Learning to Merge Word Senses

Rion Snow **Sushant Prakash**
Computer Science Department
Stanford University
Stanford, CA 94305 USA
{rion,sprakash}@cs.stanford.edu

Daniel Jurafsky
Linguistics Department
Stanford University
Stanford, CA 94305 USA
jurafsky@stanford.edu

Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305 USA
ang@cs.stanford.edu

Abstract

It has been widely observed that different NLP applications require different sense granularities in order to best exploit word sense distinctions, and that for many applications WordNet senses are too fine-grained. In contrast to previously proposed automatic methods for sense clustering, we formulate sense merging as a supervised learning problem, exploiting human-labeled sense clusterings as training data. We train a discriminative classifier over a wide variety of features derived from WordNet structure, corpus-based evidence, and evidence from other lexical resources. Our learned similarity measure outperforms previously proposed automatic methods for sense clustering on the task of predicting human sense merging judgments, yielding an absolute F-score improvement of 4.1% on nouns, 13.6% on verbs, and 4.0% on adjectives. Finally, we propose a model for clustering sense taxonomies using the outputs of our classifier, and we make available several automatically sense-clustered WordNets of various sense granularities.

1 Introduction

Defining a discrete inventory of senses for a word is extremely difficult (Kilgarriff, 1997; Hanks, 2000; Palmer et al., 2005). Perhaps the greatest obstacle is the dynamic nature of sense definition: the correct granularity for word senses depends on the application. For language learners, a fine-grained set of word senses may help in learning subtle distinctions, while coarsely-defined senses are probably more useful in NLP tasks like information retrieval (Gonzalo et al., 1998), query expansion (Moldovan and Mihalcea, 2000), and WSD (Resnik and Yarowsky, 1999; Palmer et al., 2005).

Lexical resources such as WordNet (Fellbaum, 1998) use extremely fine-grained notions of word sense, which carefully capture even minor distinctions between different possible word senses (e.g., the 8 noun senses of *bass* shown in Figure 1). Producing sense-clustered inventories of arbitrary sense granularity is thus crucial for tasks which depend on lexical resources like WordNet, and is also important for the task of automatically constructing new WordNet-like taxonomies. A solution to this problem must also deal with the constraints of the WordNet taxonomy itself; for example when clustering two senses, we need to consider the transitive effects of merging synsets.

The state of the art in sense clustering is insufficient to meet these needs. Current sense clustering algorithms are generally unsupervised, each relying on a different set of useful features or hand-built rules. But hand-written rules have little flexibility to produce clusterings of different granularities, and previously proposed methods offer little in the direction of intelligently combining and weighting the many proposed features.

In response to these challenges, we propose a new algorithm for clustering large-scale sense hierarchies like WordNet. Our algorithm is based on a supervised classifier that learns to make graduated judgments corresponding to the estimated probability that each particular sense pair should be merged. This classifier is trained on gold standard sense clustering judgments using a diverse feature space. We are able to use the outputs of our classifier to produce a ranked list of sense merge judgments by merge probability, and from this create sense-clustered inventories of arbitrary sense granularity.[1]

In Section 2 we discuss past work in sense cluster-

[1]We have made sense-clustered Wordnets using the algorithms discussed in this paper available for download at **http://ai.stanford.edu/∼rion/swn**.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1005–1014, Prague, June 2007. ©2007 Association for Computational Linguistics

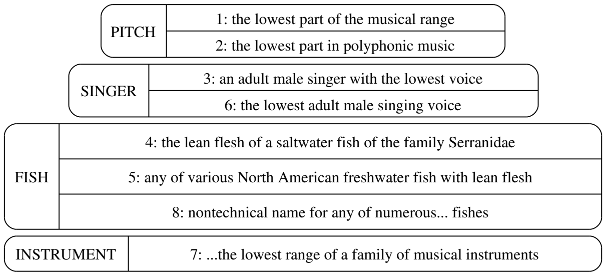

Figure 1: Sense clusters for the noun *bass*; the eight WordNet senses as clustered into four groups in the SENSEVAL-2 coarse-grained evaluation data

ing, and the gold standard datasets that we use in our work. In Section 3 we introduce our battery of features; in Section 4 we show how to extend our sense-merging model to cluster full taxonomies like WordNet. In Section 5 we evaluate our classifier against thirteen previously proposed methods.

