subset of 33 nouns covering a range of frequencies and degrees of ambiguity (Ide, *et al.,* 2001).

For each noun in the sample, we extracted all sentences from the English *Nineteen Eighty-Four* containing the lemma in question, together with the parallel sentences from each of the six translations. The aligned sentences were automatically scanned to extract translation equivalents.[1] A vector was then created for each occurrence, representing all possible lexical translations in the six parallel versions: if a given word is used to translate that occurrence, the vector contains a 1 in the corresponding position, 0 otherwise. The vectors for each ambiguous word were fed to an agglomerative clustering algorithm (Stolcke, 1996), where the resulting clusters are taken to represent different senses and sub-senses of the word in question.

The clusters produced by the algorithm were compared with sense assignments made by two human annotators on the basis of WordNet 1.6.[2] In order to compare the algorithm results with the annotators' sense assignments, we normalized the data as follows: for each annotator and the algorithm, each of the 33 words was represented as a vector of length $n(n\text{-}1)/2$, where n is the number of occurrences of the word in the corpus. The positions in the vector represent a "yes-no" assignment for each pair of occurrences, indicating whether or not they were judged to have the same sense (the same WordNet sense for the annotators, and the same cluster for the algorithm). Representing the clustering algorithm results in this form required some means to "flatten" the cluster hierarchies, which typically extend to 5 or 6 levels, to conform more closely to the completely flat WordNet-based data. Therefore, clusters with a minimum distance value (as assigned by the clustering algorithm) at or below 1.7 were combined, and each leaf of the resulting collapsed tree was treated as a different sense. This yielded a set of sense distinctions for each word roughly similar in number to those assigned by the annotators.[3]

The cluster output for *glass* in Figure 1 is an example of the results obtained from the clustering algorithm. For clarity, the occurrences have been manually labeled with WordNet 1.6 senses (Figure 2). The tree shows that the algorithm correctly

[1] Sentences in which more than one translation equivalent appears were eliminated (cca. 5% of the translations).

[2] Originally, the annotators attempted to group occurrences without reference to an externally defined sense set, but this proved to be inordinately difficult and produced highly variable results and was eventually abandoned.

[3] We used the number of senses annotators assigned rather than the number of WordNet senses as a guide to determine the minimum distance cutoff, because many WordNet senses are not represented in the corpus.

grouped occurrences corresponding to WordNet sense 1 (a solid material) in one of the two main branches, and those corresponding to sense 2 (drinking vessel) in the other. The top group is further divided into two sub-clusters, the lower of which refer to a looking glass and a magnifying glass, respectively. While this is a particularly clear example of good results from the clustering algorithm, results for other words are, for the most part, similarly reasonable.

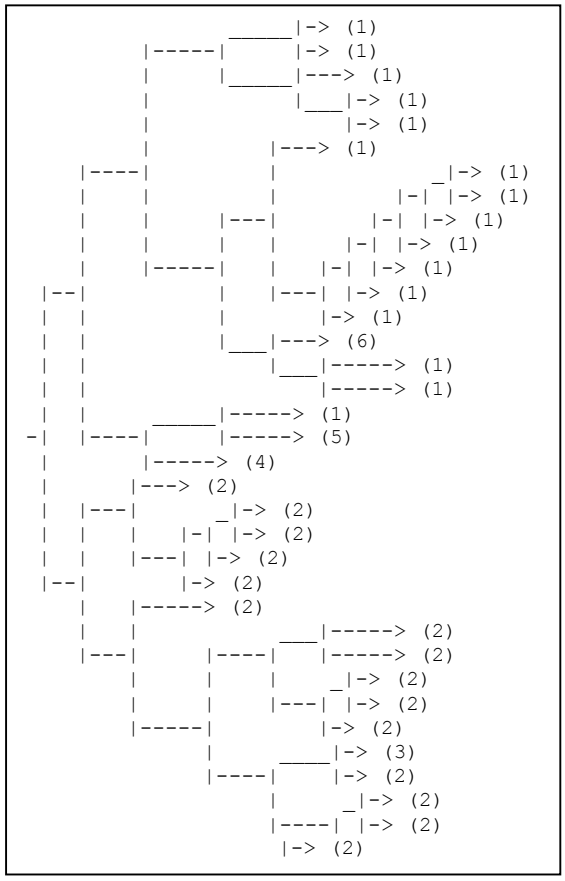

Figure 1 : Output of the clustering algorithm

1. a brittle transparent solid with irregular atomic structure
2. a glass container for holding liquids while drinking
3. the quantity a glass will hold
4. a small refracting telescope
5. a mirror; usually a ladies' dressing mirror
6. glassware collectively; "She collected old glass"

