A Quantitative Assessment of SENSATIONAL with an Exploration of Its Applications

Wei Xiong, Min Song, Lori Watrous-deVersterre

Department of Information Systems
New Jersey Institute of Technology
University Heights, Newark, NJ 07102, USA
{wx7, min.song, llw2}@njit.edu

Abstract

Word sense disambiguation is the problem of selecting a sense for a word from a set of predefined possibilities. This is a significant problem in the biomedical domain where a single word may be used to describe a gene, protein, or abbreviation. In this paper, we evaluate SENSATIONAL, a novel unsupervised WSD technique, in comparison with two popular learning algorithms, support vector machines (SVM) and K means. Based on the accuracy measure, our results show that SENSATIONAL outperforms SVM and K means by 2% and 17% respectively. In addition, we develop a polysemy based search engine and an experimental visualization application that utilizes SENSATIONAL clustering technique.

1. Introduction

SENSATIONAL is a novel unsupervised Word Sense Disambiguation (WSD) technique proposed recently by (Duan, Song, and Yates 2009), whose original study presented impressive accuracy results. Our research contributes by benchmarking SENSEATIONAL against two well-received algorithms, Support Vector Machines (SVM) and K-means. Furthermore, we discuss SENSATIONAL's data preprocessing benefits related to reduced manual effort. These characteristics make SENSATIONAL's novel approach to WSD a very attractive application to a number of real-world problems in the area of search and data visualization that our research piloted for further exploration.

In the biomedical domain, WSD is a central problem. Many names of proteins and genes, abbreviations, and general biomedical terms have multiple meanings. These ambiguous words make it difficult for NLP applications and, in some cases, humans to correctly interpret the appropriate meaning.

In general terms, WSD involves the problem of determining the correct meaning an ambiguous word bears in a given context. This has been regarded as a crucial problem in many natural language processing (NLP) applications. For example, an information retrieval system could perform better if the ambiguities among queries were reduced. Other applied NLP applications that have benefited from WSD include information extraction (Stokoe, Oakes and Tait 2003), question answering (Pasca and Harabagiu 2001), and machine translation (Vickrey et al. 2005).

There are three types of WSD techniques (Ide and Veronis 1998): supervised learning, unsupervised learning and knowledge-based WSD. Supervised techniques need manually-labeled examples for each ambiguous term in the data set to predict the correct sense of the same word in a new context. This is referred to as training material so their corpus may build up a classification scheme based on this set of feature-encoded inputs and their appropriate sense label or category.

Knowledge-based WSD systems are similar to supervised learning because they use established, external knowledge, such as databases and dictionaries to disambiguate words. However, both of these approaches need extensive manual effort to create external resources. This can be time-consuming and expensive.

Unsupervised WSD techniques do not require the creation of these training sets or predefined knowledge bases. Instead, they are based on unlabeled corpora and use a training set of feature-encoded inputs but do not have these mapped to an appropriate sense label or category. While this technique reduces the manual, time-consuming effort, unsupervised WSD often generates less accurate results (Ide and Veronis 1998).

To respond to the costly manual effort while maintaining a high accuracy, (Duan, Song, and Yates 2009) proposed a novel unsupervised WSD technique, called SENSATIONAL. It uses a single parameter which is not tied to the vocabulary thus making the resulting system extendable to new terms. This means the system does not have to be retrained when ported to a new document collection. While the technique requires a small sample data set for initialization, the effort is smaller compared to supervised WSD techniques. SENSATIONAL is a relatively new technique with limited literature to compare it with other, more popular WSD techniques. Our object is to compare SENSATIONAL with two well-received WSD techniques, one of which is SVM, and the second which is unsupervised, K-means.

The paper continues with a description of related work in Section 2. In Section 3 we introduce the SENSATIONAL algorithm for clustering word senses and provide a search engine and an experimental visualization application that utilizes this automatic clustering technique. Section 4 outlines our experimental setup and presents results. We discuss the future work for this ongoing research in Section 5.

