Extracting Glosses to Disambiguate Word Senses

Weisi Duan
Carnegie Mellon University
Language Technologies Institute
5000 Forbes Ave.
Gates Hillman Complex 5407
Pittsburgh, PA 15213
wduan@cs.cmu.edu

Alexander Yates
Temple University
Computer and Information Sciences
1805 N. Broad St.
Wachman Hall 303A
Philadelphia, PA 19122
yates@temple.edu

Abstract

Like most natural language disambiguation tasks, word sense disambiguation (WSD) requires world knowledge for accurate predictions. Several proxies for this knowledge have been investigated, including labeled corpora, user-contributed knowledge, and machine readable dictionaries, but each of these proxies requires significant manual effort to create, and they do not cover all of the ambiguous terms in a language. We investigate the task of automatically extracting world knowledge, in the form of glosses, from an unlabeled corpus. We demonstrate how to use these glosses to automatically label a training corpus to build a statistical WSD system that uses no manually-labeled data, with experimental results approaching that of a supervised SVM-based classifier.

1 Introduction

For many semantic natural language processing tasks, systems require world knowledge to disambiguate language utterances. Word sense disambiguation (WSD) is no exception — systems for WSD require world knowledge to figure out which aspects of a word's context indicate one sense over another. A fundamental problem for WSD is that the required knowledge is open-ended. That is, for every ambiguous term, new kinds of information about the world become important, and the knowledge that a system may have acquired for previously-studied ambiguous terms may have little or no impact on the next ambiguous term. Thus open-ended knowledge acquisition is a fundamental obstacle to strong performance for this disambiguation task.

Researchers have investigated a variety of techniques that address this knowledge acquisition bottleneck in different ways. Supervised WSD techniques, for instance, can learn to associate features in the context of a word with a particular sense of that word. Knowledge-based techniques rely on machine-readable dictionaries or lexical resources like WordNet (Fellbaum, 1998) to provide the necessary knowledge. And most recently, systems have used resources like Wikipedia, which contain user-contributed knowledge in the form of sense-disambiguated links, to acquire world knowledge for WSD. Yet each of these approaches is limited by the amount of manual effort that is needed to build the necessary resources, and as a result the techniques are limited to a subset of English words for which the manually-constructed resources are available.

In this work we investigate an alternative approach that attacks the problem of knowledge acquisition head-on. We use information extraction (IE) techniques to extract *glosses*, or short textual characterizations of the meaning of one sense of a word. In the ideal case, we would extract full logical forms to define word senses, but here we instead focus on a more feasible, but still very useful, sub-task: for a given word sense, extract a collection of terms that are highly correlated with that sense and no other sense of the ambiguous word. Our system requires as input only an unlabeled corpus of documents that each contain the ambiguous term of interest.

In experiments, we demonstrate that our gloss extraction system can often determine key aspects of a word's senses. In one experiment our system was able to extract glosses with 60% precision for 20 ambiguous biomedical terms, while discovering 7 senses of those terms that never appeared in a widely-used dictionary of biomedical terminology. In addition, we demonstrate that our extracted glosses are useful for real WSD problems: our sys-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 627–635,
Los Angeles, California, June 2010.

tem outperforms a state-of-the-art unsupervised system, and it comes close to the performance of a supervised WSD system on a challenging dataset.

In the next section, we describe previous work. In Section 3, we formally define the gloss extraction task and refine it into a sub-task that is feasible for an IE approach, and Section 5 presents our technique for using extracted glosses in a WSD task. Section 6 discusses our experiments and results.

2 Previous Work

Many previous systems (Cui et al., 2007; Androutsopoulos and Galanis, 2005) have studied the related task of answering definitional questions on the Web, such as "What does cold mean?". Such systems are focused on information retrieval for human consumption, and especially on recall of definitional information (Velardi et al., 2008). They generally do not consider the problem of how to merge the large number of similar extracted definitions into a single item (Fujii and Ishikawa, 2000), so that the overall result contains one definition per sense of the word. A separate approach (Pasca, 2005) relies on the WordNet lexical database to supply the set of senses, and extracts alternate glosses for the senses that have already been defined. When glosses are to be used by computational methods, as in a WSD system in our case, it becomes critical that the system extract one coherent gloss per sense. As far as we are aware, no previous system has extracted glosses for word sense disambiguation.

