A study of polysemy judgements and inter-annotator agreement

Jean Véronis

Laboratoire Parole et Langage
Université de Provence & CNRS
13621 Aix-en-Provence Cedex 1, France
Jean.Veronis@lpl.univ-aix.fr

Abstract. This paper describes two experiments on polysemy judgement and sense annotation. The first experiment enabled us to select the most polysemous words which were used in the second experiment, and which serve as test words for the evaluation of WSD systems. We show that this selection method yields results different from selecting words on the basis of their number of senses in a dictionary, and is more appropriate in the context. Both experiments show considerable disagreement among the human judges. Disagreement on sense annotation is particularly concerning, since it sheds some doubt on the very possibility of evaluating WSD systems. However, we show that a lot of the disagreement is due to the too fine granularity of sense definitions in dictionaries. We also show that clustering techniques can enable us to re-code the individual annotations according to a small set of "super-tags", and obtain a quite satisfactory level of inter-annotator agreement. Surprisingly enough, the natural clustering provided by lexicographers in the hierarchy of dictionary entries does not provide a substantial disagreement reduction, whereas "blind" data-oriented techniques produce satisfactory results. Beyond its practical goals, this paper therefore probably raises some more general questions about lexicographic practice and the adequacy of dictionaries to NLP tasks.

1. Introduction

Word sense disambiguation (WSD) has been recognised as a central (and difficult) problem in the very first paper on computer treatment of language, Weaver's memorandum (Weaver, 1949). Since then, there has been continuous research on WSD, in the context of various sub-fields (machine translation, information retrieval, content analysis, natural language understanding, etc. — for a recent survey, see Ide and Véronis, 1998). An impressive array of methods has been proposed, and occasionally rediscovered over the years, and various claims of efficiency have been made. However, it is extremely difficult to compare the results, and therefore the methods: the texts, test words and sense lists used are widely different across studies, as well as the evaluation protocols and metrics.

Under the auspices of ACL-SIGLEX and EURALEX, the SENSEVAL evaluation exercise[1] is attempting for the first time to run an ARPA-like competition between WSD systems. Discussions among the SENSEVAL program committee members pointed out the differences in existing linguistic resources (corpora, dictionaries, etc.) between English and other languages, and decided to organise within SENSEVAL a specific competition for Romance languages, called ROMANSEVAL[2]. A six-month test campaign was planned in co-ordination with the ARCADE project[3] on multilingual text alignment, whose word track uses the same corpus and test words.

[1] See `http://www.itri.bton.ac.uk/events/senseval`
[2] See `http://www.lpl.univ-aix.fr/projects/romanseval`
[3] See `http://www.lpl.univ-aix.fr/projects/arcade`

Running an evaluation exercise is a labour-consuming activity, especially in terms of the preparation of test material, and it is highly desirable that this material be chosen according to a rigorous methodology in order to make the evaluation results valuable and useful. In particular, the selection of the test words should not rely on the experimentator's intuition. It is also necessary that human agreement is carefully measured, in order to set an upper bound to the efficiency measures: it would be futile to expect computers to agree more with the reference corpus that human annotators among them.

Experimental data is lacking. Only a handful of researchers have studied systematically the problems of polysemy judgements and inter-annotator agreement (for example Amsler and White, 1979; Ahlswede, 1993, 1995; Ahlswede and Lorand, 1993; Jorgensen, 1990; Bruce and Wiebe; 1998), but these studies either are informal and/or involved only a few words or annotators.

This paper describes the results of a systematic study of polysemy judgements and inter-annotator agreement in the context of the ROMANSEVAL French material preparation. Two tasks are reported: (1) a polysemy judgement task, which involved six judges, 600 words belonging to three different parts-of-speech (POS), and 36000 contexts; (2) a sense tagging task, which involved six other judges, 60 words in three different POS, and 3700 contexts.

2. Material

ROMANSEVAL uses a multilingual parallel corpus, which will enable comparison of results across languages, and a study of relationships between sense tagging and translation (in relation with ARCADE). The corpus is composed of written questions asked by members of the European Parliament on a wide variety of topics (health, education, environment, economy, etc.) and corresponding answers from the European Commission in nine parallel versions (ca. 1.1 million words per language). It was collected and prepared within the MLCC-MULTEXT projects.

