Making Sense Out of the Web

Rada Mihalcea

University of North Texas
Department of Computer Science
rada@cs.unt.edu

Abstract. In the past few years, we have witnessed a tremendous growth of the World Wide Web, both in terms of number of Web pages accessible online – resulting in what represents today the largest publicly available corpus, and in terms of number of Web users – who now form the world's largest pool of knowledge. The growth of the Web on these two main dimensions – pages and users – has opened the doors to a realm of new approaches to data-hungry and knowledge-hungry language processing applications. Among these, Word Sense Disambiguation is one of the applications that has the potential of benefiting the most from the large amounts of Web-based data and from the availability of inexpensive Web user supervision. In this paper, I overview the main lines of research in deriving efficient Word Sense Disambiguation methods that exploit: (1) the *Web as a corpus* – which represents a view of the Web seen as an enormous collection of Web pages; and (2) the *Web as collective mind* – where the Web is regarded as a large group of Web users who can contribute their knowledge to the process of identifying word meanings.

1 Introduction

One notoriously difficult problem in understanding text is Word Sense Disambiguation (WSD). Ambiguity is very common, especially among the most common words – think for instance of *table*, or *computer fan* – but humans are so good at figuring it out from context that they usually do not even notice it. While a large number of efficient WSD algorithms have been designed and implemented to date within the recent SENSEVAL evaluation frameworks and elsewhere, the availability of sense tagged data still represents a significant problem in the development of accurate large scale WSD systems.

Most of the efforts in WSD were concentrated so far toward *supervised* learning algorithms, which usually achieve the best performance at the cost of low recall. In this class of algorithms, all sense tagged occurrences of a particular word are transformed into feature vectors, which are then used to build a classifier that can automatically identify the correct meaning of the word in any new unseen text. The main drawback of these methods is however the lack of widely available semantically tagged corpora and the strong dependence between the disambiguation accuracy and the size of training data. For instance, one study reports that high precision WSD requires an average of 500 examples per ambiguous word, depending on the word entropy [9]. At a throughput of one tagged

example per minute [5], and with about 20,000 ambiguous words in the common English vocabulary, a simple calculation leads to about 160,000 hours of tagging, which is nothing less than 80 man-years of human annotation work. Moreover, the size of the problem multiplies by an order of magnitude when languages other than English are considered. The study of a new language (among the approximately 7,000 languages spoken worldwide) implies a similar amount of work to label the data required to build a supervised WSD system for the new language.

In this paper, I overview work done for overcoming the data bottleneck problem faced by many WSD systems, which exploits: (1) the *Web as a corpus* – where the Web is seen as a very large collection of Web pages, and (2) the *Web as collective mind* – an alternative view of the Web, seen as a group of Web users who can contribute their knowledge to the WSD process.

2 The Web as a Corpus

With more than four billion pages currently indexed by search engines, in more than 1,500 languages, the Web represents today the largest textual corpus publicly available. In addition to its main asset – the size, the Web has several other appealing features, such as free access, good coverage of various genres, domains, and languages, and efficient access via fast search engines such as Google or AltaVista, which makes this corpus the resource of choice for solving the data bottleneck problem for a large range of language processing applications.

Some of the first attempts for using the Web as a textual corpus [7], [8] were concerned in fact with the WSD problem – as one of the language processing tasks facing some of the most acute data bottleneck problems. As mentioned earlier, WSD is a particularly difficult learning problem since it requires the construction of a large number of classifiers (one classifier for each ambiguous word), with each classifier requiring a large number of annotated examples. Being a data-hungry application, WSD has the potential of benefiting the most from the large amounts of data available on the Web. In this section, I overview three main research directions in using the Web as a corpus for WSD: (1) The collection of Web counts starting with carefully constructed search engine queries, to the end of identifying the most likely meaning for a given ambiguous word; (2) The unsupervised construction of semantically annotated corpora using words or word phrases that are semantically related to an ambiguous word and at the same time exhibit only one possible sense (monosemous relatives); and (3) The automatic bootstrapping of large sense tagged corpora starting with few sense-centric seed examples provided for a given ambiguous word.