2 Background

A wide number of manual and automatic techniques have been proposed for clustering sense inventories and mapping between sense inventories of different granularities. Much work has gone into methods for measuring synset similarity; early work in this direction includes (Dolan, 1994), which attempted to discover sense similarities between dictionary senses. A variety of synset similarity measures based on properties of WordNet itself have been proposed; nine such measures are discussed in (Pedersen et al., 2004), including gloss-based heuristics (Lesk, 1986; Banerjee and Pedersen, 2003), information-content based measures (Resnik, 1995; Lin, 1998; Jiang and Conrath, 1997), and others. Other approaches have used specific cues from WordNet structure to inform the construction of semantic rules; for example, (Peters et al., 1998) suggest clustering two senses based on a wide variety of structural cues from WordNet, including if they are *twins* (if two synsets share more than one word in their synonym list) or if they represent an example of *autohyponymy* (if one sense is the direct descendant of the other). (Mihalcea and Moldovan, 2001) implements six semantic rules, using *twin* and *autohyponym* features, in addition to other WordNet-structure-based rules such as whether two synsets share a *pertainym*, *antonym*, or are clustered together in the same *verb group*.

A large body of work has attempted to capture corpus-based estimates of word similarity (Pereira et al., 1993; Lin, 1998); however, the lack of large sense-tagged corpora prevent most such techniques from being used effectively to compare different senses of the same word. Some corpus-based attempts that are capable of estimating similarity between word senses include the *topic signatures* method; here, (Agirre and Lopez, 2003) collect contexts for a polysemous word based either on sense-tagged corpora or by using a weighted agglomeration of contexts of a polysemous word's monosemous relatives (i.e., single-sense synsets related by hypernym, hyponym, or other relations) from some large untagged corpus. Other corpus-based techniques developed specifically for sense clustering include (McCarthy, 2006), which uses a combination of word-to-word distributional similarity combined with the JCN WordNet-based similarity measure, and work by (Chugur et al., 2002) in finding co-occurrences of senses within documents in sense-tagged corpora. Other attempts have exploited disagreements between WSD systems (Agirre and Lopez, 2003) or between human labelers (Chklovski and Mihalcea, 2003) to create synset similarity measures; while promising, these techniques are severely limited by the performance of the WSD systems or the amount of available labeled data.

Some approaches for clustering have made use of regular patterns of polysemy among words. (Peters et al., 1998) uses the COUSIN relation defined in WordNet 1.5 to cluster hyponyms of categorically related noun synsets, e.g., "container/quantity" (e.g., for clustering senses of "cup" or "barrel") or "organization/construction" (e.g., for the building and institution senses of "hospital" or "school"); other approaches based on systematic polysemy include the hand-constructed CORELEX database (Buitelaar, 1998), and automatic attempts to extract patterns of systematic polysemy based on minimal description length principles (Tomuro, 2001).

Another family of approaches has been to use either manually-annotated or automatically-constructed mappings to coarser-grained sense inventories; an attempt at providing coarse-grained sense distinctions for the SENSEVAL-1 exercise included a mapping between WordNet and the Hector lexicon (Palmer et al., 2005). Other attempts in this vein include mappings between WordNet and PropBank (Palmer et al., 2004) and mappings to Levin classes (Levin, 1993; Palmer et al., 2005). (Navigli, 2006) presents an automatic approach for mapping between sense inventories; here similarities in gloss definition and structured relations between the two sense inventories are exploited in order to map between WordNet senses and distinctions made within the coarser-grained Oxford English Dictionary. Other work has attempted to exploit translational equivalences of WordNet senses in other languages, for example using foreign language WordNet interlingual indexes (Gonzalo et al., 1998; Chugur et al., 2002).

2.1 Gold standard sense clustering data

Our approach for learning how to merge senses relies upon the availability of labeled judgments of sense relatedness. In this work we focus on two datasets of hand-labeled sense groupings for WordNet: first, a dataset of sense groupings over nouns, verbs, and adjectives provided as part of the SENSEVAL-2 English lexical sample WSD task (Kilgarriff, 2001), and second, a corpus-driven mapping of nouns and verbs in WordNet 2.1 to the Omega Ontology (Philpot et al., 2005), produced as part of the ONTONOTES project (Hovy et al., 2006).

A wide variety of semantic and syntactic criteria were used to produce the SENSEVAL-2 groupings (Palmer et al., 2004; Palmer et al., 2005); this data covers all senses of 411 nouns, 519 verbs, and 257 adjectives, and has been used as gold standard sense clustering data in previous work (Agirre and Lopez, 2003; McCarthy, 2006)[2]. The number of judgments within this data (after mapping to WordNet 2.1) is displayed in Table 1.

Due to a lack of interannotator agreement data for this dataset, (McCarthy, 2006) performed an annotation study using three labelers on a 20-noun subset of the SENSEVAL-2 groupings; the three labelers were given the task of deciding whether the 351 potentially-related sense pairs were “Related”, “Unrelated”, or “Don’t Know”.[3] In this task the pairwise interannotator F-scores were (0.4874, 0.5454, 0.7926), for an average F-score of 0.6084.