The number of test words and test contexts were determined according to feasibility constraints. The ROMANSEVAL program committee decided upon 60 words (i.e. 20 nouns, 20 adjectives and 20 verbs), and an average of ca. 60 different contexts for each. The senses list is drawn from a common dictionary available at low price on CD-ROM (*Petit Larousse* for French), so that no participant gets an advantage by using their own dictionary/ontology and that the dictionary is available to all. Participants will need to map their dictionary/ontology to the test dictionary. In addition, this dictionary is familiar to most educated people, and therefore no particular training will be required from annotators.

3. Polysemy judgements

The choice of test words is particularly difficult. Words should not be chosen according to intuition: intuition proves wrong in many cases when semantics is concerned. Chances are great that experimentators will pick special cases or to the contrary trivial ones, and the selection is likely not to correctly reflect the real difficulty of the WSD task. Unbiased selection criteria are not easy to find. For example, frequency alone is not a good criterion, since it was repeatedly noted since the fifties that words tend to be mostly monosemic in a given text or domain. Random selection according to frequency criteria would therefore result in a very large proportion of non-interesting words for a test based on probing a small number of words. Another possibility would be to chose the test words according to their number of senses in a given dictionary, but in this case chances are great that most of these senses do not appear in the test corpus.

We therefore proposed a selection process based on judgements by human informants of the polysemy of words in the test corpus. A subset of 600 words (200 nouns, 200 adjectives and 200 verbs) was first selected on frequency criteria, and then submitted to a panel of informants who were asked to judge whether the words were polysemic in the corpus.

It was important that the entire process was as cheap as possible in terms of manual labour. The corpus was word-segmented, and three subsets of word forms were automatically extracted corresponding respectively to nouns, adjectives and verbs that are not POS ambiguous in a large dictionary (the MULTEXT French dictionary, comprising 350,000 word forms), in order to eliminate the need for POS tagging of the corpus (and the corresponding costly hand-validation).

We decided to avoid the problem of context selection, which creates biases and problems of its own, by choosing word forms with comparable frequencies in the corpus, around the desired number of 60, so that, for each test word, all its contexts will be used. In each of the three POS subsets, a 200-word frequency slice was therefore chosen such that the mean frequency of the slice is close to 60. When different morphological forms of the same word appeared in the frequency slice, they were pooled together.

Concordance lines were printed for each of the 600 words (i.e. a total of around 36000 contexts), and manually checked to eliminate a few undesirable cases (auxiliary verbs, POS ambiguities not recorded in the lexicon), which were replaced to keep the total at 600. Concordance sets for each word were fitted on separate page, and were given to six informants. The question asked to them was "According to you, does the word X have one sense or several senses in the following contexts?", and they were invited to tick the corresponding box or a "don't know" box. Informants were linguistic students, but had never received any lexicographic training. They received a small payment for the task, and could accomplish the task in free time, but had to give the results back within a week.

Somewhat to our surprise, none of the informants found the task difficult. This is confirmed by the rate of "don't know" responses, which is particularly low (4.05%). Most words were judged has having only one sense (73.0%), but there are substantial differences among categories: nouns are judged more polysemous than verbs, in turn judged more polysemous than adjectives (Figure 1). This difference is statistically significant at $p<10^{-4}$ ($\chi^2 = 67.87$; $\nu = 4$).

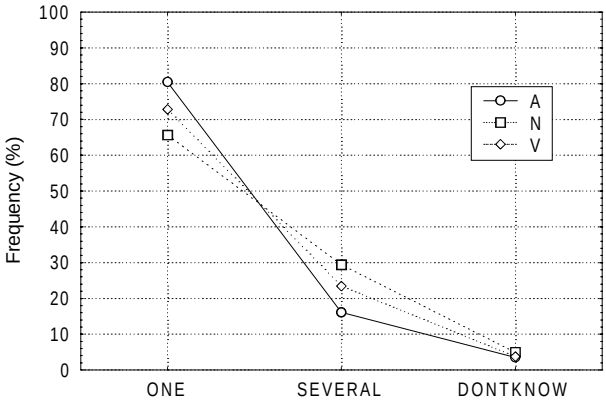

Figure 1. Polysemy judgements per POS category

However, despite the low rate of "don't know" responses, agreement between informants is also low. It seems that individual informants have no difficulty in making spontaneous judgements, but different informants tend to make different judgements. Altogether, all six judges agreed only in 45 % of the cases (Table 1). Full agreement on polysemy was achieved on only 4.5% of the words. Conversely, 40.8% of words were judged as having only one sense by all judges — the rest receiving mixed judgements. This measure is striking, but is biased with the number of judges: it tends to decrease