Gloss extraction is related to the task of ontology extraction, in which systems extract hierarchies of word classes (Snow et al., 2006; Popescu et al., 2004). Gloss extraction differs from ontology extraction in that it extracts definitional information characterizing senses of a single word, rather than trying to place a word in a hierarchy of other words.

Most WSD systems have relied on hand-labeled training examples (Leroy and Rindflesch, 2004; Joshi et al., 2005; Mohammad and Pedersen, 2004) or on dictionary glosses (Lesk, 1986; Stevenson and Wilks, 2001) or the WordNet hierarchy (Boyd-Graber et al., 2007) to help make disambiguation choices. In recent coarse-grained evaluations, such systems have achieved accuracies of close to 90% (Pradhan et al., 2007; Agirre and Soroa, 2007; Schijvenaars et al., 2005). However, by some estimates, English contains over a million word types, and new words and new senses are added to the language every day. It is unreasonable to expect that any system will have access to hand-labeled training examples or useful dictionary glosses for each of them.

More recent techniques based on user-contributed knowledge (Mihalcea, 2007; Chklovski and Mihalcea, 2002; Milne and Witten, 2008), such as that found in Wikipedia, suffer from similar problems – Wikipedia contains many articles on well known entities, categories, and events, but very few articles that disambiguate verbs, adjectives, adverbs, and certain kinds of nouns which are poorly represented in an encyclopedia.

On the other hand, word usages in large corpora like the Web reflect nearly all of the word senses in use in English today, albeit without manually-supplied labels. Unsupervised approaches to WSD use clustering techniques to group instances of words into clusters that correspond to different senses (Pantel and Lin, 2002). While such systems are more general than supervised and dictionary-based approaches in that they can handle any word type and word sense, they have lagged behind other approaches in terms of accuracy thus far – for example, the best system in the recent word sense induction task of Semeval 2007 (Agirre and Soroa, 2007) achieved an F1 score of 78.7, slightly below the baseline (78.9) in which all instances of a word are part of a single cluster. Part of the problem is that the clustering techniques operate in a bag-of-words-like representation. This is an extremely high-dimensional space, and it is difficult in such a space to determine which dimensions are noise and which ones correlate with different senses. Our gloss extraction technique helps to address this curse of dimensionality by reducing the large vocabulary of a corpus to a much smaller set of terms that are highly relevant for WSD. Others (Kulkarni and Pedersen, 2005) have used feature selection techniques like mutual information to reduce dimensionality, but so far these techniques have only been able to find features that correlate with an ambiguous term. With gloss extraction, we are able to find features that correlate with individual senses of a term.

3 Overview: The Gloss Extraction Task

Given an input corpus $\mathcal{C}$ of documents where each document contains at least one instance of a *keyword* k, a Gloss Extraction system should produce a set of *glosses* $G = \{g_i\}$, where each g_i is a logical expression defining the meaning of a particular sense s_i of

Glosses:
1. cold(a) $\equiv$ isA(a, b) $\wedge$ disease(b) $\wedge$ symptom(a, c) $\wedge$ possibly-includes(c, d) $\wedge$ fever(d)
2. cold(a) $\equiv$ isA(a, b) $\wedge$ physical-entity(b) $\wedge$ temperature(a, c) $\wedge$ less-than(c, 25C)

Sense Indicators:
1. common cold, virus, symptom, fever
2. hot, ice cold, lukewarm, cold room, room temperature

Figure 1: **Example glosses and sense indicators for two senses of the word *cold*.**

k, to the exclusion of all other senses of k. Note that the system must discover the appropriate number of senses in addition to the gloss for each sense.