SENSEVAL-2

POS	Total Pairs	Merged Pairs	Proportion
Nouns	16403	2593	0.1581
Verbs	30688	3373	0.1099
Adjectives	8368	2209	0.2640

ONTONOTES

POS	Total Pairs	Merged Pairs	Proportion
Nouns	3552	347	0.0977
Verbs	4663	1225	0.2627

Table 1: Gold standard datasets for sense merging; only sense pairs that share a word in common are included; proportion refers to the fraction of synsets sharing a word that have been merged

POS	Overlap	ON-True		ON-False		F-Score
		S-T	S-F	S-T	S-F	
Nouns	2116	121	55	181	1759	0.5063
Verbs	3297	351	503	179	2264	0.5072

Table 2: Agreement data for gold standard datasets

The ONTONOTES dataset[4] covers a smaller set of nouns and verbs, but it has been created with a more rigorous corpus-based iterative annotation process. For each of the nouns and verbs in question, a 50-sentence sample of instances is annotated using a preliminary set of sense distinctions; if the word sense interannotator agreement for the sample is less than 90%, then the sense distinctions are revised and the sample is re-annotated, and so forth, until an interannotator agreement of at least 90% is reached.

We construct a combined gold standard set from these SENSEVAL-2 and ONTONOTES groupings, removing disagreements. The overlap and agreement/disagreement data between the two groupings is given in Table 2; here, for example, the column with **ON-True** and **S-F** indicates the count of senses that ONTONOTES judged as positive examples of sense merging, but that SENSEVAL-2 data did not merge. We also calculate the F-score achieved by considering only one of the datasets as a gold standard, and computing precision and recall for the other. Since the two datasets were created independently, with different annotation guidelines, we can-

[2]In order to facilitate future work in this area, we have made cleaned versions of these groupings available at **http://ai.stanford.edu/∼rion/swn** along with a “diff” with the original files.

[3]McCarthy’s gold standard data is available at **ftp://ftp.informatics.susx.ac.uk/pub/users/dianam/relateGS/**.

[4]The OntoNotes groupings will be available through the LDC at **http://www.ldc.upenn.edu**.

not consider this as a valid estimate of interannotator agreement; nonetheless the F-score for the two datasets on the overlapping set of sense judgments (50.6% for nouns and 50.7% for verbs) is roughly in the same range as those observed in (McCarthy, 2006).

3 Learning to merge word senses

3.1 WordNet-based features

Here we describe the feature space we construct for classifying whether or not a pair of synsets should be merged; first, we employ a wide variety of linguistic features based on information derived from WordNet. We use eight similarity measures implemented within the WordNet::Similarity package[5], described in (Pedersen et al., 2004); these include three measures derived from the paths between the synsets in WordNet: HSO (Hirst and St-Onge, 1998), LCH (Leacock and Chodorow, 1998), and WUP (Wu and Palmer, 1994); three measures based on information content: RES (Resnik, 1995), LIN (Lin, 1998), and JCN (Jiang and Conrath, 1997); the gloss-based Extended Lesk Measure LESK, (Banerjee and Pedersen, 2003), and finally the gloss vector similarity measure VECTOR (Patwardan, 2003). We implement the TWIN feature (Peters et al., 1998), which counts the number of shared synonyms between the two synsets. Additionally we produce pairwise features indicating whether two senses share an ANTONYM, PERTAINYM, or derivationally-related forms (DERIV). We also create the verb-specific features of whether two verb synsets are linked in a VERBGROUP (indicating semantic similarity) or share a VERBFRAME, indicating syntactic similarity. Also, we encode a generalized notion of siblinghood in the MN features, recording the distance of the synset pair's nearest least common subsumer (i.e., closest shared hypernym) from the two synsets, and, separately, the maximum of those distances (in the MAXMN feature.

Previous attempts at categorizing systematic polysemy patterns within WordNet has resulted in the COUSIN feature[6]; we create binary features which indicate whether a synset pair belong to hypernym ancestries indicated by one or more of these COUSIN features, and the specific cousin pair(s) involved. Finally we create sense-specific features, including SENSECOUNT, the total number of senses associated with the shared word between the two synsets with the highest number of senses, and SENSENUM, the specific pairing of senses for the shared word with the highest number of senses (which might allow us to learn whether the most frequent sense of a word has a higher chance of having similar derivative senses with lower frequency).

3.2 Features derived from corpora and other lexical resources

In addition to WordNet-based features, we use a number of features derived from corpora and other lexical resources. We use the publicly available topic signature data[7] described in (Agirre and Lopez, 2004), yielding representative contexts for all nominal synsets from WordNet 1.6. These topic signatures were obtained by weighting the contexts of monosemous relatives of each noun synset (i.e., single-sense synsets related by hypernym, hyponym, or other relations); the text for these contexts were extracted from snippets using the Google search engine. We then create a sense similarity feature by taking a thresholded cosine similarity between pairs of topic signatures for these noun synsets.

