Number or Nuance: Which Factors Restrict Reliable Word Sense Annotation?

Susan Windisch Brown, Travis Rood, and Martha Palmer
University of Colorado
296 Hellems, Boulder, CO, 80309, U.S.A.
E-mail: susan.brown@colorado.edu, travis.rood@colorado.edu, martha.palmer@colorado.edu

Abstract

This study attempts to pinpoint the factors that restrict reliable word sense annotation, focusing on the influence of the number of senses annotators use and the semantic granularity of those senses. Both of these factors may be possible causes of low interannotator agreement (ITA) when tagging with fine-grained word senses, and, consequently, low WSD system performance (Ng et al., 1999; Snyder & Palmer, 2004; Chklovski & Mihalcea, 2002). If number of senses is the culprit, modifying the task to show fewer senses at a time could improve annotator reliability. However, if overly nuanced distinctions are the problem, then more general, coarse-grained distinctions may be necessary for annotator success and may be all that is needed to supply systems with the types of distinctions that people make. We describe three experiments that explore the role of these factors in annotation performance. Our results indicate that of these two factors, only the granularity of the senses restricts interannotator agreement, with broader senses resulting in higher annotation reliability.

1. Introduction

An accurate means of performing word sense disambiguation (WSD) would improve many NLP applications, such as information extraction, information retrieval, and any task that requires more complex knowledge representation and reasoning (Sanderson, 2000; Stokoe, Oakes & Tait, 2003; Chan, Ng & Chiang, 2007). A fundamental problem for WSD is choosing the set of senses to be distinguished. One common difference between sense inventories is the level of generality of the senses. WordNet, a very common sense inventory for WSD, has been described as having fine-grained senses, whereas OntoNotes' sense inventory, which was created by clustering WordNet senses, has more coarse-grained senses. For example, WordNet lists the following senses for the verb *control:*

1. exercise authoritative control or power over
2. lessen the intensity of; temper
3. handle and cause to function
4. control (others or oneself) or influence skillfully
5. check or regulate (a scientific experiment) by conducting a parallel experiment
6. verify by using a duplicate register for comparison
7. be careful or certain to do something
8. have a firm understanding or knowledge of

OntoNotes lists the following:

1. exercise power or influence over; hold within limits
2. verify something by comparing to a standard

As you can see, the OntoNotes senses are more general, while the WordNet senses are more nuanced.

Variations in sense inventories have a great effect on WSD performance. When systems attempt to identify broad, general senses, they are much more accurate than when they attempt to distinguish between narrow, nuanced senses. Using homonym-level distinctions, supervised systems achieve over 90% accuracy, whereas with fine sense distinctions, systems achieve 60-70% accuracy (Ide & Wilks, 2006). This point is crucial, because WSD is not an end in itself, and it seems that WSD systems must achieve greater than 90% accuracy in order to contribute in any meaningful way to more complex tasks. In fact, WSD at lower accuracy levels seems to hurt the performance of such applications (e.g., information retrieval; Voorhees, 1999).

WSD systems are unlikely to achieve higher accuracy than the annotations they train on. However, standard annotation with WordNet senses usually achieves ITAs of only about 70% (Ng, Lim & Foo, 1999; Snyder & Palmer, 2004; Chklovski & Mihalcea, 2002). Resolving the problem of low annotation reliability is key to improving WSD systems.

One reason for low agreement may be that people are unable to distinguish senses with such subtle differences. If people don't make fine word sense distinctions, perhaps such distinctions are not necessary for computer applications. Coarse-grained annotation may result in higher ITAs and may be more appropriate for the text-processing tasks we want computers to perform. Another reason may be that fine-grained distinctions result in too many senses for annotators to keep straight. The 50 most frequently used nouns have an average of over 8 senses in WordNet, while for verbs, the average is 17 senses. A word with over 40 senses in WordNet is not unusual. Selecting the appropriate sense from a list of dozens of possibilities may exceed annotators' cognitive capacities. If the problem is the sheer number of senses rather than the subtle distinctions between them, then we cannot assume that we do not need our computers to make these distinctions. In this case, altering the task to reduce the cognitive load on annotators may improve reliability.