While extraction technology has made impressive advancements, it is not yet at a stage where it can produce full logical forms for sense glosses. As a first step towards this goal, we introduce the task of Sense Indicator Extraction, in which each gloss g_i consists of a set of features that, when present in the context of an instance of k, strongly indicate that the instance has sense s_i, and no other sense. Examples of both tasks are given in Figure 1. The Sense Indicator Extraction task represents a nontrivial extraction challenge, but it is much more feasible than full Gloss Extraction. And the task preserves key properties of Gloss Extraction: the results are quite useful for word sense disambiguation. The results are also readily interpreted upon inspection, making it easy to monitor a system's accuracy.

4 Extracting Word Sense Glosses

We present the GLOSSY system, an unsupervised information extraction system for Sense Indicator Extraction. GLOSSY proceeds in two phases: a *collocation detection* phase, in which the system detects components of the glosses, and an *arrangement* phase, in which the system decides how many distinct senses there are, and puts together the components of the glosses.

4.1 Collocation Detection

The first major challenge to a Gloss Extraction system is that the space of possible features is enormous, and almost all of them are irrelevant to the task at hand. Supervised techniques can use labeled examples to provide clues, but in an unsupervised setting the curse of dimensionality can be overwhelming. Indeed, unsupervised WSD techniques suffer from exactly this problem.

GLOSSY's answer to this problem is based on the following observation: pairs of potential features which rarely or never co-occur in the same document in a large corpus are likely to represent features for two distinct senses. The well-known observation that words rarely exhibit more than one sense per discourse (Yarowsky, 1995) implies that features closely associated with a particular sense have a low probability of appearing in the same document as features associated with another sense. Features that are independent of any particular sense of the keyword, on the other hand, have no such restriction, and are just as likely to appear in the context of one sense as any other. As a consequence, a low count for the co-occurrence of two potential features over a large corpus of documents for keyword k is a reliable indicator that the two features are part of the glosses of two distinct senses of k.

GLOSSY's collocation detector begins by indexing the corpus and counting the frequency of each vocabulary word. Using the index, the collocation detector determines all pairs of potential features such that each feature appears at least T times, and the pair of features never co-occurs in the same document. We call the pairs that this step finds the "non-overlapping" features. Finally, we rank the feature pairs according to the total number of documents they appear in, and choose the most frequent N pairs. This excludes non-overlapping pairs that have not been seen often enough to provide reliable evidence that they are features of different senses, and it cuts down on processing time for the next phase of the algorithm. The collocation detector outputs the set of features $F = \{f | \exists_{f'}(f, f')$ or (f', f) is one of the top N non-overlapping pairs$\}$. The GLOSSY system uses stems, words, and bigrams as potential features. We use $N = 100$ and $T = 50$ in our experiments. Figure 2 shows an example corpus and the set of features that the collocation detector would output.

Corpus of documents for term `cold`:
DOCUMENT 1: "Symptoms of the common cold may include fever, headache, sore throat, and coughing."
DOCUMENT 2: "Hibernation is a common response to the cold winter weather of temperate climates."

Non-overlapping feature pairs:
(symptoms,temperate) (headache, climate) (cold winter, common cold) (*response*, headache)

Detected collocations:
symptoms, temperate, headache, climate, cold winter, common cold, *response*

Arranged glosses:
`cold 1`: symptoms, common cold, headache
`cold 2`: temperate, climate, cold winter

Figure 2: **Example operation of the** GLOSSY **extraction system.** The collocation detector finds potential features using its non-overlapping pair heuristic. The arranger selects a subset of the potential features (in this example, it drops the feature *response*) and clusters them to produces glosses containing sense indicators.

4.2 Arranging Glosses

Given the corpus $\mathcal{C}$ for keyword k and the features F that GLOSSY's collocation detector has discovered, the arrangement phase groups these features into coherent sense glosses. Figure 2 shows an example of how the features found during collocation detection may be arranged to form coherent glosses for two senses of the word "cold."

