Criteria for the Manual Grouping of Verb Senses

Cecily Jill Duffield, Jena D. Hwang, Susan Windisch Brown,
Dmitriy Dligach, Sarah E.Vieweg, Jenny Davis, Martha Palmer
Departments of Linguistics and Computer Science
University of Colorado
Boulder, C0 80039-0295, USA
{cecily.duffield, hwangd, susan.brown, dmitry.dligach,
sarah.vieweg, jennifer.davis, martha.palmer}@colorado.edu

Abstract

In this paper, we argue that clustering WordNet senses into more coarse-grained groupings results in higher inter-annotator agreement and increased system performance. Clustering of verb senses involves examining syntactic and semantic features of verbs and arguments on a case-by-case basis rather than applying a strict methodology. Determining appropriate criteria for clustering is based primarily on the needs of annotators.

1 Credits

We gratefully acknowledge the support of the National Science Foundation Grant NSF-0415923, Word Sense Disambiguation, and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, a subcontract from the BBN-AGILE Team.

2 Introduction

Word sense ambiguity poses significant obstacles to accurate and efficient information extraction and automatic translation. Successful disambiguation of polysemous words in NLP applications depends on determining an appropriate level of granularity of sense distinctions, perhaps more so for distinguishing between multiple senses of verbs than for any other grammatical category. WordNet, an important and widely used lexical resource, uses fine-grained distinctions that provide subtle information about the particular usages of various lexical items (Felbaum, 1998). When used as a resource for annotation of various genres of text, this fine level of granularity has not been conducive to high rates of inter-annotator agreement (ITA) or high automatic tagging performance. Annotation of verb senses as described by coarse-grained Proposition Bank framesets may result in higher ITA scores, but the blurring of distinctions between verb senses with similar argument structures may fail to alleviate the problems posed by ambiguity. Our goal in this project is to create verb sense distinctions at a middle level of granularity that allow us to capture as much information as possible from a lexical item while still attaining high ITA scores and high system performance in automatic sense disambiguation. We have demonstrated that clear sense distinctions improve annotator productivity and accuracy. System performance typically lags around 10% behind ITA rates. ITA scores of at least 90% for a majority of our sense-groupings result in the expected corresponding improvement in system performance. Training on this new data, Chen et al., (2006) report 86.7% accuracy for verbs using a smoothed maximum entropy model and rich linguistic features. (Also Semeval07[1]) They also report state-of-the-art performance on fine-grained senses, but the results are more than 16% lower. We begin by describing the overall process.

3 The Grouping and Annotation Process

The process for building our database with the appropriate level of verb sense distinctions

[1] Task 17, http://nlp.cs.swarthmore.edu/semeval/.

Proceedings of the Linguistic Annotation Workshop, pages 49–52,
Prague, June 2007.

involves two steps: sense grouping and annotation (Figure 1). During our sense grouping process, linguists (henceforth, "groupers") cluster fine-grained sense distinctions listed in WordNet 2.1 into more coarse-grained groupings. These rough clusters of WordNet entries are based on speaker intuition. Other resources, including PropBank, VerbNet (based on Levin's verb classes (Levin, 1993)), and online dictionaries are consulted in further refining the distinctions between senses (Palmer, et. al., 2005**,** Kipper et al., 2006). To aid annotators in understanding the distinctions, sense groupings are ordered according to saliency and frequency. Detailed information, including syntactic frames and semantic features, is provided as commentary for the groupings. We also provide the annotators with simple example sentences from WordNet as well as syntactically complex and ambiguous attested usages from Google search results. These examples are intended to guide annotators faced with similar challenges in the data to be tagged.

Completed verb sense groupings are sent through sample-annotation and tagged by two annotators. Groupings that receive an ITA score of 90% or above are then used to annotate all instances of that verb in our corpora in actual-annotation. Groupings that receive less than 90% ITA scores are regrouped (Hovy et al., 2006). Revisions are made based on a second grouper's evaluation of the original grouping, as well as patterns of annotator disagreement. Verb groupings receiving ITA scores of 85% or above are sent through actual-annotation. Verbs scoring below 85% are regrouped by a third grouper, and in some cases, by the entire grouping team. It is sometimes impossible to get ITA scores over 85% for high frequency verbs that also have high entropy. These have to be carefully adjudicated to produce a gold standard. Revised verbs are then evaluated and either deemed ready for actual-annotation or are sent for a third and final round of sample-annotation. Verbs subject to the re-annotation process are tagged by different annotators. Data from actual-annotation is examined by an adjudicator who resolves remaining disagreements between annotators. The adjudicated data is then used as the gold standard for automatic annotation. The final versions of the sense groupings are mapped to VerbNet and FrameNet and linked to the Omega Ontology (Philpot et al., 2005).