Additionally, we use the WordNet domain dataset described in (Magnini and Cavaglia, 2000; Bentivogli et al., 2004). This dataset contains one or more labels indicating of 164 hierarchically organized "domains" or "subject fields" for each noun, verb, and adjective synset in WordNet; we derive a set of binary features from this data, with a single feature indicating whether or not two synsets share a domain, and one indicator feature per pair of domains indicating respective membership of the sense pair within those domains.

Finally, we use as a feature the mappings produced in (Navigli, 2006) of WordNet senses to Oxford English Dictionary senses. This OED dataset was used as the coarse-grained sense inventory in the Coarse-grained English all-words task of SemEval-

[5] We choose not to use the PATH measure due to its negligible difference from the LCH measure.

[6] This data is included in the WordNet 1.6 distribution as the "cousin.tops" file.

[7] The topic signature data is available for download at **http://ixa.si.ehu.es/Ixa/resources/sensecorpus**.

2007[8]; we specify a single binary feature for each pair of synsets from this data; this feature is true if the words are clustered in the OED mapping, and false otherwise.

3.3 Classifier, training, and feature selection

For each part of speech, we split the merged gold standard data into a part-of-speech-specific training set (70%) and a held-out test set (30%). For every synset pair we use the binary "merged" or "not-merged" labels to train a support vector machine classifier[9] (Joachims, 2002) for each POS-specific training set. We perform feature selection and regularization parameter optimization using 10-fold cross-validation.

4 Clustering Senses in WordNet

The previous section describes a classifier which predicts whether two synsets should be merged; we would like to use the pairwise judgments of this classifier to cluster the senses within a sense hierarchy. In this section we present the challenge implicit in applying sense merging to full taxonomies, and present our model for clustering within a taxonomy.

4.1 Challenges of clustering a sense taxonomy

The task of clustering a sense taxonomy presents certain challenges not present in the problem of clustering the senses of a word; in order to create a consistent clustering of a sense hierarchy an algorithm must consider the transitive effects of merging synsets. This problem is compounded in sense taxonomies like WordNet, where each synset may have additional structured relations, e.g., hypernym (IS-A) or holonym (is-part-of) links. In order to consistently merge two noun senses with different hypernym ancestries within WordNet, for example, an algorithm must decide whether to have the new sense inherit both hypernym ancestries, or whether to inherit only one, and if so it must decide which ancestry is more relevant for the merged sense.

Without strict checking, human labelers will likely find it difficult to label a sense inventory with

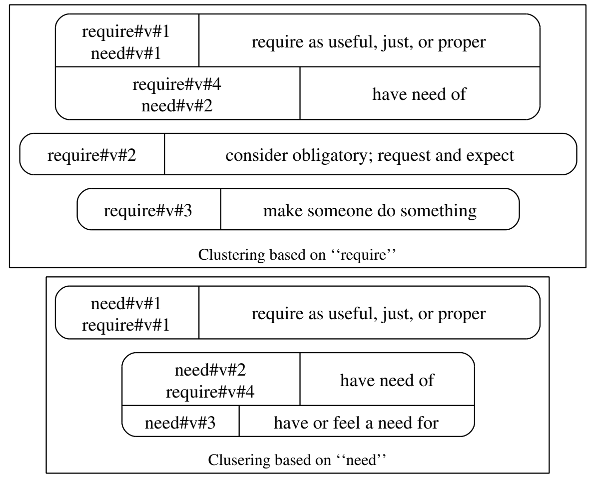

Figure 2: Inconsistent sense clusters for the verbs *require* and *need* from SENSEVAL-2 judgments

transitively-consistent judgments. As an example, consider the SENSEVAL-2 clusterings of the verbs *require* and *need*, as shown in Figure 2. In WN 2.1 *require* has four verb senses, of which the first has synonyms {*necessitate*, *ask*, *postulate*, *need*, *take*, *involve*, *call for*, *demand*}, and gloss "require as useful, just, or proper"; and the fourth has synonyms {*want*, *need*}, and gloss "have need of."

Within the word *require*, the SENSEVAL-2 dataset clusters senses 1 and 4, leaving the rest unclustered. In order to make a consistent clustering with respect to the sense inventory, however, we must enforce the transitive closure by merging the synset corresponding to the first sense (*necessitate*, *ask*, *need* etc.), with the senses of *want* and *need* in the fourth sense. In particular, these two senses correspond to WordNet 2.1 senses *need#v#1* and *need#v#2*, respectively, which are **not** clustered according to the SENSEVAL-2 word-specific labeling for *need* – *need#v#1* is listed as a singleton (i.e., unclustered) sense, though *need#v#2* is clustered with *need#v#3*, "have or feel a need for."