2. Related work

In this section, we will review Word Sense Disambiguation techniques that are widely used in the biomedical domain.

2.1. Supervised learning techniques

One of the most popular supervised learning techniques used for WSD is Support Vector Machines (SVMs) (Vapnik 1995). (Joshi, Pedersen, Maclin 2005) compare SVM with other four well-known supervised learning algorithms: Naïve Bayes, decision trees, decision lists and boosting approaches on a subset of the NLM-WSD data set. They converted the NLM formatted data into SENSEVAL-2 format data which is an XML format with certain pre-defined markup tags. The statistical significance test of the log likelihood measure was employed to identify bigrams that occur together more often than by chance. Their evaluation results indicated that SVM obtained the best performance with unigram features selected using a frequency cut-off of four.

Naïve Bayes is another popular supervised learning technique widely used in biomedical domain. (Leroy and Rindflesch 2005) use the naïve Bayes classifier from the Weka data mining suite. Their experiments were performed with incremental feature sets, thus evaluating the contribution of new features over the previous ones. They achieved convincing improvements over the majority sense baseline in some cases, but observed degradation of performance in others. A comparative study conducted by (Pedersen and Bruce 1997) also shows that the Naïve Bayes classifier achieves a high level of accuracy using a model of low complexity.

Decision list learning is a rule-based approach. (Frank and Witten 1998) propose an approach for learning decision lists based on the repeated generation of partial decision trees in a 'separate-and-conquer' manner. They demonstrate rule sets can be learned one rule at a time without any need for global optimization. Decision Lists were one of the most successful systems on the 1st Senseval competition for WSD (Kilgarriff and Rosenzweig 2000).

Boosting approach is based on the observation that finding many rough rules of thumb can be much easier than finding a single, highly accurate predication rule (Schapire 2003). (Escudero, Marquez and Rigau 2000) apply the boosting algorithm to WSD problem, and compare it with Naive Bayes and Exemplar-based approaches. Their experiments on a set of 15 selected polysemous words show that the boosting approach outperforms its rivals.

2.2. Unsupervised learning techniques

The K-means clustering (MacQueen 1967) is a common clustering algorithm used to automatically partition a data set into k groups. (Schütze 1998) proposed an unsupervised technique for word sense disambiguation based on a vector representation of word senses that were induced from a corpus without labeled training instances or other external knowledge sources. The K-means vector model was used and demonstrated good performance of context-group discrimination for a sample of natural and artificial ambiguous words.

(Bhattacharya, Getoor and Bengio 2004) propose two unsupervised WSD systems: 'Sense Model' and 'Concept Model'. Their experimental results show that the concept model improved performance on the word sense disambiguation task over the previous approaches participated in 21 Senseval-2 English All Word competition. (Yarowsky 1995) propose an unsupervised learning algorithm to perform WSD, which is based on two powerful constraints: that words tend to have one sense per discourse and one sense per collocation. When trained on unannotated English text, the experimental results indicate that his algorithm is able to compete with some unsupervised learning technique.

An unsupervised approach for WSD which exploits translation correspondences in parallel corpora is presented by (Diab and Resnik 2002). The idea is that words having the same translation often share some dimension of meaning, leading to an algorithm in which the correct sense of a word is reinforced by the semantic similarity of other words with which it shares those dimensions of meaning. Based on fair comparison using community-wide test data, the performance of their algorithm has been evaluated.

2.3. Knowledge-based WSD

The availability of extensive knowledge source such as Unified Medical Language System (UMLS) and WordNet has been widely utilized to tackle WSD problem. (Widdows et al. 2003) propose their system for word sense disambiguation of English and German medical documents using UMLS. (Liu, Johnson and Friedman 2002) used UMLS as the ontology and identified UMLS concepts in abstracts and analyzed the co-occurrence of these terms with the term to be disambiguated. (Leroy and Rindflesch 2005) studies the effect of different types of symbolic information for terms in medical text by mapping sentences to the UMLS. They use Naïve Bayes classifier to disambiguate medical terms and the UMLS for its symbolic knowledge. (Mihalcea and Moldovan 1998) present a method for WSD that is based on measuring the conceptual density between words using WordNet. (Inkpen and Hirst 2003) use WordNet to disambiguate near-synonyms in dictionary entries. Their approach is based on the overlap of words in the dictionary description and the WordNet glosses, synsets, antonyms, and polysemy information.