This paper will describe three experiments designed to elucidate the differences in reliability between fine-grained and coarse-grained sense annotation and the factors causing those differences. They do this by:

1. Comparing fine-grained sense annotation of text with coarse-grained annotation of the same text.
2. Focusing on the effect of the number of senses by holding the degree of sense granularity the same but varying the number of senses the annotators use.
3. Focusing on the effect of sense nuance by comparing fine-grained annotation and coarse-grained annotation when the numbers of senses are closely matched.

2. Related Research

There have been a few studies that have compared word sense annotation or WSD system performance with coarse-grained and fine-grained sense inventories. Their results give us good reason to believe that coarse-grained annotation is more reliable than fine-grained and that system performance is better when trained with coarse-grained senses. However, none have directly compared annotation for the same set of words on the same corpus. Such a comparison is necessary to eliminate the possibilities that (1) the words in one group are more difficult to annotate or (2) the instances in one of the corpora are more difficult to annotate. In addition, the sense inventories should, as much as possible, differ only in the granularity of the senses. Most importantly, none of the studies was designed in way to discover why there is a difference in annotation reliability.

Task 17 at the SemEval 2007 competition included an all-words WSD task using fine-grained WordNet sense annotation and a lexical-sample WSD task using coarse-grained OntoNotes annotation (Pradhan et al., 2007). The WordNet annotation had a 72% ITA for verbs and an 86% ITA for nouns. The OntoNotes annotation had over 90% ITA across the 100 words used in the task. Although these can be generally compared, the annotation was done on different sections of the WSJ corpus and the sets of words being annotated were not the same.

SemEval 2007, Task 7, was another all-words WSD task that used largely the same corpus as the one in Task 17, but it used coarse-grained senses rather than fine-grained (Navigli, Litkowski, & Hargraves, 2007). For Task 7, WordNet senses were automatically clustered by mapping WordNet senses to the Oxford Dictionary of English (ODE), using a method described in Navigli (2006). A portion of the data was double annotated, with an ITA rate of 93.8%. This can be compared to the Task 17 annotation with WordNet senses, with ITA rates of 72% for verbs and 86% for nouns. However, there are some important differences between the two sets of annotation. Task 17 used a larger corpus by adding 2 articles to the original set, and its fine-grained annotation only included verbs and nouns, whereas the coarse-grained annotation also included adverbs and adjectives. In addition, WordNet and the ODE are two independently created sense inventories, which divide the semantic coverage of a word in different ways. Even if one can be generally considered “fine-grained” and the other “coarse-grained”, mapping from one to the other often does not result in a coarse-grained sense from one resource cleanly subsuming the more nuanced senses from the other resource (Ide & Véronis, 1990).

Two comparisons have been done with the same words and the same corpus, although both compared manual fine-grained annotation to coarse-grained annotation derived by automatically retagging with clustered senses. The first (Ng, Lim & Foo, 1999) calculated ITA rates for double annotated instances from the Brown corpus. The study looked at instances of 191 nouns and verbs that had been annotated with fine-grained WordNet senses. The authors found an average ITA rate of 57%. They then collapsed senses using an algorithm that maximized agreement rates. It progressively combined senses in a way that eliminated some of the instance disagreements between the annotators. The algorithm continued until chance-adjusted agreement (kappa) for a particular word reached 80%. Human judgments of some of the resulting coarse-grained senses suggested that the automatically derived senses were made up of fine-grained senses that were semantically closely related.

The human assessments of the automatically clustered senses indicate that clustered fine-grained senses can be intuitively correct or appropriate. However, the study cannot answer the questions posed here because no manual annotation was ever done with the coarse-grained senses. Direct annotation of the instances with these senses may not have been as reliable as their results suggest, given that the coarse-grained results are based not on actual agreements between annotators but on senses created exactly so as to maximize agreement on existing annotations. More importantly, without a comparison between manual annotations, we cannot discover what factors affect human annotation reliability.