GLOSSY's Arranger component uses a combination of a small set of statistics to determine whether a particular arrangement of the features into glosses is warranted, based on the given corpus. Let $\mathcal{A} \subset 2^F$ be an arrangement of the features into clusters representing glosses. We require that clusters in $\mathcal{A}$ be disjoint, but we do not require every feature in F to be included in a cluster in $\mathcal{A}$ — in other words, $\mathcal{A}$ is a partition of a subset of F. We define a scoring function S that is a linear interpolation of several statistics of the arrangement $\mathcal{A}$ and the corpus $\mathcal{C}$:

$$S(\mathcal{A}|\mathcal{C}, \mathbf{w}) = \sum_i w_i f_i(\mathcal{A}, \mathcal{C}) \qquad (1)$$

After experimenting with a number of options, we settled on the following for our statistics f_i:

NUMCLUSTERS: the number of clusters in $\mathcal{A}$. We use a negative weight for this statistic to favor fewer senses and encourage clustering.

DOCSCOVERED: the total number of documents in $\mathcal{C}$ in which at least one feature from $\mathcal{A}$ appears. We use this statistic to encourage the Arranger to find an arrangement that explains the sense of as many examples of the keyword as possible.

BADOVERLAPS: the number of pairs of features that co-occur in at least one document in $\mathcal{C}$, and that belong to different clusters of $\mathcal{A}$. A negative weight for this statistic encourages overlapping feature pairs to be placed in the same cluster.

BADNONOVERLAPS: the number of pairs of features that never co-occur in $\mathcal{C}$, and that belong to the same cluster in $\mathcal{A}$. A negative weight for this statistic encourages non-overlapping feature pairs to be placed in different clusters.

Given such an optimization function, the Arranger attempts to maximize its value by searching for an optimal $\mathcal{A}$. Note that this is a structured prediction task in which the choice for some sub-component of $\mathcal{A}$ can greatly affect the choice of other clusters and features. GLOSSY addresses this optimization problem with a greedy hill-climbing search with random restarts. Each round of hill-climbing is initialized with a randomly chosen subset of features, which are then all assigned to a single cluster. Using a randomly chosen search operator from a pre-defined set, the search procedure attempts to move to a new arrangement $\mathcal{A}'$. It accepts the move to $\mathcal{A}'$ if the optimization function gives a higher value than at the previous state; otherwise, it continues from the previous state. Our set of search operators include a move that splits a cluster; a move that joins two clusters; a move that swaps a feature from one cluster to another; a move that removes a feature from the arrangement altogether; and a move

that adds a feature from the pool of unused features. We used 100 rounds of hill-climbing, and found that each round converged in fewer than 1000 moves.

To estimate the weights w_i for each of the four features of the Arranger, we use a development corpus consisting of 185 documents each containing the same ambiguous term, and each labeled with sense information. Because of the small number of parameters, we performed a grid search on the development data for the optimal values of the weights.

5 A Bootstrapping WSD System

Yarowsky (1995) first recognized that it is possible to use a small number of features for different senses to bootstrap an unsupervised word sense disambiguation system. In Yarowsky's work, his system requires an initial, manually-supplied collocation as a feature for each sense of a keyword. In contrast, we can use GLOSSY's extracted glosses to supply starter features fully automatically, using only an unlabeled corpus. Thus GLOSSY complements the efforts of Yarowsky and other bootstrapping techniques for WSD (Diab, 2004; Mihalcea, 2002).

Building on their efforts, we design a bootstrapping WSD system using GLOSSY's extracted glosses as follows. Let $\mathcal{A}$ be the arranged features representing glosses for a keyword. We first retrieve all the documents from our unlabeled corpus which contain features in $\mathcal{A}$. We then label appearances of the target word according to the cluster of the features that appear in that document. If features for more than one cluster appear in the same document, we discard it. The result is an automatically labeled corpus containing examples of all the extracted senses.

We use this automatically labeled "bootstrap corpus" to perform supervised WSD. This allows our system a great deal of flexibility once the bootstrap corpus is created: we can use any features of the corpus, plus the labels, in our classifier. Importantly, this means we do not need to rely on just the features in the extracted glosses. We use a multi-class SVM classifier with a linear kernel and default parameter settings. We use LibSVM (Chang and Lin, 2001) for all of our experiments. We use standard features for supervised WSD (Liu et al., 2004): all stems, words, bigrams, and trigrams within a context window of 20 words surrounding the ambiguous term.