3. SENSATIONAL and its Application

This section provides an overview of SENSATIONAL's clustering algorithm. It also suggests a polysemy-based search engine and an experimental visualization application that utilizes SENSATIONAL clustering technique.

3.1. SENSATIONAL Clustering

(Duan, Song, and Yates 2009) proposed a novel and efficient graph-based algorithm to cluster words into groups that have the same meaning. Their system, called SENSATIONAL, is built on the principle of margin-maximization. This principle finds a surface in the space of data points that separates the points in such a way that maximizes the smallest distance between points on opposite sides of the surface. Although it works accurately and effectively, max-margin clustering is computationally expensive.

To overcome this drawback, Duan et al. used a novel approximation algorithm for finding max-margin clusters in a document collection based on minimum spanning trees (MST). A MST is an undirected graph that can be computed efficiently and provides the smallest set of edges in the tree to connect all the data points together. A second characteristic of a MST is that for each data point, or node, there will be exactly one path to every other node. Therefore, we can divide a MST into sub-graphs by simply removing one edge.

To use MST in max-margin cluster, each word in a set of documents containing that word is stored in an *index*, which can be used to prune the set of edges that are added to a graph. After a weighted graph that represents the set of mentions of an ambiguous term is built, the MST for the graph is constructed, and the largest edge is removed from it. This splits the graph into two clusters which can then be further analyzed to determine if additional segmentation is needed. The largest edge of the MST provides a large margin between two clusters of the weighted graph. This insight forms the basis of the SENSATIONAL algorithm.

One weakness with this algorithm is the potential for outliers to skew the clusters. The SENSATIONAL system corrects this from happening by finding the roots of all the sub-trees and determines a path or backbone between these sub-trees. This Backbone-Finding algorithm identifies the core of the MST and helps prevent outliers skewing the clusters that are developed.

The beauty of SENSATIONAL is that the algorithm assumes essentially no input requiring significant manual input to construct, such as manually-labeled training examples for supervised learning algorithms or manually-constructed and manually-curated databases containing structured knowledge. It needs a single free parameter, which is essentially threshold that could be set by hand, but also could be trained with only a few hundred manually-labeled examples of one ambiguous term in context. This parameter is not tied to the vocabulary, and the system does not have to be retrained when ported to a new document collection. This means, after being trained, the system can be applied to any new term. We consider SENSATIONAL as an unsupervised system because it is common to refer to systems with a small number of free parameters as "unsupervised" (Davidov and Rappaport 2008).

3.2. Polysemy Extraction-based Search Engine

SENSATIONAL's benchmarking results and data preprocessing features make it a very attractive technique for a number of real-world problems. The following described a prototype of two novel approaches to search and data visualization.

In the case of search, it is well known that as our corpora continually grow, the current retrieval systems, modeled after card catalog systems, will become cumbersome and inefficient (Korfhage 1991). New ways to determine what end-users are searching for need to be developed that don't just present all possible query keyword matches, but also attempts to provide them with information to support iterative search requests that will isolate the documents relevant to their particular inquiry. This iterative 3-step process starts with a broad query which, in successive queries drills-down by either zooming in on wanted concepts or filtering out unwanted concepts. Finally details on specific content are examined and the process begins again (Koshman 2006).