asymptotically towards zero as the number of judges increases. We therefore used also pairwise agreement, which is a stable measure. Table 1 shows that pairwise agreement reaches 73%.

POS	*Full*	*Pairwise*	κ
A	0.54	0.78	0.67
N	0.37	0.68	0.36
V	0.45	0.74	0.37
All	0.45	0.73	0.49

Table 1. Measures of agreement on polysemy judgements

Agreement measures should however always be corrected for chance agreement: it is obvious that some agreement would be reached even if annotators were responding at random. The kappa statistics, proposed by Cohen (1960) (see also Carletta, 1996) is a standard measure of "true" agreement, i.e. of the proportion of agreement above what would be expected by chance:

$$\kappa = \frac{p_{\text{observed}} - p_{\text{expected}}}{1 - p_{\text{expected}}}$$

where p_{expected} is computed on the basis of the marginal frequencies. This coefficient ranges between 0 when agreement is no better than chance and 1 when there is perfect agreement (it can also become negative in case of systematic disagreement).

The κ values are low, since they range from 0.37 to 0.67 depending on the POS category, with a value of 0.49 for all categories combined. Such values are generally considered in the literature as indicative of poor agreement (see Krippendorff, 1980). The correction due to κ is particularly noticeable: adjectives and verbs have very close absolute values of pairwise agreement, although "true" agreement is almost twice as important for adjectives than verbs (this is due to the highest rate of "one sense" responses for adjectives).

These differences among POS categories remain to be explained. It is interesting to note that the dictionary does not have the same perception of polysemy differences among POS. Table 2 gives the average number of senses in the *Petit Larousse* for each POS. Adjectives have less senses than other categories, which is consistent with the polysemy judgements. However, verbs have more senses than nouns, although they were felt less polysemous by informants.

POS	*Senses*
A	2.4
N	4.6
V	5.8

Table 2. Average number of senses in dictionary (all 600 words)

A score was then attributed to each word by summing up the responses (1=several senses, 0=don't know, -1=one sense). The 20 words with the highest score were selected as test words for ROMANSEVAL. Within each POS category, there is some correlation with the number of senses in the dictionary (r = 0.68 for adjectives, 0.48 for nouns, 0.51 for verbs). Although significant (all $p < 10^{-3}$), the values are low, and there is a lot of residual noise as shown in the scatterplots of Figure 2. The divergences between scores and number of senses can be explained at least by two factors, whose relative importance remains to be determined: (1) it is possible that dictionaries do not reflect well the polysemy judgements of naïve users; (2) the corpus acts as a "filter" that eliminates many of the possible senses of a word, and this filtering can be different for different words.

Table 3 gives the average number of senses for the 60 words selected. This number is, not surprisingly, higher than that of Table 2. All of these words had at least two senses in the dictionary.

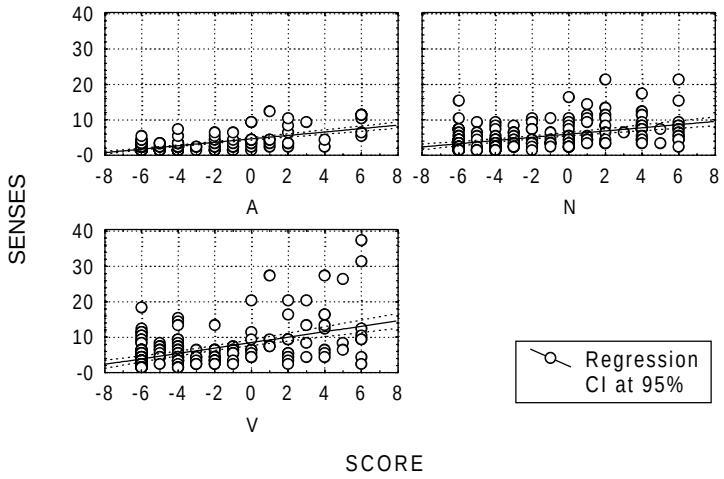

Figure 2. Score vs. number of senses in dictionary (all 600 words)