6 Experiments

We ran two types of experiments, one to measure the accuracy of our sense gloss extractor, and one to measure the usefulness of the extracted knowledge for word sense disambiguation.

6.1 Data

We use a dataset of biomedical literature abstracts from Duan *et al.*(2009). The data contains a set of documents for 21 ambiguous terms. We reserved one of these terms ("MCP") for setting parameters, and ran our algorithms on the remaining keywords. The ambiguous terms vary from acronyms (7 terms), which are common and important in biomedical literature, to ambiguous biomedical terminology (3 terms), to terms like "culture" and "mole" that have some biomedical senses and some senses that are part of the general lexicon (11 terms). There were on average 271 labeled documents per term; the smallest number of documents for a term is 125, and the largest is 503. For every ambiguous term, we added on average 9625 (minimum of 1544, maximum of 15711) unlabeled documents to our collection by searching for the term on PubMed Central and downloading additional PubMed abstracts.

6.2 Extracting Glosses

We measured the performance of GLOSSY's gloss extraction by comparing the extracted glosses with definitions contained in the Unified Medical Language System (UMLS) Metathesaurus. First, for each ambiguous term, we looked up the set of exact matches for that term in the Metathesaurus, and downloaded definitions for all of the different senses listed under that term. Wherever possible, we used the MeSH definition of a sense; when that was unavailable, we used the definition from the NCI Thesaurus; and when both were unavailable, we used the definition from the resource listed first. 34 senses (40%) had no available definitions at all, but in all cases, the Metathesaurus lists a short (usually 1-3 word) gloss of the sense, which we used instead.

We manually aligned extracted glosses with UMLS senses in a way that maximizes the number of matched senses for every ambiguous term. We consider an extracted gloss to match a UMLS sense when the extracted gloss unambiguously refers to a single sense of the ambiguous term, and that sense matches the definition in UMLS. Typically, this means that the extracted features in the gloss

overlap content words in the UMLS definition (*e.g.*, the extracted feature "symptoms" for the "common cold" sense of the term "cold"). In some cases, however, there was no strict overlap in content words between the extracted gloss and the UMLS definition, but the sense of the extracted gloss still unambiguously matched a unique UMLS sense: *e.g.*, for the term "transport," the extracted gloss "intracellular transport" was matched with the UMLS sense of "Molecular Transport," which the NCI Thesaurus defines as, "Any subcellular or molecular process involved in translocation of a biological entity, such as a macromolecule, from one site or compartment to another." In the end, such matchings were determined by hand. Table 1 shows extracted glosses and UMLS definitions for the term "mole."

For each ambiguous term, we measure the number of extracted glosses, the number of UMLS senses, and the number of matches between the two. We report on the precision (number of matches / number of extracted glosses), recall (number of matches / number of UMLS senses), and F1 score (harmonic mean of precision and recall). Table 2 shows the average of the precision and recall numbers over all terms. Since these terms have different numbers of senses, we can compute this average in two different ways: a Macro average, in which each term has equal weight in the average; and a Micro average, in which each term's weight in the average is proportional to the number of senses (extracted senses for the precision, and UMLS senses for the recall). We report on both.

A strict matching between GLOSSY's glosses and UMLS senses is potentially unfair to GLOSSY in several ways: GLOSSY may discover valid senses that happen not to appear in UMLS; UMLS senses may overlap one another, and so multiple UMLS senses may match a single GLOSSY gloss; and the two sets of senses may differ in granularity. For the sake of repeatable experiments, in this evaluation we make no attempt to change existing UMLS senses.

However, to highlight one particular strength of the Gloss Extraction paradigm, we do consider a separate evaluation that allows for new senses that GLOSSY discovers, but do not appear in UMLS. For instance, "biliopancreatic diversion" and "bipolar disorder" are both valid senses for the acronym "BPD." GLOSSY discovers both, but UMLS does not contain entries for either, so in our original evaluation both senses would count against GLOSSY's precision. To correct for this, our second evaluation adds senses to the list of UMLS senses whenever GLOSSY discovers valid entries missing from the Metathesaurus. The last five columns of Table 2 show our results under these conditions.