Figure 2 : WordNet 1.6 senses for *glass* (noun)

The results of the first experiment are summarized in Table 1, which shows the percentage of agreement between the cluster algorithm and each annotator, between the two annotators, and for the algorithm and both annotators taken together.[4] The percentages are similar to those reported in earlier work; for example, Ng *et al.* (1999) achieved a raw percentage score of 58% agreement among annotators tagging nouns with WordNet 1.6 senses.

Cluster/Annotator 1	66.7%
Cluster/Annotator 2	63.6%
Annotator 1/Annotator 2	76.3%
Cluster/Annotator 1/ Annotator 2	53.4%

Table 1 : Levels of agreement

2.2 Second experiment

Comparison of sense differentiation achieved using translation equivalents, as determined by the clustering algorithm, with those assigned by human annotators suggests that use of translation equivalents for word sense tagging and disambiguation is worth pursuing. Agreement levels are comparable to (and in some cases higher than) those obtained in earlier studies tagging with WordNet senses. Furthermore, the pairwise difference in agreement between the human annotators and the annotators and the clustering algorithm is only 10-13%, which is also similar to scores obtained in other studies.

In the second phase, the experiment was broadened to include 76 nouns from the multi-lingual lexicon, including words with varying ambiguity (the range in number of WordNet senses is 2 to 29, average 7.09) and semantic characteristics (e.g., abstract vs. concrete: “thought”, “stuff”, “meaning”, “feeling” vs. “hand”, “boot”, “glass”, “girl”, etc.). We chose nouns that occur a minimum of 10 times in the corpus, have no undetermined translations and at least five different translations in the six non-English languages, and have the log likelihood score of at least 18; that is:

$$LL(T_T, T_S) = 2 * \sum_{j=1}^{2} \sum_{i=1}^{2} n_{ij} * \log \frac{n_{ij} * n_{**}}{n_{i*} * n_{*j}} \geq 18$$

where n_{ij} stands for the number of times T_T and T_S have been seen together in aligned sentences, n_{i*} and n_{*j} stand for the number occurrences of T_T and T_S, respectively, and n_{**} represents the total

[4] We computed raw percentages only; common measures of annotator agreement such as the Kappa statistic (Carletta, 1996) proved to be inappropriate for our two-category (“yes-no”) classification scheme.

number of potential translation equivalents in the parallel corpus. The LL score is set at a maximum value to ensure high precision for the extracted translation equivalents, which minimizes sense clustering errors due to incorrect word alignment. Table 2 summarizes the data.

No. of words	76
No. of example sentences	2399
Average examples/word	32
No. of senses (annotator 1)	241
No. of senses (annotator 2)	280
No. of senses (annotator 3)	213
No. of senses (annotator 4)	232
No. of senses (all annotators)	345
Average senses per word	4.53
Percentage of annotator agreement:	
Full agreement (4/4)	54.27
75% agreement (3/4)	28.13
50% agreement (2/4)	16.92
No agreement	0.66

Table 2 : Summary of the data

In this second experiment, we increased the number of annotators to four. The results of the clustering algorithm and the sense assignments made by the human annotators were normalized differently than in the earlier experiment, by ignoring sense numbers and interpreting the annotators' sense assignments as clusters only. To see why this was necessary, consider the following set of sense assignments for the seven occurrences of "youth" in *Nineteen Eighty-Four*:

OCC	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Ann1	3	1	6	3	6	3	1
Ann2	2	1	4	2	6	2	1

Agreement is 43%; however, both annotators classify occurrences 1, 4, and 6 as having the same sense, although each assigned a different sense number to the group. If we ignore sense numbers and consider only the annotators' "clusters", the agreement rate is much higher,[5] and the data is more comparable to that obtained from the cluster algorithm.