POS	*Senses*
A	6.3
N	7.6
V	12.6

Table 3. Average number of senses in dictionary (60 selected words)

The rankings of words according either to their number of senses in the dictionary or to the polysemy judgements by informants are quite different. The rank correlation coefficient (Spearman's ρ) is relatively low (0.51 for adjectives, 0.50 for nouns, 0.47 for verbs) although significant (all $p < 10^{-4}$). Selecting 20 test words in each POS category on the basis of their number of senses in the dictionary would have resulted in a significantly different list: the overlap is only of 24 words out of 60 (40%). Moreover, 18 (30%) of the words which would have been chosen would have been majoritarily judged as monosemous by the informants (score < 0), and in six cases (10%), all informants where in full agreement on the words having one sense only.

4. Inter-annotator agreement

The 60 selected words were sense-tagged by six annotators in parallel according to the *Petit Larousse* sense list. The annotators were linguistic students with no training in lexicography, different from the set of informants used in the previous task, in order to avoid cross-task biases. They were paid for the task.

For each of the 3724 occurrences of the selected words in the corpus, the corresponding paragraph of context was displayed in a spreadsheet, with the word to tag highlighted. All occurrences of the same word were grouped on the same spreadsheet, and annotators were asked to mark the senses in additional columns. They had therefore all occurrences of the same word available on the screen, and could mark them in any order, and revise their judgement as they were going along.

Annotators were instructed to chose either one sense, or several if they felt that more than one were appropriate in the given context. They could also chose no sense at all, if they felt that no sense in the dictionary was appropriate in the context. In the latter case, they were instructed to write down a question mark in the sense column. In the subsequent study, the question mark was treated as an additional sense for each word, grouping all meanings that were not found in the dictionary.

The annotators gave more senses per context for the verbs than for adjectives and nouns (Table 4, column *Nsen*). This is likely a result of the larger number of senses offered for verbs for the dictionary (see discussion above and Table 3). The average number of senses (used by a single judge in a given context) per POS category is not very high, which shows that annotators have a tendency to avoid multiple answers (as said above, the "no sense" answer is counted as a special sense). However, the average per POS category masks important differences between words: the average number of responses per word ranges from 1 to to 1.311 (verb *comprendre*). In some cases, annotators used up to six senses in a single response for a given context.

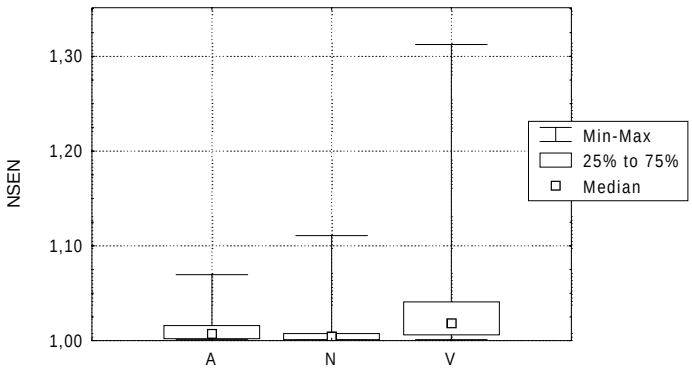

Figure 3. Average number of responses per word

Agreement was computed according to several measures (summarised in Table 4):

(1) Full agreement among the six annotators. Two variants were computed:

Min	Counts agreement when judges agree on all senses proposed for a given context
Max	Counts agreement when judges agree on at least one of the senses proposed for a given context

The difference between the *min* and *max* measures is not very important, apart from a few words (*sûr, comprendre, importer*). This is due to the fact that the average number of senses given by judges is close to 1 (Table 4, column *Nsen*). Of course, these measures are biased with the number of judges, as mentioned above. However, it is still striking to note that for some words (*correct, historique, économie, comprendre*) there was full agreement on none of the contexts for that word.