While one might hope that such disagreements between sense clusterings are rare, we found 178 such transitive closure disagreements in the SENSEVAL-2 data. The ONTONOTES data is much cleaner in this respect, most likely due to the stricter annotation standard (Hovy et al., 2006); we found only one transitive closure disagreement

[8] **http://lcl.di.uniroma1.it/coarse-grained-aw/index.html**

[9] We use the SVM^{perf} package, freely available for non-commercial use from **http://svmlight.joachims.org**; we use the default settings in v2.00, except for the regularization parameter (set in 10-fold cross-validation).

in the OntoNotes data, specifically WordNet 2.1 synsets (*head#n#2*, *lead#n#7*: "be in charge of") and (*head#n#3*, *lead#v#4*: "travel in front of") are clustered under *head* but not under *lead*.

4.2 Sense clustering within a taxonomy

As a solution to the previously mentioned challenges, in order to produce taxonomies of different sense granularities with consistent sense distinctions we propose to apply agglomerative clustering over all synsets in WordNet 2.1. While one might consider recalculating synset similarity features after each synset merge operation, depending on the feature set this could be prohibitively expensive; for our purposes we use average-link agglomerative clustering, in effect approximating the the pairwise similarity score between a given synset and a merged sense as the average of the similarity scores between the given synset and the clustered sense's component synsets. Further, for the purpose of sense clustering we assume a zero sense similarity score between synsets with no intersecting words.

Without exploiting additional hypernym or coordinate-term evidence, our algorithm does not distinguish between judgments about which hypernym ancestry or other structured relationships to keep or remove upon merging two synsets. In lieu of additional evidence, for our experiments we choose to retain only the hypernym ancestry of the sense with the highest frequency in SEMCOR, breaking frequency ties by choosing the first-listed sense in WordNet. We add every other relationship (meronyms, entailments, etc.) to the new merged sense (except in the rare case where adding a relation would cause a cycle in acyclic relations like hypernymy or holonymy, in which case we omit it). Using this clustering method we have produced several sense-clustered WordNets of varying sense granularity, which we evaluate in Section 5.3.

5 Evaluation

We evaluate our classifier in a comparison with thirteen previously proposed similarity measures and automatic methods for sense clustering. We conduct a feature ablation study to explore the relevance of the different features in our system. Finally, we evaluate the sense-clustered taxonomies we create on the problem of providing improved coarse-grained sense distinctions for WSD evaluation.

5.1 Evaluation of automatic sense merging

We evaluate our classifier on two held-out test sets; first, a 30% sample of the sense judgments from the merged gold standard dataset consisting of both the SENSEVAL-2 and ONTONOTES sense judgments; and, second, a test set consisting of only the ONTONOTES subset of our first held-out test set. For comparison we implement thirteen of the methods discussed in Section 2. First, we evaluate each of the eight WordNet::Similarity measures individually. Next, we implement cosine similarity of topic signatures (TOPSIG) built from monosemous relatives (Agirre and Lopez, 2003), which provides a real-valued similarity score for noun synset pairs.

Additionally, we implement the two methods proposed in (Peters et al., 1998), namely using metonymy clusters (MetClust) and generalization clusters (GenClust) based on the COUSIN relationship in WordNet. While (Peters et al., 1998) only considers four cousin pairs, we re-implement their method for general purpose sense clustering by using all 226 cousin pairs defined in WordNet 1.6, mapped to WordNet 2.1 synsets. These methods each provide a single clustering of noun synsets.

Next, we implement the set of semantic rules described in (Mihalcea and Moldovan, 2001) (MIMO); this algorithm for merging senses is based on 6 semantic rules, in effect using a subset of the TWIN, MAXMN, PERTAINYM, ANTONYM, and VERBGROUP features; in our implementation we set the parameter for when to cluster based on number of twins to $K = 2$; this results in a single clustering for each of nouns, verbs, and adjectives. Finally, we compare against the mapping from WordNet to the Oxford English Dictionary constructed in (Navigli, 2006), equivalent to clustering based solely on the OED feature.

Considering merging senses as a binary classification task, Table 3 gives the F-score performance of our classifier vs. the thirteen other classifiers and an uninformed "merge all synsets" baseline on our held-out gold standard test set. This table shows that our SVM classifier outperforms all implemented methods on the basis of F-score on both datasets

	SENSEVAL-2 + ONTONOTES			ONTONOTES	
Method	Nouns	Verbs	Adj	Nouns	Verbs
SVM	**0.4228**	**0.4319**	**0.4727**	**0.3698**	**0.4545**
RES	0.3817	0.2703	—	0.2807	0.3156
WUP	0.3763	0.2782	—	0.3036	0.3451
LCH	0.3700	0.2440	—	0.2857	0.3396
OED	0.3310	0.2878	0.3712	0.2183	0.3962
LESK	0.3174	0.2956	0.4323	0.2914	0.3774
HSO	0.3090	0.2784	0.4312	0.3025	0.3156
TOPSIG	0.3072	—	—	0.2581	—
VEC	0.2960	0.2315	0.4321	0.2454	0.3420
JCN	0.2818	0.2292	—	0.2222	0.3156
LIN	0.2759	0.2464	—	0.2056	0.3471
Baseline	0.2587	0.2072	0.4312	0.1488	0.3156
MIMO	0.0989	0.2142	0.0759	0.1833	0.2157
GenClust	0.0973	—	—	0.0264	—
MetClust	0.0876	—	—	0.0377	—