Users search for different reasons. Two basic purposes are to find explicit information and to perform exploratory

browsing. When searching for explicit information, we may or may not know the specific terms to use. Initial results returned must be examined and if inadequate summaries provided, the user must open up documents and examine content. This can be time consuming and frustrating. Exploratory browsing can be even worse. As a user moves from document to document, the "bigger picture" or reason for the browsing can soon be forgotten leading to wasted time and effort. Even worse, potential new discoveries can be hidden in the maze of lengthy lists of keyword matched content.

An exploratory polysemy extraction-based search engine was built to determine if SENSATIONAL could be used to identify clusters based on existing corpora. Preliminary results show we can identify the clusters via the Backbone-Finding algorithm and present topical terms.

The following example is a visual representation for the query "cold". A search was conducted on the PubMed collection using the query, "cold". A limit was placed on the search to retrieve up to 300 documents. The polysemy-based search engine identified 4 different senses. Figure 1 is a display of the results which shows the number of resources associated with each sense along with a machine generated meaning of the sense based on latent semantic indexing. For example, the first sense, labeled, "vitamin common cold revisited", includes 148 documents related to managing a common cold while the second sense contains 11 documents describing an object used for pain relief.

This display can help a user focus their search which can reduce the amount of content an individual must examine. It may also help a user determine if a repository contains the type of content they are interested in. For users performing an explicit or exhaustive search, this format also provides information on additional appropriate terms.

Figure 1. Two-level search results display

If a user decides they are interested in information about managing the common cold they can drill down to discover the Cluster Map shown in Figure 2. The Map shows 3 subtopics which are defined in the ovals beginning with the topic identifiers 1, 2, and 3. The identifiers are followed by a set of terms generated by latent semantic indexing and end with an overall count of the number of documents associated with the topic id. Since documents may have an overlap between these subtopics, the Cluster Map shows this by color coding the groupings and indicating an overlap by having the connecting pathways having the different colors. In the case of label 2, all the documents are also grouped within label 3 while label 1's document set has 7 of its documents overlapping with subtopic 3 and a single outlier shown at the bottom of the display not associated at all.

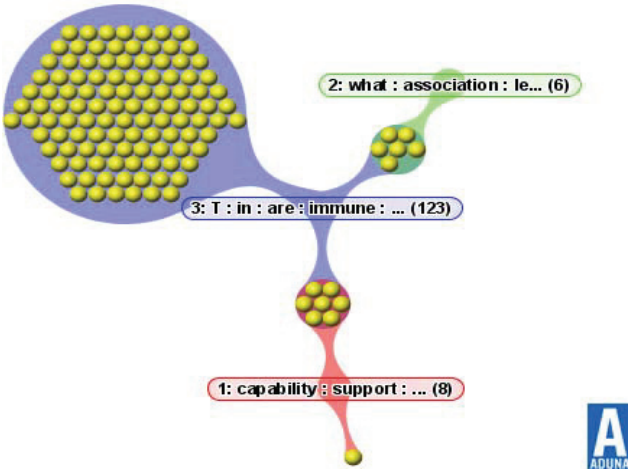

Figure 2. Visual representation of SENSATIONAL generated terms in existing corpora

This display not only identifies all the topical terms for a query but their inter-relationship within the collection being searched. This visual representation may help users understand meaning and relationships of ambiguous terms. For users unfamiliar with the multiple meanings of a term or performing exploratory browsing, this high-level overview provides an organizational structure to help users focus on the intent of their search before opening any documents to determine if it meets their needs or not.

4. Experiment and results

We performed an experiment to compare SENSATIONAL system against both SVM and K-means.

4.1. Experimental Setup

We evaluated SENSATIONAL, SVM and K-means on the same data set that (Duan, Song, and Yates 2009) used to evaluate SENSATIONAL and we show their results below. The data set of keywords is from the National Library of Medicine (NLM) data set, plus a set of additional terms, including a number of acronyms. They collected a data set of PubMed abstracts for these terms. On average, 271 documents per keyword were collected; no keyword had fewer than 125 documents, and the largest collection was 503 documents. They filtered out abstracts that were less

than 15 words and manually labeled each occurrence of each term with an identifier indicating its sense in the given context. They collected data for a total of 21 keywords. Two of these were used for training, and the other 19 for tests.