The second comparison was the SensEval 2 all-words competition (Edmonds & Cotton, 2001). It evaluated WSD machine learning systems based on both fine-grained and coarse-grained tags. Like Ng, Lim & Foo (1999), only the fine-grained tagging was done by human annotators. The systems trained only on fine-grained tags and labeled the test data with fine-grained labels. To evaluate the systems for coarse-grained labeling, the systems’ fine-grained tags on the test data were automatically mapped to clusters of those tags, representing more coarse-grained senses. Interestingly, for the English tasks, the coarse-grained scoring did not on average result in higher system scores for precision or accuracy, although it did for systems working in some of the other languages. Had the systems trained on coarse-grained tags, their performance may have been higher. This test would have required training data annotated with the coarse-grained tags, rather than clustering the test data and the systems' tags post hoc.

System performance is closely tied to annotation reliability, as measured by ITA. Although coarse-grained annotation seems to be more reliable than fine-grained, only a strict apple-to-apples comparison can confirm this impression or provide the controlled data needed to investigate the causes of any differences in ITA. The first experiment described here provides just such a comparison.

3. Experiment 1

This experiment compared annotator reliability for tagging with fine-grained WordNet senses to tagging with coarse-grained OntoNotes senses. To our knowledge no one has done an apples-to-apples comparison of manual annotations with the same words on the same corpus. Such a comparison is necessary to eliminate the possibilities that (1) the words annotated in one group are more difficult to annotate or (2) the instances in one of the corpora are more difficult to annotate. In addition, the sense inventories should, as much as possible, differ only in the granularity of the senses.

The OntoNotes project has annotated a large corpus (1.1 million words of English text, 1.3 million words of Chinese text, and 200,000 words of Arabic text) with multiple layers of semantic and syntactic information (Hovy et al., 2006). Word sense annotation is a key component of the project, and a coarse-grained sense inventory has been created for that purpose. The English portion of the corpus includes the TreeBanked section of the WSJ, the Broadcast News corpus, Broadcast Conversation, and WebText. In addition to word senses, the corpus is being TreeBanked and annotated with PropBank semantic roles and co-reference information.

The coarse-grained senses are developed by manually clustering related WordNet senses. Subcategorization frames and semantic classes of arguments play major roles in determining the verb groupings (Duffield et al., 2007). For each word, a sample of 50 corpus instances is annotated using a preliminary set of clustered WordNet senses. If ITA is greater than 90%, the clustered senses are used to annotate the corpus. If ITA is less than 90%, the sense groupings are revised and a new set of corpus instances are annotated. If a revised grouping fails to achieve 90% ITA, it is revised again and that third revision is used to annotate the corpus.

Each grouped sense lists the WN senses on which it is based, provides a gloss and example sentences, and maps to corresponding VerbNet classes, PropBank rolesets and FrameNet frames, if any exist. As of 2009, approximately 2,000 of the most frequent verbs in the data had been grouped and double annotated with at least 87% inter-annotator agreement.

Given the method of its creation, the OntoNotes sense inventory is an excellent resource for testing the influence of sense granularity. A coarse-grained lexicon created independently from WordNet would most likely divide some words into senses based on different attributes than WordNet did and result in WordNet senses not fitting neatly into the coarser-grained senses. In that case, a difference in ITA rates could be influenced by much more than just sense granularity. Because the senses of each verb in the OntoNotes lexicon are built from the more nuanced senses in WordNet, the chance of very different semantic criteria for sense distinctions is eliminated. With a few exceptions, the OntoNotes senses cleanly subsume clusters of WordNet senses.

3.1 Method

For this experiment the fine-grained annotation was done with WordNet senses for 40 verbs; the coarse-grained annotation was done with OntoNotes senses for the same 40 verbs. The selected verbs represent a wide range of polysemy. In WordNet, the number of senses for these verbs ranges from 3 to 36, with an average of 14.6 WordNet senses. In OntoNotes, the number of senses ranges from 2 to 15 senses, with an average of 6.2 OntoNotes senses. Although WordNet has many verbs with two senses, these were not considered for inclusion. If OntoNotes clustered the two senses into one, no ITA could be calculated for the OntoNotes annotation. Conversely, if OntoNotes preserved the two WordNet senses in its lexicon, there would be no difference in the tasks for that word between WordNet tagging and OntoNotes tagging. Nothing would be gained by comparing them.