Table 3: F-score sense merging evaluation on hand-labeled testsets

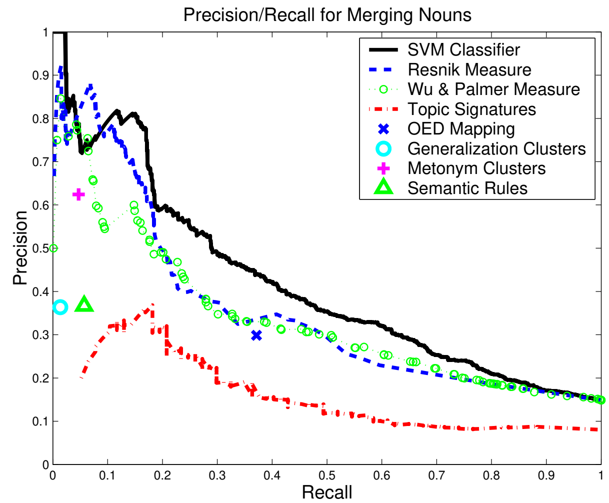

Figure 3: Precision/Recall plot for noun sense merge judgments

for all parts of speech. In Figure 3 we give a precision/recall plot for noun sense merge judgments for the SENSEVAL-2 + ONTONOTES dataset. For sake of simplicity we plot only the two best measures (RES and WUP) of the eight WordNet-based similarity measures; we see that our classifier, RES, and WUP each have higher precision all levels of recall compared to the other tested measures.

Of the methods we compare against, only the WordNet-based similarity measures, (Mihalcea and Moldovan, 2001), and (Navigli, 2006) provide a method for predicting verb similarities; our learned measure widely outperforms these methods, achieving a 13.6% F-score improvement over the LESK similarity measure. In Figure 4 we give a precision/recall plot for verb sense merge judgments, plotting the performance of the three best WordNet-based similarity measures; here we see that our classifier has significantly higher precision than all other tested measures at nearly every level of recall.

Only the measures provided by LESK, HSO, VEC, (Mihalcea and Moldovan, 2001), and (Navigli, 2006) provide a method for predicting adjective similarities; of these, only LESK and VEC outperform the uninformed baseline on adjectives, while our learned measure achieves a 4.0% improvement over the LESK measure on adjectives.

5.2 Feature analysis

Next we analyze our feature space. Table 4 gives the ablation analysis for all features used in our system as evaluated on our held-out test set; here the quantity listed in the table is the F-score loss obtained by removing that single feature from our feature space, and retraining and retesting our classifiers, keeping everything else the same. Here negative scores correspond to an *improvement* in classifier performance with the removal of the feature.

For noun classification, the three features that yield the highest gain in testset F-score are the topic signature, OED, and derivational link features, yielding a 4.0%, 3.6%, and 3.5% gain, respectively.

For verb classification, we find that three features yield more than a 5% F-score gain; by far the largest single-feature performance gain for verb classification found in our ablation study was the DERIV feature, i.e., the count of shared derivational links between the two synsets; this single feature improves our maximum F-score by 9.8% on the testset. This is a particularly interesting discovery, as none of the referenced automatic techniques for sense clustering presently make use of this very useful feature. We also achieve large gains with the LIN and LESK similarity features, with F-score improvement of 7.4% and 5.4% gain respectively.

For adjective classification again the DERIV feature proved very helpful, with a 3.5% gain on the testset. Interestingly, only the DERIV feature and the SENSECNT features helped across all parts of speech; in many cases a feature which proved to be

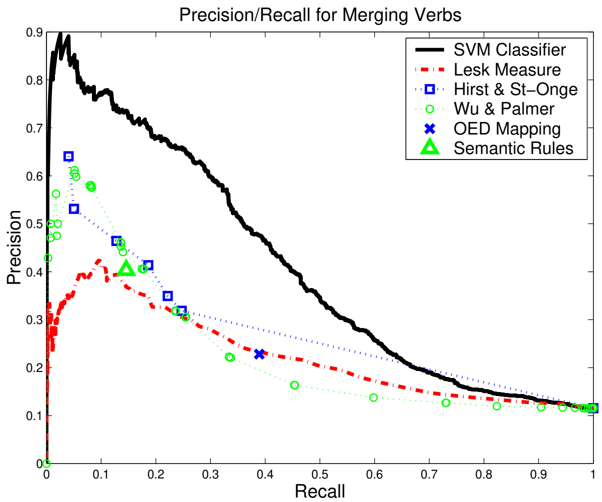

Figure 4: Precision/Recall plot for verb sense merge judgments

very helpful for one part of speech actually hurt performance on another part of speech (e.g., LIN on nouns and OED on adjectives).