We also addressed the issue of the appropriate point at which to cut off the clustering by the algorithm. Our use of a pre-defined minimum distance value to determine the number of clusters (senses) in the earlier experiment yielded varying results for different words (especially words with significantly different numbers of translation equivalents) and we sought a more principled means to determine the cut-off value. The clustering algorithm was therefore modified to compute the correct number of clusters automatically by halting the clustering process when the number of clusters reached a value similar to the average number obtained by the annotators.[6] As criteria, we used the minimum distance between existing clusters at each iteration, which determines the two clusters to be joined, where minimum distance is computed between two vectors v_1, v_2 length n as:

$$\sqrt{\sum_{i=1}^{n} (v_1(i) - v_2(i))^2}$$

Best results were obtained when the clustering was stopped at the point where:

$$(\text{dist}(k)-\text{dist}(k+1))/\text{dist}(k+1) < 0.12$$

where dist(k) is the minimal distance between two clusters at the kth iteration step.

We defined a "gold standard" annotation by taking the majority vote of the four annotators (in case of ties, the annotator closest to the majority vote in the greatest number of cases was considered to be right). Using this heuristic, the clustering algorithm assigned the same number of senses as the gold standard for 41 words. However, overall agreement was much worse (67.9%) than when the number of clusters was pre-specified. The vast majority of clustering errors occurred when sense distributions are skewed; we therefore added a post-processing phase in which the smallest clusters are eliminated and their members included in the largest cluster when the number of occurrences in the largest cluster is at least ten times that of any other cluster.[7]

With this new heuristic, the algorithm produced the same number of clusters as the gold standard for only 15 words, but overall agreement reached 74.6%. Mismatching clusters typically included

[5] In fact, the only remaining disagreement is that Annotator 1 assigns occurrences 3 and 5 together, whereas Annotator 2 assigns a different sense to occurrence 3—in effect, Annotator 2 makes a finer distinction than Annotator 1 between occurrences 3 and 5.

[6] In principle, the upper limit for the number of senses for a word is the number of senses in WordNet 1.6; however, there was no case in which all WordNet senses appeared in the text.

[7] The factor of 10 is a conservative threshold; additional experiments might yield evidence for a lower value.

only one element. There were only five words for which a difference in the number of clusters assigned by the gold standard vs. the algorithm significantly contributed to the 2.7% depreciation in agreement.

We also experimented with eliminating the data for "non-contributing" languages (i.e., languages for which there is only one translation for the target word); this was ultimately abandoned because it worsened results by amplifying the effect of synonymous translations in other languages. Finally, we compared the use of weighted vs. unweighted clustering algorithms (see, e.g., Yarowsky and Florian, 1999) and determined that results were improved using weighted clustering.

The clusters produced by each pair of classifiers (human or machine) were mapped for maximum overlap; differences were considered as divergences. The agreement between two different classifications was computed as the number of common occurrences in the corresponding clusters of the two classifications divided by the total number of the occurrences of the target word. For example, the word *movement* occurs 40 times in the corpus; both the "gold standard" and the algorithm identified four clusters, but the distribution of the 40 occurrences was substantially different, as summarized in Table 3. Thirty-four of the 40 occurrences appear in the clusters common to the two classifications; therefore, the agreement rate is 85%.

CLUSTER	1	2	3	4
Gold standard	28	6	3	3
Algorithm	25	7	6	2
Intersection	24	6	3	1

Table 3 : Gold standard vs. algorithm clustering for *movement*

2.3 Results

The results of our second experiment are summarized in Table 4, which gives the agreement rate between baseline clustering (B), in which it is assumed all occurrences are labeled with the same sense; each pair of human annotators (1-4); the gold standard (G); and the clustering algorithm (A). The table shows that agreement rates among the human annotators, as compared to those between the algorithm and all but one annotator, are not significantly different, and that the algorithm's highest level of agreement is with the baseline. This is not surprising because of the second heuristic used. However, the second best agreement rate for the algorithm is with the gold standard, which suggests that sense distinctions determined using the algorithm are almost as reliable as sense distinctions determined manually. The agreement of the algorithm with the gold standard falls slightly below that of the human annotators, but is still well within the range of acceptability. Also, given that the gold standard was computed on the basis of the human annotations, it is understandable that these annotations do better than the algorithm.