Approximately 70 instances of each verb were chosen from the English portion of the OntoNotes corpus. First, 35 instances of the dominant OntoNotes sense were selected randomly, to be used not only in this part of the experiment, but in the other two parts as well. Second, instances covering all of the other OntoNotes senses of the verb were chosen, up to an additional 35. Some senses were quite rare in the 1.1m-word OntoNotes corpus, however, and a total of 35 additional instances could not always be acquired. In addition, instances representing every sense of a word could occasionally not be found. Within these criteria, the instances were otherwise randomly drawn. At least three sentences of context were given to annotators for each instance.

Each instance was annotated twice with OntoNotes senses as part of the GALE OntoNotes project.

The two annotators did not consult each other when choosing a sense for a particular instance of a word. Disagreements were later resolved by an adjudicator for the final version of the corpus (Duffield et al., 2007). Here, however, the unadjudicated annotations from the original annotators were used to calculate ITA rates. Multiple pairs of annotators worked on the various verbs in this study.

For this experiment, the same instances were then double annotated with WordNet senses by a new set of annotators. This annotation was also divided among several pairs of annotators.

We used the same computer interface for the WordNet annotation as was used for the OntoNotes annotation to avoid introducing any new factors that could affect the outcome. The annotation tool, STAMP[1], required some adaptation to accommodate the slightly different background information provided for the WordNet senses and the occasionally large number of WordNet senses. Every effort was made, however, to maintain the same appearance and functionality.

3.2 Results

Not surprisingly, interannotator agreement was significantly higher ($F_{(1,79)} = 166.42$, $p< .0001$) for the coarse-grained OntoNotes annotation (mean: 91%) than for the fine-grained WordNet annotation (mean: 56%) (Figure 1).

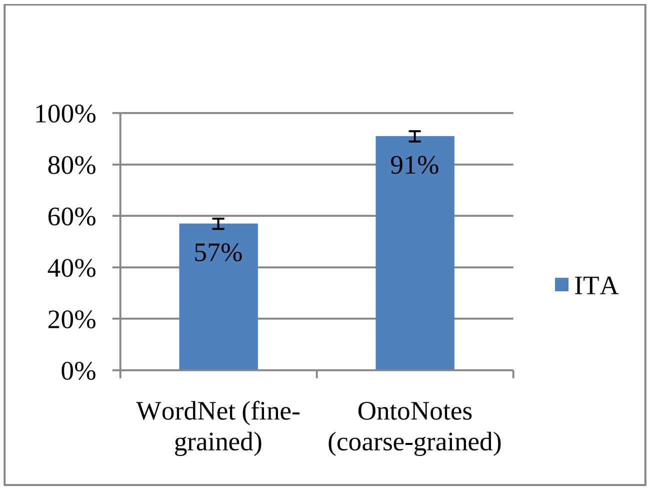

Figure 1: ITA rates for experiment 1

Deciding between 30 senses may be a much more difficult task than choosing between 7, and that difficulty may affect annotators' ability to make fully considered or consistent decisions. In addition, the number of senses varied substantially within each group. When judging the relative influence of number of senses and sense granularity on annotation, we found it more helpful to control for those factors directly.

To do this, a regression analysis was done that predicted ITA based on the number of senses annotators had to choose between and the granularity of those senses (fine or coarse). This analysis showed that when controlling for the granularity of the senses, the number of senses annotators had to choose between was not a significant factor in determining ITA ($t_{(79)} = -1.28$, $p = .206$). However, when controlling for number of senses, granularity was a significant factor ($t_{(79)} = 10.39$, $p < .0001$). The results show that when comparing annotations with the number of senses held constant, we would expect the ITAs for coarse-grained senses to be 16.2 percentage points higher than those for fine-grained senses.