We used LIBSVM by Chih-Chung Chang and Chih-Jen Lin (available for download at http://www.csie.ntu.edu.tw/~cjlin/libsvm). (Xu et al. 2006) applied SVM classifiers to perform WSD tasks on an automatically generated data set that contains ambiguous biomedical abbreviations. Their results indicated that there was no statistical difference between results when using a five-fold or ten-fold cross-validation method. In our case, for SVM we adopted the linear kernel and default parameter values, and ran a five-fold cross-validation.

4.2. Results

We evaluated SENSATIONAL against SVM and K-means based on the standard measure of accuracy, which is the percentage of the correctly classified instances.

Results for our comparison appear in Table 1. SENSATIONAL is able to outperform both SVM and K-means, by 2% and 17% on average across the keywords respectively. Considering that the size of our data set of keywords is smaller than 30, we performed K-S test for normality of distributions and the results (significance level > 0.05) suggest that the distributions are normal. The performance of SENSATIONAL compared to K-means and SVM is statistically significant at $p<0.05$ and $p=0.495$ respectively, using a paired t-test.

Table 1: Comparison with SVM and K-means

Keyword	SVM	K-means	Sensational
ANA	0.82	0.72	1.0
BPD	0.97	0.42	0.53
BSA	0.99	1.0	0.95
CML	0.99	0.60	0.90
cold	0.68	0.45	0.67
culture	0.59	0.58	0.82
discharge	0.71	0.43	0.95
fat	0.51	0.53	0.53
fluid	0.92	0.60	0.99
glucose	0.51	0.58	0.51
inflammation	0.42	0.45	0.50
inhibition	0.50	0.67	0.54
MAS	1.0	0.51	1.0
mole	0.78	0.53	0.96
nutrition	0.53	0.44	0.55
pressure	0.82	0.68	0.86
single	0.95	0.84	0.99
transport	0.51	0.52	0.57
VCR	0.80	0.64	0.64
average	**0.74**	**0.59**	**0.76**

It is interesting to notice that some ambiguous words were more troublesome to the classifiers than others. Most words only had 2 senses in the data, with four exceptions: "BPD", "cold", "inflammation", and "nutrition" had 3 senses each. The WSD performance of these exceptions was generally poorer than others, which confirmed that the number of senses could be one of the determinants of the word ambiguity (Leroy and Rindflesch 2005).

5. Discussion and future work

We evaluated SENSATIONAL, a novel unsupervised WSD technique, in comparison with two popular learning algorithms, SVM and K-means. We manually curated the data set collected from National Library of Medicine (NLM). In addition, we develop a polysemy-based search engine and an experimental visualization application that utilizes SENSATIONAL clustering technique. These applications could help users understand the meaning and relationships of ambiguous terms, choose their next level of search and reduce the amount of content the individuals must wade through to find what is relevant to them.

The experiment that we have performed demonstrated that compared with K-means, SENSATIONAL is able to achieve a better accuracy. It outperforms K-means by 17%. However, compared with SVM, the performance of SENSATIONAL is not statistical significant, given that the significance level $p=0.495$. In addition, the performance of SVM could be improved by optimizing its parameter values.

So far the performance of SENSATIONAL is very encouraging, given that it assumes essentially no inputs that require significant manual input to construct. SENSATONAL's Max-margin technique combined with its Backbone-Finding algorithm is not only able to outperform the state-of-the-art unsupervised WSD technique, but also competitive against supervised learning algorithm. In future work, we plan to examine if the good performance of SENSATIONAL in medical domain will translate into general English word sense disambiguation, especially compared with other well-known machine learning algorithms.

6. References

Duan, W., Song, M. and Yates A. 2009. Fast max-margin clustering for unsupervised word sense disambiguation in biomedical texts. *BMC Bioinformatics* 10: S4.