Despite the difficulty of the task, GLOSSY is able to find glosses with 53% precision and 47% recall (Macro average, no discovered senses) using only unlabeled corpora as input, and it is extracting roughly the right number of senses for each ambiguous term. In addition, GLOSSY is able to identify 7 valid senses missing from UMLS for the 20 terms in our evaluation. Including these senses in the evaluation increases GLOSSY's F1 by 6.2 points Micro (4.7 Macro). We are quite encouraged by the results, especially because they hold promise for WSD. Note that in order to improve upon a WSD baseline which tags all instances of a word as the same sense, GLOSSY only needs to be able to separate one sense from the rest. GLOSSY is finding between 1.85 and 2.2 correct glosses per term, more than enough to help with WSD.

6.3 WSD with Extracted Glosses

While extracting glosses is an important application in its own right, we also aim to show that this extracted knowledge is useful for an established application: namely, word sense disambiguation. Our next experiment compares the performance of our WSD system with an established unsupervised algorithm, and with a supervised technique — support vector machines (SVMs).

Using the same dataset as above, we trained GLOSSY on the ambiguous term "MCP", and tested it on the remaining ones. For comparison, we also report the state-of-the-art results of Duan *et al.*'s (2009) SENSATIONAL system, and the results of a BASELINE system that lumps all documents into a single cluster. SENSATIONAL is a fast clustering system based on minimum spanning trees and a pruning mechanism that eliminates noisy points from consideration during clustering. Since SENSATIONAL uses both "MCP" and "white" to train a small set of parameters, we leave "white" out of our comparison as well. We measure accuracy by aligning each system's clusters with the gold standard clusters in such a way as to maximize the number of elements that belong to aligned clusters. We use an implementation of the MaxFlow algorithm to determine this alignment. We then compute accuracy

GLOSSY		UMLS	
1.	choriocarcinoma, invasive, complete, hydatidiform mole, hydatidiform	1.	Hydatidiform Mole – Trophoblastic hyperplasia associated with normal gestation, or molar pregnancy. ... Hydatidiform moles or molar pregnancy may be categorized as complete or partial based on their gross morphology, histopathology, and karyotype.
2.	grams per mole	2.	Mole, unit of measurement – A unit of amount of substance, one of the seven base units of the International System of Units. It is the amount of substance that contains as many elementary units as there are atoms in 0.012 kg of carbon-12.
3.	mole fractions	-	
-		3.	Nevus – A circumscribed stable malformation of the skin ....
-		4.	Talpidae – Any of numerous burrowing mammals found in temperate regions ...

Table 1: GLOSSY**'s extracted glosses and UMLS dictionary entries for the example term "mole".**

		Without Discovered Senses					**With Discovered Senses**				
	GLOSSY Senses	UMLS Senses	Matches	P	R	F1	UMLS Senses	Matches	P	R	F1
Macro Avg	4.35	4.25	1.85	53.1	47.1	49.9	4.6	2.2	60.6	49.7	54.6
Micro Avg	N/A	N/A	N/A	42.5	43.5	43.0	N/A	N/A	50.6	47.8	49.2

Table 2: GLOSSY **can automatically discover glosses that match definitions in an online dictionary.** "Without Discovered Senses" counts only the senses that are listed in the UMLS Metathesaurus; "With Discovered Senses" enhances the Metathesaurus with 7 new senses that GLOSSY has automatically discovered.

as the percentage of elements that belong to aligned clusters. This metric is very similar to the so-called "supervised" evaluation of Agirre *et al.* (2006).

The first four columns of Table 3 show our results. Clearly, both SENSATIONAL and GLOSSY outperform the BASELINE significantly, and traditionally this is a difficult baseline for unsupervised WSD systems to beat. SENSATIONAL outperforms GLOSSY by approximately 6%. There appear to be two reasons for this. In other experiments, SENSATIONAL has been shown to be competitive with supervised systems, but only when the corpus consists mostly of two, fairly well-balanced senses, as is true for this particular dataset, where the two most common senses always covered at least 70% of the examples for every ambiguous term.