Verbs are selected based on frequency of appearance in the WSJ corpus. As the most frequent verbs are also the most polysemous, the number of sense distinctions per verb as well as the number of instances to be tagged decreases as the project continues. The 740 most frequent verbs in the WSJ corpus were grouped in order of frequency. They have an average polysemy of 7 senses in WordNet; our sense groups have reduced the polysemy to 3.75 senses. Of these, 307 verb groupings have undergone regrouping to some extent. A total of 670 verbs have completed actual-annotation and adjudication. The next 660 verbs have been divided into rough semantic domains based on VerbNet classes, and grouping will proceed according to these semantic domains rather than by verb frequency. As groupers create sense groupings for new verbs, old verb sense groupings in the same semantic domain are consulted. This organization allows for more consistent grouping methodologies, as well as more efficiency in integrating our sense groupings into the Ontology.

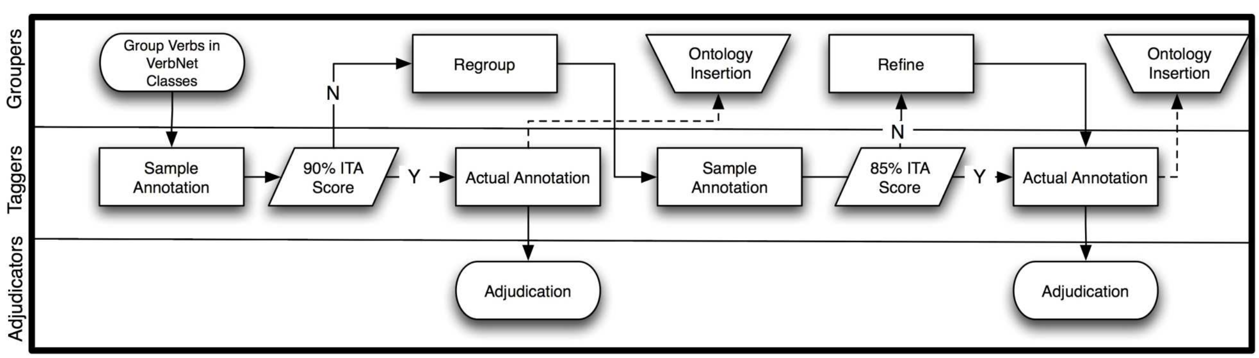

Figure 1: The grouping and annotation process.

4 Grouping Methodology

Various criteria are considered when disambiguating senses and creating sense groupings for the verbs, including frequent lexical usages and collocations, syntactic features and alternations, and semantic features, similarly to Senseval2 (Palmer, et. al. 2006). Because these criteria do not apply uniformly to every verb, groupers take various approaches when creating sense groupings. Groupers recognize that there are many alternate ways to cluster senses at this level of granularity; each grouping represents only one possible clustering as a middle ground between PropBank and WordNet senses for each verb. Our highest priority is to then create clear distinctions among sense groupings that will be easily understood by the annotators and consequently result in high ITA scores. Initial clustering is based on groupers' intuitions of the most salient categories. Many verb groupings, such as that for the verb *kill*, provide little detailed syntactic or semantic analysis and yet have received high ITA scores. The success of these intuitive sense groupings is not due to lack of polysemy; *kill* has 15 WordNet senses and 2 multi-word expressions clustered into 9 sense groupings, yet it received 94% ITA in first round sample-annotation.

While annotators have little trouble tagging text with verb senses that fall neatly into intuitive categories, many verbs have fine-grained WordNet senses that fall on a continuum between two distinct lexical usages. In such cases, syntactic and semantic aspects of the verb and its arguments help groupers cluster senses in such a way that annotators can make consistent decisions in tagging the text.

Syntactic criteria: Annotators have found syntactic frames, such as those defining VerbNet classes, to be useful in understanding boundaries between sense groupings. For example, *split* was originally grouped with consideration for the units resulting from a *splitting* event (i.e. whether a whole unit had been split into incomplete portions of the whole, or into smaller, but complete, individual units.) This grouping proved difficult for annotators to distinguish, with an ITA of 42%. Using the causative/inchoative alternation for verbs in the "break-45.1" class to regroup resulted in higher consistency among annotators, increasing the ITA score to 95%.