Ide, N., and Véronis, J. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics* 24: 2–40.

Stokoe, C.; Oakes, M. P.; and Tait, J. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th annual international ACM SIGIR*

conference on Research and development in informaion retrieval 166.

Pasca, M. and Harabagiu, S. 2001. The informative role of WordNet in open-domain question answering. *Workshop on WordNet and Other Lexical resources at NAACL*.

Vapnik, V. 1995. *The Nature of Statistical Learning* Theory. Spring, New York.

Vickrey, D.; Biewald, L.; Teyssier, M.; and Koller, D.2005. Word-sense disambiguation for machine translation. In *Proceedings of HLT/EMNLP* vol. 5.

Joshi, M.; Pedersen, T.; and Maclin, R. 2005. A comparative study of support vector machines applied to the supervised word sense disambiguation problem in the medical domain. In *Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI'05)*, 3449–3468.

Leroy, G., and Rindflesch, T. C. 2005. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *International Journal of Medical Informatics* 74: 573–585.

Pedersen, T., and Bruce, R. 1997. Knowledge lean word-sense disambiguation. In *Proceedings of the National Conference of Artificial Intelligence*, 814–814.

Frank, E., and Witten, I. H. 1998. Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 144–151.

Kilgarriff, A., and Rosenzweig, J. 2000. English SENSEVAL: Report and results. *LREC, Athens*, 265–283.

Schapire, R. E. 2003. The boosting approach to machine learning: An overview. *LECTURE NOTES IN STATISTICS NEW YORK SPRINGER VERLAG*, 149–172.

Escudero, G.; Marquez, L.; and Rigau, G. 2000. Boosting applied to word sense disambiguation. In *Proceedings of ECML 00, 11th European Conference on Machine Learning* (Barcelona, Spain, 2000), 129–141.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics* 24: 97–123.

Bhattacharya, I.; Getoor, L.; and Bengio, Y. 2004. Unsupervised sense disambiguation using bilingual probabilistic models. *Proceedings of the Meeting of the Association for Computational Linguistics*.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* 189–196.

Diab, M. and Resnik, P. 2002. An unsupervised method for word sense tagging using parallel corpora. *Proc. ACL*.

Wilks, Y.; Fass, D.; Guo, C.; MacDonald, J.; Plate, T.; and Slator, B. 1990. Providing Machine Tractable Dictionary Tools, MIT Press.

Harley, A., and Glennon, D. 1997. Sense tagging in action. In *Proceedings of SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How*.

Humphrey, S. M.; Rogers, W. J.; Kilicoglu, H.; Demner-Fushman D.; and Rindflesch, T. C. 2006. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *J. Am. Soc. Inf. Sci.* 57: 96-113.

Widdows, D.; Peters, S.; Cederberg, S.; Chan, C. K.; and Steffen, D. 2003. Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using UMLS. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine Volume 13* 9–16.

Liu, H.; Johnson, S. B.; and Friedman, C. 2002. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *Journal of the American Medical Informatics Association* 9: 621.

Mihalcea, R., and Moldovan, D. 1998. Word sense disambiguation based on semantic density. In *Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing* 16–22.

Inkpen, D. Z., and Hirst, G. 2003. Automatic sense disambiguation of the near-synonyms in a dictionary entry. *Lecture Notes in Computer Science* 258–267.

Xu, H., Markatou, M.; Dimova, R.; Liu, H.; and Friedman, C. 2006. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC bioinformatics* 7: 334.

Davidov, D. and Rappaport, A. 2008 Unsupervised Discovery of Generic Relationships Using Pattern Clusters and its Evaluation by Automatically Generated SAT Analogy Questions. *Proceedings of the Annual Meeting of the Association of Computational Linguistics.*

Korfhage, R.R. 1991. To see, or not to see – IS that the query? *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM: 134-141.

Koshman, S. 2006. Visualization-based information retrieval on the Web. Library & Information Science Research 28(2): 192-207