A more serious problem for GLOSSY is that the unlabeled corpus that it extracts glosses from may not match well with the labeled test data. If the relative frequency of senses in the unlabeled documents does not match the relative frequency of senses in the labeled test set, GLOSSY may not extract the right set of glosses. Manual inspection of the extracted glosses shows that this is indeed a problem: for example, the labeled data contains two senses of the word "mole": a discolored area of skin (78%), and a burrowing mammal (22%); our unlabeled data contains both of these senses, but the additional sense of "mole" as a unit of measurement is by far predominant. GLOSSY manages to extract glosses for "skin" and "unit of measurement," but misses out on "mammal" as a result of the skew in the data.

Note that this problem, though serious for our experiments, is largely artificial from the point of view of applications. In a typical usage of a WSD system, there is a supply of data that the system needs to disambiguate, and accuracy is measured on a labeled sample of this data. Here, we started from a sample of labeled data, constructed a larger corpus that does not necessarily match it, and then ran our algorithm.

To correct for the artificial bias in our experiment, we ran a second test in which we manually labeled a random sample of 100 documents for each ambiguous term from the larger unlabeled corpus. We used a subset of 14 of the 21 keywords in the original dataset. As before, we compared our system against SENSATIONAL and the most-frequent-sense BASELINE. We also compare against an SVM system using 3-fold cross-validation. We use a linear kernel SVM, with the same set of features that are available

	Duan *et al.*(2009) Data				Sampled Data				
Keyword	Num. senses	BASE-LINE	GLOSSY	SENSE-ATIONAL	Num. senses	BASE-LINE	SENSE-ATIONAL	GLOSSY	SVM
ANA	2	63.1	87.9	100	13	75	79	74	75.8
BPD	3	39.8	71.6	52.9	7	33	48	85	66.7
BSA	2	50.1	77.9	94.7	5	97	53	89	87.9
CML	2	55.0	99.2	89.5	4	81	75	84	75.8
MAS	2	50.0	100	100	35	46	90	67	66.7
VCR	2	79.2	79.2	64.0	8	72	32	72	75.8
cold	3	37.1	73.3	66.8	3	87	81	44	90.9
culture	2	52.0	67.1	81.7	3	74	39	62	66.7
discharge	2	66.3	82.4	95.1	5	57	41	84	54.5
fat	2	50.6	50.1	53.2	2	97	60	97	97.0
mole	2	78.3	71.3	95.8	7	78	47	57	84.8
pressure	2	52.1	69.8	86.4	5	47	60	65	75.8
single	2	50.0	59.7	99.5	4	53	63	37	45.4
white	-	-	-	-	7	32	33	58	51.5
fluid	2	64.3	83.5	99.6	-	-	-	-	-
glucose	2	50.5	64.5	50.5	-	-	-	-	-
inflammation	3	35.5	52.8	50.4	-	-	-	-	-
inhibition	2	50.4	50.4	54.2	-	-	-	-	-
nutrition	3	38.8	53.8	54.9	-	-	-	-	-
transport	2	50.6	41.1	56.8	-	-	-	-	-
AVERAGE	2.16	53.4	70.3	76.1	7.71	66.3	57.2	69.6	72.5
Diff from BL	-	0.0	+16.9	+22.7	-	0.0	-9.1	+3.3	+6.2

Table 3: GLOSSY**'s extracted glosses can be used to create an unsupervised WSD system that achieves an accuracy within 3% of a supervised system.** Our WSD system outperforms our BASELINE system, widely recognized as a difficult baseline for unsupervised WSD, by 16.9% and 3.3% on two different datasets.

to the SVM in the GLOSSY system. We run our unsupervised systems on all of the unlabeled data, and then intersect the resulting clusters with the document set that we randomly sampled.