(2) Paiwise agreement. Three variants were computed:

Min	Counts agreement when judges agree on all senses proposed for a given context
Max	Counts agreement when judges agree on at least one of the senses proposed for a given context
Weighted	Accounts for partial agreement using the Dice coefficient: $$Dice = 2\frac{\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$$

Again, there is not much difference between the measures, apart from a few words, interestingly enough not exactly the same as before (*chef, comprendre, connaître*).

(3) Agreement corrected for chance. The measures above are not completely satisfactory, because they do not enable comparison of observed agreement and agreement that would be obtained by pure chance. The κ statistics mentioned above enables such a comparison. In order to account for partial agreement, κ was computed on the weighted pairwise measure using the extension proposed in Cohen (1968).

It is interesting to note that κ ranges between 0.92 (noun *détention*) and 0.007 (adjective *correct*). In other terms, there is no more agreement than chance for some words. Figure 4 shows the range of κ per POS category. The overall values of κ are low, below 50%, which indicates a great amount of disagreement among judges.

		Full		*Pairwise*			
POS	*Nsen*	*min*	*max*	*min*	*max*	*wgh*	κ
A	1.013	0.43	0.46	0.69	0.72	0.71	0.41
N	1.009	0.44	0.45	0.72	0.74	0.73	0.46
V	1.045	0.29	0.34	0.60	0.65	0.63	0.41

Table 4. Agreement measures per POS catagory

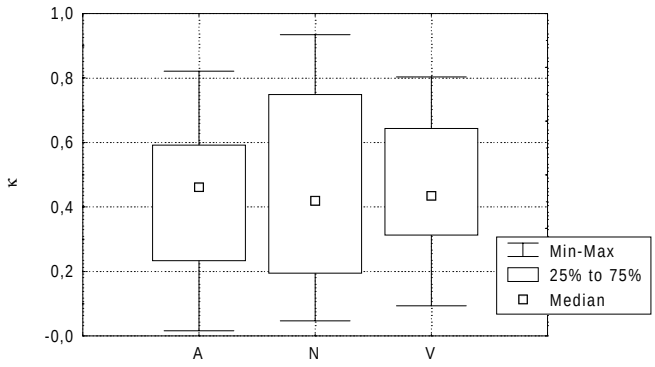

Figure 4. Distribution of κ per POS category

5. Towards robust tagging

The substantial amount of disagreement on sense tagging raises some questions about the feasibility of WSD evaluation. If human annotators do not agree much more than chance on many words, it seems that systems that produce random sense tags for these words should be considered as satisfactory. In turn another question is raised, of a more general nature: if humans do not need to perform a precise WSD to understand what they read, why should NLP systems do? We will obviously not attempt at answering this question here.

However, it is possible that the low level of agreement results from a number of controllable factors. In particular:

(1) It is possible that common dictionaries, despite the fact that they result from centuries of lexicographic tradition and work on sense delineation, are not an appropriate resource for WSD. It was

already pointed out (e.g. Kilgarriff, 1993) that dictionaries have a different goal from NLP systems, and that what appears as shortcomings for processing text, are not necessarily shortcomings for the dictionary users' goals. It is not clear, however, what kind of resources are useful for NLP tasks. The only large-scale resources available at the moment (such as *WordNet*) are very similar to common dictionaries in terms of sense divisions. It should be noted that lexicographic practice has been changing recently, following the pioneering effort of the COBUILD team to base sense distinctions and entry organisation on the strict analysis of corpora. Hopefully, better resources will therefore emerge.

(2) It is also possible that the sense divisions contained in dictionaries are too fine-grained for NLP purposes. This argument has also been made many times, and many WSD systems have been restricted to homograph level or broad sense distinctions.

In order to test this second hypothesis, we have computed the degree of inter-annotator agreement when their responses are reduced to the top-level distinctions made in the dictionary (French dictionaries are traditionally much more hierarchical than English ones, due to different lexicographic traditions). The improvement was measured as the reduction of disagreement once corrected for chance, i.e.:

$$\Delta = 1 - \frac{1-\kappa_2}{1-\kappa_1}$$

The results are disappointing: the disagreement reduction is only of 8% for adjectives and 9% for verbs. It is higher for nouns, but reaches only 25% (Table 5).