We considered other factors as possible predictors of ITA or as interacting with the number of senses variable. Given the strong correlation of frequency to the familiarity and level of polysemy in a word, we tested the frequency of the word in the British National Corpus as a predictor of ITA. However, it was not a significant predictor either alone or as part of an interaction with number of senses. Because of the Zipfian nature of word frequency, we transformed the BNC frequencies of the words into a ranking, but this did not prove significant either. We repeated these tests in the other two parts of this experiment, but they never proved to be significant, so no mention of them will occur in those sections.

4. Experiment 2

4.1 Method

To focus more exclusively on number of senses as a limiting factor, we compared annotations using a full set of WordNet senses to annotations using a restricted set of WordNet senses. The degree of sense granularity remained constant because all annotators were using fine-grained WordNet senses. The two conditions differed, however, in the number of senses the annotators had to choose between.

The same verbs and instances from Experiment 1 were used. For each verb, the most frequent coarse-grained OntoNotes sense was chosen. From the original 70 instances for that verb, the 35 instances tagged (and adjudicated) with that OntoNotes sense were used as the dataset for the restricted set condition. The WordNet senses that had been clustered to form that OntoNotes sense made up the restricted set of WordNet senses (see Figure 2).

For example, for the hypothetical verb in Figure 2, OntoNotes grouped sense B could be chosen. In this case,

[11] Many thanks to Benjamin Snyder for the creation of STAMP and the use of his computer code.

the restricted set of WordNet senses would include WordNet senses 3, 7, 8, 13, and 14. One pair of annotators would tag with the full set of WordNet senses for the verb (i.e., senses 1-14), and another pair would tag with the restricted set of WordNet senses (i.e., senses 3, 7, 8, 13, 14).

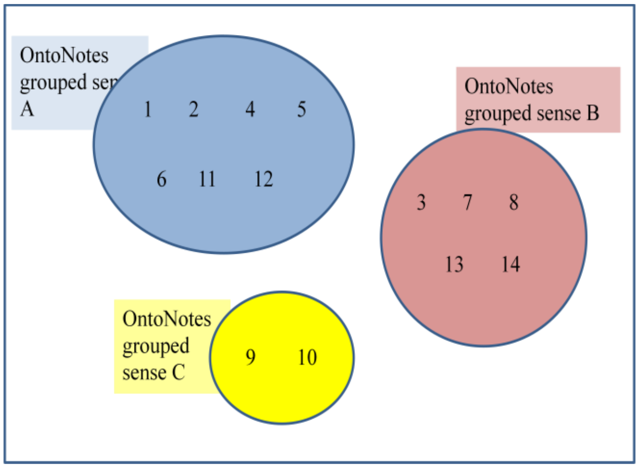
OntoNotes grouped sense A
1 2 4 5
6 11 12
OntoNotes grouped sense B
3 7 8
13 14
OntoNotes grouped sense C
9 10

Figure 2: Clustering of related WordNet senses into OntoNotes senses

For the annotators using the restricted set, the available tags were reduced but still appropriate for the instances, because each instance chosen for this part of the experiment had already been labeled with the OntoNotes sense that covered the meanings in the restricted set of WordNet senses.

The annotators using the full set of WordNet senses tagged all 70 instances for each verb. The 35 extra instances covered senses that were not part of the restricted set. The data from these instances were not used in this analysis, but acted as fillers that ensured that the annotators in this group were considering the entire set of WordNet senses. With this design, one pair of annotators tagged with a large number of fine-grained senses, while the second pair of annotators focused on only a few fine-grained sense distinctions. The number of senses varied substantially between the two annotation conditions, with the full-set condition having an average of 14.6 senses and the restricted-set condition having an average of 5.6 senses.

Using the restricted set of senses simplified the annotation task in two primary ways. First, annotators had fewer senses to consider when labeling an instance of a verb. Second, the senses they were considering were more homogenous, so they could focus on the nuanced distinctions rather than first eliminating the more clearly inappropriate senses, which were sometimes numerous.

4.2 Results

The results again show that the number of senses annotators have to deal with has no impact on the reliability of the annotations. A comparison of ITAs for the two groups (Figure 3) show no significant difference (Full set mean: 59%, Restricted set mean: 53%; $F_{(1,79)}$ = 1.75, p = .19). A simple regression using the number of senses annotators had to choose between to predict ITA was also not significant ($F_{(1,79)}$ = 0.03, p = .86).