5.3 Evaluation of sense-clustered Wordnets

Our goal in clustering a sense taxonomy is to produce fully sense-clustered WordNets, and to be able to produce coarse-grained Wordnets at many different levels of resolution. In order to evaluate the entire sense-clustered taxonomy, we have employed an evaluation method inspired by Word Sense Disambiguation (this is similar to an evaluation used in Navigli, 2006, however we do not remove monosemous clusters). Given past system responses in the SENSEVAL-3 English all-words task, we can evaluate past systems on the same corpus, but using the coarse-grained sense hierarchy provided by our sense-clustered taxonomy. We may then compare the scores of each system on the coarse-grained task against their scores given a random clustering at the same resolution. Our expectation is that, if our sense clustering is much better than a random sense clustering (and, of course, that the WSD algorithms perform better than random guessing), we will see a marked improvement in the performance of WSD algorithms using our coarse-grained sense hierarchy.

We consider the outputs of the top 3 all-words WSD systems that participated in Senseval-3: Gambl (Decadt et al., 2004), SenseLearner (Mihalcea and Faruque, 2004), and KOC University (Yuret,

	Nouns	Verbs	Adjectives
F-SCORE	0.4228	0.4319	0.4727
Feature	F-Score Ablation Difference		
TOPSIG	0.0403	—	—
OED	0.0355	0.0126	-0.0124
DERIV	0.0351	0.0977	0.0352
RES	0.0287	0.0147	—
TWIN	0.0285	0.0109	-0.0130
MN	0.0188	0.0358	—
LESK	0.0183	0.0541	-0.0250
SENSENUM	0.0155	0.0146	-0.0147
SENSECNT	0.0121	0.0160	0.0168
DOMAIN	0.0119	0.0082	-0.0265
LCH	0.0099	0.0068	—
WUP	0.0036	0.0168	—
JCN	0.0025	0.0190	—
ANTONYM	0.0000	0.0295	0.0000
MAXMN	-0.0013	0.0179	—
VEC	-0.0024	0.0371	-0.0062
HSO	-0.0073	0.0112	-0.0246
LIN	-0.0086	0.0742	—
COUSIN	-0.0094	—	—
VERBGRP	—	0.0327	—
VERBFRM	—	0.0102	—
PERTAINYM	—	—	-0.0029

Table 4: Feature ablation study; F-score difference obtained by removal of the single feature

2004). A guess by a system is given full credit if it was either the correct answer or if it was in the same cluster as the correct answer.

Clearly any amount of clustering will only increase WSD performance. Therefore, to account for this natural improvement and consider only the effect of our particular clustering, we also calculate the expected score for a random clustering of the same granularity, as follows: Let C represent the set of clusters over the possible N synsets containing a given word; we then calculate the expectation that an incorrectly-chosen sense and the actual correct sense would be clustered together in the random clustering as $\frac{\sum_{c \in C} |c|(|c|-1)}{N(N-1)}$.

Our sense clustering algorithm provides little improvement over random clustering when too few or too many clusters are chosen; however, with an appropriate threshold for average-link clustering we find a maximum of 3.55% F-score improvement in WSD over random clustering (averaged over the decisions of the top 3 WSD algorithms). Table 5 shows the improvement of the three top WSD algorithms given a sense clustering created by our algorithm vs. a random clustering at the same granularity.

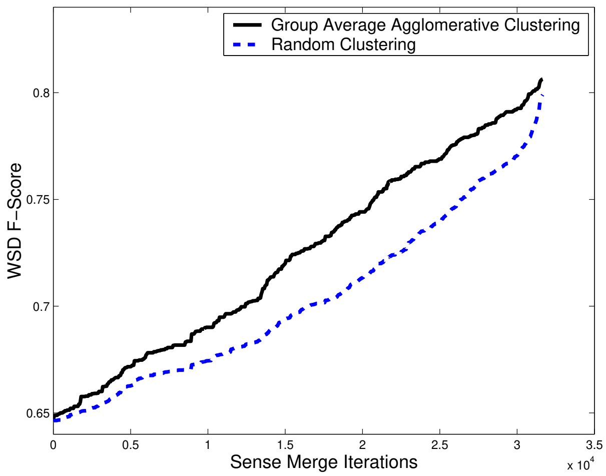

Figure 5: WSD Improvement with coarse-grained sense hierarchies

System	F-score	Avg-link	Random	Impr.
Gambl	0.6516	0.7702	0.7346	0.0356
SenseLearner	0.6458	0.7536	0.7195	0.0341
KOC Univ.	0.6414	0.7521	0.7153	0.0368

Table 5: Improvement in SENSEVAL-3 WSD performance using our average-link agglomerative clustering vs. random clustering at the same granularity

6 Conclusion

We have presented a classifier for automatic sense merging that significantly outperforms previously proposed automatic methods. In addition to its novel use of supervised learning and the integration of many previously proposed features, it is interesting that one of our new features, the DERIV count of shared derivational links between two synsets, proved an extraordinarily useful new cue for sense-merging, particularly for verbs.