Semantic criteria: When senses of a verb have similar syntactic frames, and usages fall along a continuum between these senses, semantic features of the arguments, or less often, of the verb itself, can clarify these senses and help groupers draw clear distinctions between them. Argument features that are considered when creating sense groupings include [+/-attribute], [+/-patient], and [+/-locative]. It is most common for groupers to mark these features on nominal arguments, but a prepositional phrase may also be described in semantic terms. Semantic features of the verb that are considered include aspectual features, as illustrated by the use of [+/-punctual] in sense groupings for *make* (Figure 2). However, it may be argued that this feature is unnecessary for annotators to be able to distinguish between the sense groupings, as the prepositional phrase in sense 9 is a more salient feature for annotators.

Other features of the verb that were used earlier in the project include concrete/abstract, continuative, stative, and others. However, these features proved less useful than those

Sense group	Description and Commentary	WordNet 2.1 senses	Examples
8	Attain or reach something desired NP1[+agent] MAKE[+punctual] NP2[desired goal, destination, state] This sense implies the goal has been met. Includes: MAKE IT	make 13, 22, 38	- He made the basketball team. - We barely made the plane. - I made the opening act in plenty of time. - Can you believe it? We made it!
9	Move toward or away from a location NP1[+agent] MAKE[-punctual] (pronoun+way) PP/INFP	make 30, 37 make off 1 make way 1	- As the enemy approached our town, we made for the hills. - He made his way carefully across the icy parking lot. - They made off with the jewels.

Figure 2: Sense groupings 8 and 9 for "make." Senses are distinguished in part by aspectual features marked on the verb.

described above, and annotators not familiar with linguistic theory found them to be confusing. Therefore, they are now rarely used to label sense groupings. Such concepts, when used, are more likely to be described in prose commentary for the sake of the annotators.

Certain compositional features of verbs have also proven to be confusing for annotators. In several cases, attempts to distinguish sense groupings based on *manner* and *path* have resulted in increased annotator disagreement. In the first attempt at grouping *roll*, syntactic and semantic information, as well as prose commentary, was presented to help annotators distinguish the manner and path sense groupings. Despite this, the admissibility of certain prepositions in both senses ("The baby rolled over," vs "She rolled over to the wall,") may have blurred the distinction. In two rounds of sample-annotation, the greatest number of disagreements occurred with respect to these two senses for *roll*, which were then merged in the final version of the sense groupings.

5 Conclusion

Building on results in grouping fine-grained WordNet senses into more coarse-grained senses that led to improved inter-annotator agreement (ITA) and system performance (Palmer et al., 2004; Palmer et al., 2007), we have developed a process for rapid sense inventory creation and annotation of verbs that also provides critical links between the grouped word senses and the ontology (Philpot et al., 2005). This process is based on recognizing that sense distinctions can be represented by linguists in a hierarchical structure, that is rooted in very coarse-grained distinctions which become increasingly fine-grained until reaching WordNet (or similar) senses at the leaves. Sets of senses under specific nodes of the tree are grouped together into single entries, along with the syntactic and semantic criteria for their groupings, to be presented to the annotators. Criteria are applied on a case-by-case basis, considering syntactic and semantic features as consistently as possible when grouping verbs in similar semantic domains as defined by VerbNet. By using this approach when creating sense groupings, we are able to provide annotators with clear and reliable descriptions of senses, resulting in improved accuracy and performance.

References

Chen, J., A. Schein, L. Ungar and M. Palmer. 2006. An Empirical Study of the Behavior of Word Sense Disambiguation. *Proceedings of HLT-NAACL 2006.* New York, NY.

Fellbaum, C. (ed.) 1998. *WordNet: An On-line Lexical Database and Some of its Applications*. MIT Press, Cambridge, MA.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer. 2006. Extensive Classifications of English Verbs. *Proceedings of the 12th EURALEX International Congress.* Turin, Italy.

Levin, B. 1993. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago, IL.

OntoNotes, 2006. Hovy, E.H., M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. Short paper. *Proceedings of HLT-NAACL 2006.* New York, NY.

Palmer, M., O. Babko-Malaya, and H.T. Dang. 2004. Different Sense Granularities for Different Applications. *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems (HLT-NAACL 2004)*. Boston, MA.

Palmer, M., Dang, H.T., and Fellbaum, C., Making Fine-grained and Coarse-grained sense distinctions, both manually and automatically, *Journal of Natural Language Engineering* (to appear, 2007).

Palmer, M., Gildea, D., Kingsbury, P., The Proposition Bank: A Corpus Annotated with Semantic Roles, *Computational Linguistics Journal*, 31:1, 2005.

Philpot, A., E.H. Hovy, and P. Pantel. 2005. The Omega Ontology. *Proceedings of the ONTOLEX Workshop at the International Conference on Natural Language Processing (IJCNLP)*. Jeju Island, Korea.