	1	2	3	4	G	A
B	71.1	65.1	76.3	74.1	75.5	81.5
1		78.1	75.6	83.1	88.6	74.4
2			71.3	75.9	82.5	66.9
3				77.3	82.1	77.1
4					90.4	75.9
G						77.3

Table 4 Agreement rates among baseline, the four annotators, gold standard, and the algorithm

3 Discussion and Further Work

Our results show that sense distinctions based on translation variants from parallel corpora are similar to those obtained from human annotators, which suggests several potential applications. Because our approach is fully automated through all its steps, it could be used to automatically obtain large samples of "sense-differentiated" data without the high cost of human annotation. Although our method does not choose sense assignments from a pre-defined list, most language processing applications (e.g. information retrieval) do not require this knowledge; they need only the information that different occurrences of a given word are used in the same or a different sense.

A by-product of applying our method is that once words in a text in one language are tagged using this method, different senses of the corresponding translations in the parallel texts are also identified, potentially providing a source of information for use in other language processing tasks and for building resources in the parallel languages (e.g., WordNets for the Eastern European languages in our study). In addition, if different senses of target

words are identified in parallel texts, contextual information for different senses of a word can be gathered for use in disambiguating other, unrelated texts. The greatest obstacle to application of this approach is, obviously, the lack of parallel corpora: existing freely available parallel corpora including several languages are typically small (e.g., the Orwell), domain dependent (e.g. the MULTEXT *Journal of the Commission* (JOC) corpus; Ide and Véronis, 1994), and/or represent highly stylized language (e.g. the Bible; Resnik *et al.*, 1999). Appropriate parallel data including Asian languages is virtually non-existent. Given that our method applies only to words for which different senses are lexicalized differently in at least one other language, its broad application depends on the future availability of large-scale parallel corpora including a variety of language types.

Many studies have pointed out that coarser-grained sense distinctions can be assigned more reliably by human annotators than finer distinctions such as those in WordNet. In our study, the granularity of the sense distinctions was largely ignored, except insofar as we attempted to cut off the number of clusters produced by the algorithm at a value similar to the number identified by the annotators. The sense distinctions derived from the clustering algorithm are hierarchical, often identifying four or five levels of refinement, whereas the WordNet sense distinctions are organized as a flat list with no indication of their degree of relatedness. Our attempt to flatten the cluster data in fact loses much information about the relatedness of senses.[8] As a result, both annotators and the clustering algorithm are penalized as much for failing to distinguish fine-grained as coarse-grained distinctions. We are currently exploring two possible sources of information about sense relatedness: the output of the clustering algorithm itself, and WordNet hypernyms, which may not only improve but also broaden the applicability of our method.

[8] Interestingly, the clustering for "glass" in Figure 1 reveals additional sub-groupings that are not distinguished in WordNet: the top sub-group of the top cluster includes occurrences that deal with some physical aspect of the material ("texture of", "surface of", "rainwatery", "soft", etc.). In the lower cluster, the two main sub-groups distinguish a (drinking) glass as a manipulatable object (by washing, holding, on a shelf, etc.) from its sense as a vessel (mainly used as the object of "pour into", "fill", "take/pick up", etc. or modified by "empty", "of gin", etc.).

We note in our data that although it is not statistically significant, there is some correlation (-.51) between the number of WordNet senses for a word and overall agreement levels. The lowest overall agreement levels were for "line" (29 senses), "step" (10), position (15), "place" (17), and "corner" (11). Perfect agreement was achieved for several words with under 5 senses, e.g., "hair" (5), "morning" (4), "sister" (4), "tree" (2), and "waist" (2)—all of which were judged by both the annotators and the algorithm to occur in only one sense in the text. On the other hand, agreement levels for some words with under five WordNet senses had low agreement: e.g., "rubbish" (2), "rhyme" (2), "destruction" (3), and "belief" (3). Because both the algorithm (which based distinctions on translations) and the human annotators (who used WordNet senses) had low agreement in these cases, the WordNet sense distinctions may be overly fine-grained and, possibly, irrelevant to many language processing tasks.