2.1 Web counts for Word Sense Disambiguation

Corpus counts in general, and Web counts in particular, represent a powerful method for finding the meaning of a word, by identifying the most frequently used word sense given its surrounding context. Previous work in using Web

counts for WSD [7] modeled the *context* of an ambiguous word using syntactic dependencies, e.g. verb–noun, adjective–noun, etc. Given such a word–word pair, the algorithm proposed in [7] attempts to identify the meaning of the two words by following three main steps.

First, for each possible meaning of the ambiguous word currently considered, a list of semantically related words is determined using WordNet synonyms and hypernyms. Next, a search engine query is formed for each word meaning, consisting of the words in the similarity list connected with the other word in the pair using a *phrase*, AND, or NEAR relation[1]. Finally, the queries are run against the Web corpus using a search engine, and counts are collected for each query. The word meaning corresponding to the query leading to the highest count is selected as the correct one.

In the experiments reported in [7], the use of Web counts for WSD was found to lead to good disambiguation results – with precision figures in the range of 60–80% obtained on a large set of ambiguous word pairs extracted from SemCor. Similar experiments were later reported in [2] with consistently good results obtained on a different WSD data set.

2.2 Unsupervised Construction of Sense Annotated Corpora Using Monosemous Relatives

The amount of manually annotated data required to sustain accurate word sense classifiers represents a serious impediment in applying supervised learning methods to WSD. One possible solution for overcoming this drawback is to use the Web to automatically build large sense tagged corpora, using the approach suggested in [8], Their approach is mainly based on (1) information provided by WordNet, in particular semantically related words and short word definitions identified within glosses, and (2) information gathered from the Web using existing search engines. The information obtained from WordNet is used to formulate queries consisting of synonyms or short definitions that are representative for the meaning of a given ambiguous word, while the Web is used as a resource to extract text snippets relevant to such queries.

Given a word for which a sense annotated corpus is required, the first step consists of identifying the possible senses that the word might have with respect to a given sense inventory – e.g. WordNet, in the case of the experiments reported in [8]. Next, for each possible sense, a list of semantically related words or word phrases is constructed by using either the monosemous synonyms in the word synset – if such synonyms exist, or by using short representative word phrases extracted from the gloss definitions. The monosemous synonyms or the short definitional phrases constitute the basis for creating queries used to find text snippets on the Web, which are later used to extract sentences containing the search phrase. Next, in each of these sentences, the search phrase is replaced with the original word annotated with the corresponding sense, therefore resulting in

[1] The experiments reported in [7] relied on the AltaVista search engine and the *phrase*, AND, and NEAR search operators provided by this engine.

a corpus of example sentences that contain sense annotated occurrences for the given ambiguous word.

Experiments on a test set of 20 ambiguous words, as reported in [8], show that using this method more than 80,000 sense tagged examples could be automatically acquired, with an average annotation accuracy of about 91%. Moreover, recent experiments reported in [1] showed that a sense tagged corpus automatically constructed in this way can be used to build competitive WSD systems.

2.3 Web-based Bootstrapping of Large Sense Tagged Corpora

Another Web-based approach for tackling the WSD problem targets the construction of large sense tagged corpora starting with a few sense-centric seeds. The iterative generation algorithm proposed in [6] builds a large sense annotated data set by following the principles of a bootstrapping algorithm: An initial set of seeds is used (1) to extract text snippets from the Web, which are then added to the sense tagged corpus, and (2) to identify other instances of ambiguous words that can be accurately sense tagged. The newly tagged words are added to the set of seeds and the generation process continues.

The initial set of seeds is formed using existing sense tagged data – such as SemCor, or data from the SENSEVAL evaluations – or it is manually constructed. The seeds are merely formed as multiple word units that include an ambiguous word, such that the expression itself places a constraint over the possible meaning for the word of interest.

The generation algorithm was evaluated in two ways. First, the bootstrapping algorithm was used to create a corpus that was used in a disambiguation system participating in the SENSEVAL-2 English all words task, with significant improvements measured over the baseline performance. Second, the algorithm was used to build a sense tagged corpus for a subset of randomly selected words from the SENSEVAL-2 English lexical sample task. The disambiguation results obtained with a WSD system trained on the generated corpus were found to be comparable, and sometimes even better than those achieved with the same WSD system trained on manually labeled data, at a significantly lower annotation cost.