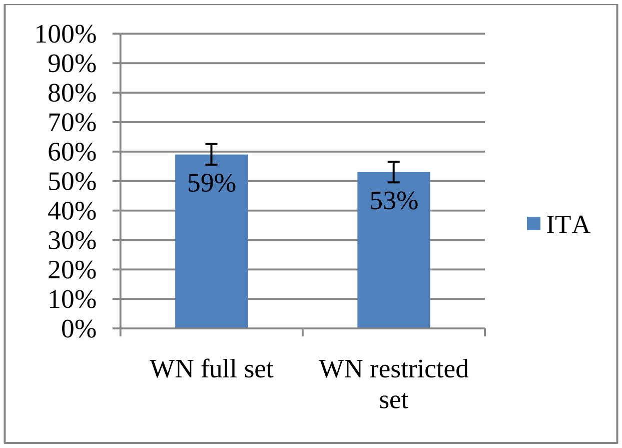

Figure 3: ITA rates for experiment 2

5. Experiment 3

5.1 Method

To focus more exclusively on granularity, a third analysis was done, comparing the ITAs of the restricted WordNet set of annotations (from comparison 2) to the ITAs of the OntoNotes annotations (from comparison 1). These two sets were more closely matched in terms of number of senses (restricted WordNet senses mean = 5.6; OntoNotes senses mean = 6.2), but differed in the granularity of the senses, with the OntoNotes senses representing broad, coarse-grained senses and the WordNet senses representing nuanced, fine-grained senses.

5.2 Results

Although the two groups had similar averages for the number of senses, within the groups there was still some variability, with WordNet senses ranging from 2 to 16 senses and OntoNotes senses ranging from 2 to 15. Therefore, our regression analysis still controlled for the number of senses annotators had to choose between for each verb.

A comparison of the ITA rates for the two groups can be seen in Figure 4. After controlling for the number of senses, sense granularity was a highly significant predictor of annotation reliability ($t_{(79)}$ = 11.2, p < .0001). When controlling for number of senses, our analysis showed that a change from fine-grained senses to coarse-grained senses would improve interannotator agreement by 19.3 percentage points.

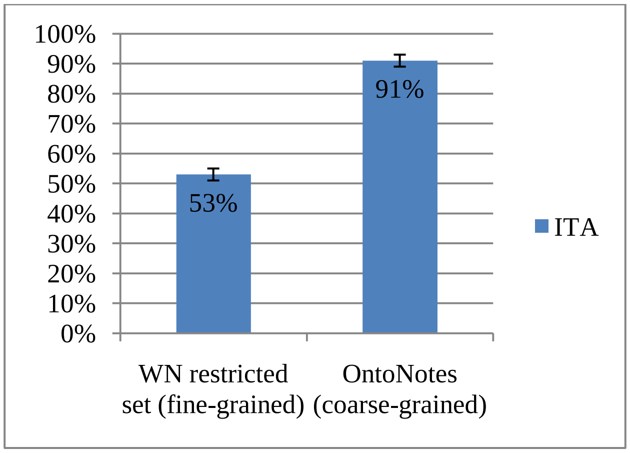

Figure 4: ITA rates for experiment 3

6. Conclusion

Our three experiments compared a variety of annotation schemes in an attempt to discover the affect of two variables on word sense annotation: the number of senses an annotator has to choose between and the nuance or level of semantic granularity of those senses. Our first analysis compared fine-grained WordNet annotation to coarse-grained OntoNotes annotation, with a large difference between the average number of senses for the two groups. The second compared annotation with a full set of WordNet senses to that with a restricted set of WordNet senses, which varied the number of senses but held the sense granularity constant. The third compared the restricted WordNet annotation with the OntoNotes annotation, where the average number of senses was matched but the sense granularity differed.

Each analysis found that number of senses was not a significant factor in ITA. Experiments 1 and 3, which had variations in sense granularity, also showed that sense granularity was a significant factor, with coarse-grained annotation resulting in higher ITAs, even when controlling for number of senses.