We continue to explore the viability of our method to automatically determine sense distinctions comparable to those achieved by human annotators. We are currently exploring methods to refine the clustering results as well as their comparison to results obtained from human annotators (e.g., the Gini Index [Boley, *et al.,* 1999]).

4 Conclusion

The results reported here represent a first step in determining the degree to which automated clustering based on translation equivalents can be used to differentiate word senses. Our work so far indicates that the method is promising and could provide a significant means to automatically acquire sense-differentiated data in multiple languages. Our current results suggest that coarse-grained agreement is the best that can be expected from humans, and that our method is capable of duplicating sense differentiation at this level.

5 Acknowledgements

Our thanks go to Arianna Schlegel, Christine Perpetua, and Lindsay Schulz who annotated the data, and to Ion Radu who modified the clustering algorithm. We would also like to thank the

anonymous reviewers for their comments and suggestions. All errors, of course, remain our own.

6 References

Boley D., Gini, M, Gross, R., Han, S.,. Hastings, K and Karypis, G., Kumar, V., Mobasher, B, Moore, J. (1999) Partitioning-Based Clustering for Web Document Categorization. *Decision Support Systems*, 27:3, 329-341.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22:2, 249-254.

Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics,* 20:4, 563-596.

Dagan, I., Itai, A., and Schwall, U. (1991). Two languages are more informative than one. *Proceedings of the 29th Annual Meeting of the ACL,* 18-21 Berkeley, California, 130-137.

Dyvik, H. (1998). Translations as Semantic Mirrors. *Proceedings of Workshop Multilinguality in the Lexicon II, ECAI 98*, Brighton, UK, 24-44.

Erjavec, T. and Ide, N. (1998). The MULTEXT-EAST Corpus. *Proceedings of the First International Conference on Language Resources and Evaluation,* Granada, 971-74.

Gale, W. A., Church, K. W. and Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities,* 26, 415-439.

Ide, N. (1999). Cross-lingual sense determination: Can it work? *Computers and the Humanities,* 34:1-2, 223-34.

Ide, N., Erjavec, T., and Tufis, D. (2001). Automatic sense tagging using parallel corpora. *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium,* Tokyo, 83-89.

Ide, N., Véronis, J. (1994). Multext (Multilingual Tools and Corpora). *Proceedings of the 14th International Conference on Computational Linguistics,* COLING'94, Kyoto, 90-96.

Miller, G. A., Beckwith, R. T. Fellbaum, C. D., Gross, D. and Miller, K. J. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography,* 3:4, 235-244.

Ng, H. T., Lim, C. Y., Foo, S. K. (1999). A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. *Proceedings of the ACL SIGLEX Workshop: Standardizing Lexical Resources*, College Park, MD, USA, 9-13.

Priest-Dorman, G.; Erjavec, T.; Ide, N. and Petkevic, V. (1997). Corpus Markup. COP Project 106 MULTEXT-East D2.3 F.

Resnik, P. and Yarowsky, D. (2000). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Journal of Natural Language Engineering*, 5(2): 113-133.

Resnik, P., Broman Olsen, M., Diab, M. (1999). Creating a Parallel Corpus from the Book of 2000 Tongues. *Computers and the Humanities,* 33:1-2. 129-153.

Resnik, Philip and Yarowsky, David (1997). A perspective on word sense disambiguation methods and their evaluation. *ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How?* Washington, D.C., 79-86.

Stolcke, Andreas (1996) Cluster 2.9. http://www.icsi.berkeley.edu/ftp/global/pub/ai/stolcke/software/cluster-2.9.tar.Z.

Tufis, D., Barbu, A.-M. (2001) Automatic Construction of Translation Lexicons. In V.Kluew, C. D'Attellis N. Mastorakis (eds.) *Advances in Automation, Multimedia and Modern Computer Science*, WSES Press, 156-172

Tufis, D., Barbu, A.-M. (2002), Revealing translators knowledge: statistical methods in constructing practical multilingual lexicons for language and speech processing. *International Journal of Speech Technology* (to appear).

Yarowsky, D., Florian. R. (1999). Taking the load off the conference chairs: towards a digital paper-routing assistant. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora,* 220-230.