We also show how to integrate this sense-merging algorithm into a model for sense clustering full sense taxonomies like WordNet, incorporating taxonomic constraints such as the transitive effects of merging synsets. Using this model, we have produced several WordNet taxonomies of various sense granularities; we hope these new lexical resources will be useful for NLP applications that require a coarser-grained sense hierarchy than that already found in WordNet.

Acknowledgments

Thanks to Marie-Catherine de Marneffe, Mona Diab, Christiane Fellbaum, Thad Hughes, and Benjamin Packer for useful discussions. Rion Snow is supported by an NSF Fellowship. This work was supported in part by the Disruptive Technology Office (DTO)'s Advanced Question Answering for Intelligence (AQUAINT) Phase III Program.

References

Eneko Agirre and Oier Lopez de Lacalle. 2003. Clustering WordNet word senses. In *Proceedings of RANLP 2003*.

Eneko Agirre and Oier Lopez de Lacalle. 2004. Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of LREC 2004*.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of IJCAI 2003*.

Lisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *Proceedings of COLING Workshop on Multilingual Linguistic Resources*, 2004.

Timothy Chklovski and Rada Mihalcea. 2003. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. In *Proceedings of RANLP 2003*.

Irina Chugur, Julio Gonzalo, and Felisa Verdejo. 2002. Polysemy and Sense Proximity in the Senseval-2 Test Suite. In *Proceedings of ACL 2002 WSD Workshop*.

Bart Decadt, Veronique Hoste, Walter Daelemans, and Antal van den Bosch. 2004. Gamble, genetic algorithm optimization of memory-based wsd. In *Proceedings of ACL/SIGLEX Senseval-3*.

William Dolan. 1994. Word Sense Ambiguation: Clustering Related Senses. In *Proceedings of ACL 1994*.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Julio Gonzalo, Felia Verdejo, Irina Chugur, and Juan Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of COLING-ACL 1998 Workshop on WordNet in NLP Systems*.

Patrick Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, **34**(1-2): 171-177.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In *WordNet: An Electronic Lexical Database*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. *Proceedings of HLT-NAACL 2006*.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, 19-33.

Thorsten Joachims. 2002. Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer, 2002.

Adam Kilgariff. 1997. I don't believe in word senses. *Computers and the Humanities*, **31**(1-2): 1-13.

Adam Kilgarriff. 2001. English lexical sample task description. In *Proceedings of the* SENSEVAL-2 *workshop*, 17-20.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC 1986*.

Beth Levin. 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago, IL.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of ICML 1998*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 1998*.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC 2000*.

Diana McCarthy. 2006. Relating WordNet Senses for Word Sense Disambiguation. In *Proceedings of ACL Workshop on Making Sense of Sense, 2006*.

Rada Mihalcea and Dan I. Moldovan. 2001. Automatic Generation of a Coarse Grained WordNet. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*.

Rada Mihalcea and Ehsanul Faruque. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of ACL/SIGLEX Senseval-3*.

Dan I. Moldovan and Rada Mihalcea. 2000. Using WordNet and lexical operators to improve Internet searches. IEEE Internet Computing, 4(1):34-43.

Roberto Navigli. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of COLING-ACL 2006*.

Martha Palmer, Olga Babko-Malaya, Hoa Trang Dang. 2004. Different Sense Granularities for Different Applications. In *Proceedings of Workshop on Scalable Natural Language Understanding*.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2005. Making fine-grained and coarse-grained sense distinctions. Journal of Natural Language Engineering.

Siddharth Patwardhan. 2003. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, Univ. of Minnesota, Duluth.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *Proceedings of NAACL 2004*.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional Clustering of English Words. In *Proceedings of ACL 1993*.

Wim Peters, Ivonne Peters, and Piek Vossen. 1998. Automatic Sense Clustering in EuroWordNet. In *Proceedings of LREC 1998.*

Andrew Philpot, Eduard Hovy, and Patrick Pantel. 2005. The Omega Ontology. In *Proceedings of the ONTOLEX Workshop at IJCNLP 2005*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the IJCAI 1995*, 448-453.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. Natural Language Engineering, 5(2):113-134.

Noriko Tomuro. 2001. Tree-cut and A Lexicon based on Systematic Polysemy. In *Proceedings of NAACL 2001*.

Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. In *Proceedings of ACL 1994*.

Deniz Yuret. 2004. Some experiments with a naive bayes wsd system. In *Proceedings of ACL/SIGLEX Senseval-3*.