		Full		*Pairwise*				
POS	*Nsen*	*min*	*max*	*min*	*max*	*wgh*	κ	Δ
A	1.010	0.55	0.57	0.78	0.80	0.79	0.46	7.9%
N	1.003	0.70	0.70	0.86	0.86	0.86	0.60	25.2%
V	1.018	0.54	0.56	0.77	0.80	0.79	0.46	8.9%

Table 5. Agreement on top-level divisions

These results show that, somewhat surprisingly, most disagreement between annotators spans across the top-level divisions of entries. They also shed some doubt on the lexicographic practice for the hierarchical organisation of entries. The difference in improvement between nouns on one hand, and adjectives and verbs on the other hand, can probably be explained by the fact that many nouns correspond to concrete objects for which sense divisions can be made in a more "natural" way than for adjectives and verbs, or at least in a way that better matches human perception.

However, the fact that the sense clustering proposed in dictionaries does not result in much improvement in terms of inter-annotator agreement does not imply that all possible clusterings would yield bad results. We therefore tried to cluster the sense tags using "blind" data analysis methods. This has been recently proposed by Bruce and Wiebe (1998), and successfully tested on the word *interest* in the *Wall Street Journal*. However, the method we use differs in that we do not attempt at clustering *senses*, but rather *annotations*, i.e. a triple composed of a context, a judge and a sense (attributed by the judge for that context). This accounts better for systematic disagreement, where annotators seem to "prefer" different senses. In addition, we use a more powerful method, composed of two steps (for lack of space, we cannot describe the details of the method here). First, a Multiple Correspondence Analysis (Benzécri, 1973; see a presentation in Lebart *et al.*, 1997) is performed on the table of annotations, in order to reduce its high dimensionality. The projections of annotations on the three first dimensions are then used as input to a standard tree-clustering technique (ascending hierarchical,

Euclidian distance)[4]. We retained only the top of the tree, in order to form three clusters of annotations for each word. The choice of the number three was made in line with Jorgensen's (1990) finding that human informants cannot easily cluster corpus citations in samples of size comparable to ours in more than three groups. The individual annotations are then re-coded using "super-tags" corresponding to the clusters (Figure 5).

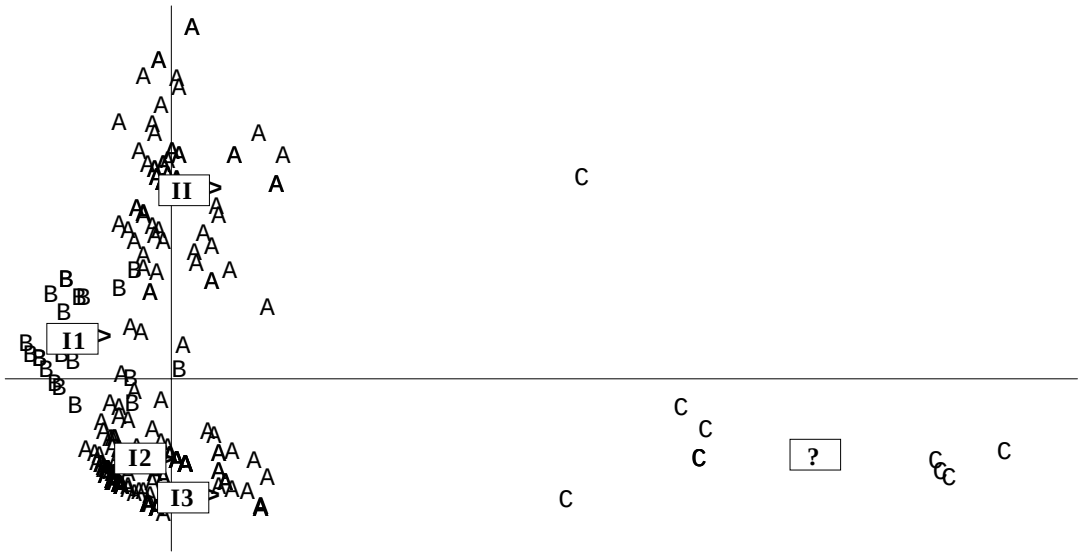

Figure 5. Clustering of annotations (word *communication*, projection on two first axes)

As opposed to sense clustering, the mapping between senses and super-tags is many-to-many, since the re-coding depends not only on the sense, but also on the judge (see Table 6).