The last four columns of Table 3 show our results. The sampled data set appears to be a significantly harder test, since even the supervised SVM achieves only a 6% gain over the BASELINE. The SENSATIONAL system does significantly worse on this data, where there is a wider variation in the distribution of senses. The GLOSSY system outperforms both the SENSATIONAL system and the BASELINE.

7 Conclusion and Future Work

Gloss Extraction is an important, and difficult task of extracting definitions of words from unlabeled text. The GLOSSY system succeeds at a more feasible refinement of this task, the Sense Indicator Extraction task. GLOSSY's extractions have proven useful as seed definitions in an unsupervised WSD task. There is a great deal of room for future work in expanding the ability of Gloss Extraction systems to extract sense glosses that more closely match the meanings of a word. An important first step in this direction is to extract relations, rather than ngrams, that make up the definition a word's senses.

Acknowledgments

Presentation of this work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080157 to Carnegie Mellon University. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. The authors thank the anonymous reviewers for their helpful suggestions and comments.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval 2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval)*, pages 7–12.

E. Agirre, O. Lopez de Lacalle, D. Martinez, and A. Soroa. 2006. Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm. In *Proceedings of the NAACL Textgraphs Workshop*.

I. Androutsopoulos and D. Galanis. 2005. A practically unsupervised learning method to identify single-snippet answers to definition questions on the web. In *Proceedings of HLT-EMNLP*, pages 323–330.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Empirical Methods in Natural Language Processing*.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

T. Chklovski and R. Mihalcea. 2002. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*.

H. Cui, M.K. Kan, and T.S. Chua. 2007. Soft pattern matching models for definitional question answering. *ACM Trans. Information Systems*, 25(2):1–30.

Mona Diab. 2004. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the ACL*.

Weisi Duan, Min Song, and Alexander Yates. 2009. Fast max-margin clustering for unsupervised word sense disambiguation in biomedical texts. *BMC Bioinformatics*, 10(S3)(S4).

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

A. Fujii and T. Ishikawa. 2000. Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of ACL*, pages 488–495.

M. Joshi, T. Pedersen, and R. Maclin. 2005. A comparative study of support vector machines applied to the supervised word sense disambiguation problem in the medical domain. In *Proceedings of the Second Indian International Conference on Artificial Intelligence*.

Anagha Kulkarni and Ted Pedersen. 2005. Name discrimination and email clustering using unsupervised clustering and labeling of similar contexts. In *Proceedings of the Second Indian International Conference on Artificial Intelligence*, pages 703–722.

Gondy Leroy and Thomas C. Rindflesch. 2004. Using symbolic knowledge in the umls to disambiguate words in small datasets with a naive bayes classifier. In *MEDINFO*.

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

Hongfang Liu, Virginia Teller, and Carol Friedman. 2004. A multi-aspect comparison study of supervised word sense disambiguation. *Journal of the American Medical Informatics Association*, 11:320–331.

Rada Mihalcea. 2002. Bootstrapping large sense-tagged corpora. In *International Conference on Languages Resources and Evaluations (LREC)*.

Rada Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *Proceedings of the NAACL*.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM)*.

S. Mohammad and Ted Pedersen. 2004. Combining lexical and syntactic features for supervised word sense disambiguation. In *Proceedings of CoNLL*.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Procs. of ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*.

Marius Pasca. 2005. Finding instance names and alternative glosses on the web: WordNet reloaded. In *Computational Linguistics and Intelligent Text Processing*, pages 280–292. Springer Berlin / Heidelberg.

Ana-Maria Popescu, Alexander Yates, and Oren Etzioni. 2004. Class extraction from the world wide web. In *AAAI-04 ATEM Workshop*, pages 65–70.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval)*.

B.J. Schijvenaars, B. Mons, M. Weeber, M.J. Schuemie, E.M. van Mulligen, H.M. Wain, and J.A. Kors. 2005. Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics*, 6.

R. Snow, D. Jurafsky, and A. Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *COLING/ACL*.

M. Stevenson and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

Paola Velardi, Roberto Navigli, and Pierluigi D'Amadio. 2008. Mining the web to create specialized glossaries. *IEEE Intelligent Systems*, 23(5):18–25.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the ACL*.