Artstein and Poesio (2008) state “Reliability is . . . a prerequisite for demonstrating the validity of the coding scheme—that is, to show that the coding scheme captures the “truth” of the phenomenon being studied. . . . If the annotators are not consistent then either some of them are wrong or else the annotation scheme is inappropriate for the data.” The consistently low ITAs for fine-grained word sense annotation are an indication that the annotation scheme is inappropriate.

Some recent studies suggest that the categorical nature of the task may be inappropriate (Erk & Pado, 2007). Investigations in how to design and efficiently execute a graded annotation scheme are still in progress.

Within the standard paradigm of categorical word sense annotation, however, our results suggest that altering the task to reduce the cognitive load of the annotators (by lowering the number of senses seen at one time) is not likely to result in higher interannotator agreement. The results do indicate that the level of sense nuance is an important limiting factor, with much more consistent annotation resulting from coarse-grained senses.

7. Acknowledgements

We gratefully acknowledge the efforts of all of the annotators and the support of the National Science Foundation Grants NSF-0415923, Word Sense Disambiguation and CISE-CRI-0551615, Towards a Comprehensive Linguistic Annotation and CISE-CRI 0709167, as well as a grant from the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022, a subcontract from BBN, Inc.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. References

Artstein, Ron, and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics.* 34(4), pp. 555–596.

Chan, Yee Seng, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 33–40, Prague, Czech Republic, June.

Chklovski, Tim, and Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. *Proc. of ACL 2002 Workshop on WSD: Recent Successes and Future Directions*. Philadelphia, PA.

Duffield, Cecily Jill, Jena D. Hwang, Susan Windisch Brown, Dmitriy Dligach, Sarah E.Vieweg, Jenny Davis, Martha Palmer. 2007. Criteria for the manual grouping of verb senses. *Linguistics Annotation Workshop, ACL-2007*. Prague, Czech Republic.

Edmonds, Philip, and Scott Cotton. 2001. Senseval-2: Overview. In Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, pp. 1–5.

Erk, Katrin, and Sebastian Pado. 2007. Towards a computational model of gradience in word sense. In Proceedings of IWCS-7. Tilburg, The Netherlands.

Hovy, Edward H., Mitch Marcus, Martha Palmer, Sameer Pradhan, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. Short paper. *Proceedings of HLT-NAACL 2006*. New York, NY.

Ide, Nancy, and Jean Véronis. 1990. Mapping dictionaries: A spreading activation approach. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING),* pp. 588–592. Kyoto, Japan.

Ide, Nancy, and Yorick Wilks. 2006. Making sense about sense. In Eneko Agirre and Philip Edmonds (eds.) *Word Sense Disambiguation: Algorithms and Applications.* Dordrecht, The Netherlands: Springer.

Navigli, Roberto. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 105–112, Sydney, July.

Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 Task 7: Coarse-grained English all words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007),* pp. 30–35, Prague, June.

Ng, Hwee T., Chung Y. Lim, and Shou K. Foo. 1999. A case study on the inter-annotator agreement for word sense disambiguation. In *Proc. of ACL Workshop: Standardizing Lexical Resources*. College Park, Maryland.

Pradhan, Sameer, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 Task 17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007),* pp. 87–92, Prague, June.

Sanderson, Mark. 2000. Retrieving with good sense. *Information Retrieval*. 2(1), pp. 49–69.

Snyder, Benjamin, and Martha Palmer. 2004. The English all-words task. *Proc. of ACL 2004 SENSEVAL-3 Workshop.* Barcelona, Spain.

Stokoe, Christopher, Michael P. Oakes, and John Tait. 2003. Word sense disambiguation and information retrieval revisited. In *Proceedings of the 26th annual ACM SIGIR conference on research and development in information retrieval*. Toronto, Canada..

Voorhees, Ellen. 1999. Natural language processing and information retrieval. *Information Extraction; Towards Scalable, Adaptable Systems,* ed. by Maria Theresa Pazienza, pp. 32–48. Germany: Springer