	Super-tag			
Sense	*A*	*B*	*C*	*TOTAL*
?			9	9
I1		22		22
I2	135	1		136
I3	51		1	52
II	87		1	88
TOTAL	273	23	11	307

Table 6. Cross-tabulation of senses and super-tags (word *communication*)

So far, the technique has been tested only on nouns. However, the results are extremely encouraging, since the new κ is as high as 0.86, a value usually considered as indicative of fair agreement, with a disagreement reduction of 74%. Several parameters can be tuned up and it is likely that the results can be slightly improved.

6. Conclusion

In this paper we have reported two experiments on polysemy judgement and sense annotation. The first experiment enabled us to select the most polysemous words which were used in the second experiment and serve as test words for the evaluation of WSD systems. We have shown that this

[4] Filtering observations by dimensionality reduction before clustering has also been proposed in information retrieval under the name of *Latent Semantic Indexing* (Deerwester *et al.*, 1990), although the technique is slightly different.

selection method yields results different from selecting words on the basis of their number of senses in a dictionary, and is more appropriate in the context. Both experiments showed considerable disagreement among the human judges. Disagreement on sense annotation is particularly concerning, since it sheds some doubt on the very possibility of evaluating WSD systems. However, we have shown that a lot of the disagreement is due to the too fine granularity of sense definitions in dictionaries. We have also shown that clustering techniques can enable us to re-code the individual annotations according to a small set of "super-tags", and obtain a quite satisfactory level of inter-annotator agreement. Surprisingly enough, the natural clustering provided by lexicographers in the hierarchy of dictionary entries does not provide a substantial disagreement reduction, whereas "blind" data-oriented techniques produce satisfactory results. Beyond its practical goals, this paper therefore probably raises some more general questions about lexicographic practice and the adequacy of dictionaries to NLP tasks.

7. Acknowledgements

I am grateful to Rebecca Bruce and Janyce Wiebe for interesting discussions on inter-annotator agreement, as well as to Jean Carletta for discussions on the kappa statistics. I would also like to thank my students Corinne Jean and Valérie Houitte for their help on the experiments.

8. References

Ahlswede, T. E. (1993). Sense Disambiguation Strategies for Humans and Machines. *Proceedings of the 9th Annual Conference on the New Oxford English Dictionary*, Oxford, England, September, 75-88.

Ahlswede, T. E. (1995). Word Sense Disambiguation by Human Informants. *Proceedings of the Sixth Midwest Artificial Intelligence and Cognitive Society Conference*, Carbondale, Illinois, April 1995, 73-78.

Ahlswede, T. E., Lorand, D. (1993). The Ambiguity Questionnaire: A Study of Lexical Disambiguation by Human Informants. *Proceedings of the Fifth Midwest Artificial Intelligence and Cognitive Society Conference*, Chesterton, Indiana, 21-25.

Amsler, R. A. and White, J. S. (1979). *Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries.* Final report on NSF project MCS77-01315. University of Texas at Austin, Austin, Texas.

Benzécri, J.-P. (1973). *La Taxinomie* (vol. I); *L'analyse des Correspondances* (vol. II). Dunod, Paris.

Bruce, R., Wiebe, J. (1998). Word sense distinguishability and inter-coder agreement. *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-98).* Association for Computational Linguistics SIGDAT, Granada, Spain, June 1998.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2), 249-254.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, (70)4, 213-220.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Ide, N., Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1), in press.

Jorgensen, J. (1990). The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19, 167-190.

Kilgarriff, A. (1993). Dictionary word sense distinctions: An inquiry into their nature. *Computers and the Humanities*, 26, 365-387.

Krippendorff, K. (1980). *Content Analysis: An introduction to its Methodology.* Sage Publications.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring textual data.* Kluwer Academic Publishers, Dordrecht.

Weaver, W. (1949). *Translation.* Mimeographed, 12 pp., July 15, 1949. Reprinted in Locke, William N. and Booth, A. Donald (1955) (Eds.), *Machine translation of languages.* John Wiley & Sons, New York, 15-23.