This line of work is related to work previously reported in [10], where few tagged seeds are used to train a decision list employed to tag new unlabeled instances. More recently, [4] analyses the various factors that affect the learning performance in the presence of automatically generated noisy sense tagged data, and shows that a supervised learning WSD system can be successfully bootstrapped starting with an unsupervised seed set.

3 The Web as Collective Mind

The Web is not only a collection of Web pages, but it is also a network of Web users who can contribute their knowledge to the data annotation process. This alternative view of the Web, seen as *collective mind*, represents the basis for another approach for building large annotated corpora, by tapping into the

world's vast pool of knowledge, and relying on a large number of Web users to build the sense annotated data required for efficient WSD methods.

To overcome the current lack of sense tagged data and the limitations imposed by the creation of such data using trained lexicographers, the *Open Mind Word Expert* system proposed in [3] enables the collection of semantically annotated corpora using volunteer contributions over the Web. Open Mind Word Expert is one of the Web-based knowledge capture systems developed under the Teach-Computers project (`http://teach-computers.org`), designed to help computers to solve the WSD problem. Any Web user can visit the Open Mind Word Expert site and contribute some knowledge about the meanings of given words in given sentences. As a result, Open Mind Word Expert creates large, sense tagged corpora that can be used to build automatic WSD systems.

When contributors visit the Open Mind Word Expert site, they are presented with a set of natural language (e.g., English) sentences that include an instance of the ambiguous word, and are asked to indicate the most appropriate meaning with respect to the definitions provided. Hundreds of thousands of tags have been collected since the site's launch two years ago. Annotations are currently being collected for building word sense disambiguation tools for English, Romanian, and for creating English-Hindi translation tools. The data collected so far is publicly available, and has been used in several tasks during the recent SENSEVAL-3 evaluations.

4 Conclusion

The Web is arguably the largest textual corpus available today – consisting of billions of Web pages that are accessible online, and at the same time it is the largest "public" pool of knowledge – formed by the large community of Web users who can contribute their knowledge to various Web-based activities. In this paper, I overviewed the main lines of research in exploiting the Web as a resource to overcome the acute data bottleneck faced by many WSD systems. When seen as a large textual *corpus*, the Web can be used to derive efficient methods for collecting Web counts, or for building large sense annotated corpora. When regarded as *collective mind*, the large community of Web users can contribute their knowledge to the process of identifying word meanings and can help constructing large semantically annotated data. Experiments reported along these lines have proved that the Web – seen either as a *corpus*, or as *collective mind* – represents an invaluable resource for tackling the WSD data bottleneck problem and for building accurate large scale sense classifiers.

References

1. AGIRRE, E., AND MARTINEZ, D. Unsupervised word sense disambiguation based on automatically retrieved examples: The importance of bias. In *Proceedings of EMNLP 2004* (Barcelona, Spain, July 2004).

2. Chao, G., and Dyer, M. Word sense disambiguation of adjectives using probabilistic networks. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)* (Saarbrucken, 2000).
3. Chklovski, T., and Mihalcea, R. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of the ACL 2002 Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"* (Philadelphia, July 2002).
4. Diab, M. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL 2004)* (Barcelona, Spain, July 2004).
5. Edmonds, P. Designing a task for Senseval-2, May 2000. Available online at http://www.itri.bton.ac.uk/events/senseval.
6. Mihalcea, R. Bootstrapping large sense tagged corpora. In *Proceedings of the Third International Conference on Language Resources and Evaluation LREC 2002* (Canary Islands, Spain, May 2002), pp. 1407–1411.
7. Mihalcea, R., and Moldovan, D. An automatic method for generating sense tagged corpora. In *Proceedings of AAAI-99* (Orlando, FL, July 1999), pp. 461–466.
8. Mihalcea, R., and Moldovan, D. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)* (College Park, MA, June 1999), pp. 152–158.
9. Ng, H. Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* (Washington, 1997), pp. 1–7.
10. Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)* (Cambridge, MA, June 1995).

This article was processed using the LaTeX macro package with LLNCS style