Chapter 6: Disambiguation

55-60

6.1 Introduction

Arthur Weasley1: "Now, Harry you must know all about Muggles2, tell me, what exactly is the function of a rubber duck?"

—JK Rowling (2002)

Harry Potter And The Chamber Of Secrets

Word sense disambiguation is the task of selecting the appropriate sense of a word in a given context. In order to do so we need to train a disambiguator to weight the context features to accurately distinguish the given word senses. Disambiguators have been constructed by two approaches: manually or by machine learning. The machine learning approach can be divided into supervised, semi-supervised and unsupervised approaches. All of them are faced with the

same task: given two or more different meanings or word senses, they need to find distinctive contexts or usages for each sense. First we give a brief overview of the ideas that were introduced already in the 50's and are still used in word sense disambiguation. For a more in-depth historical overview of how the ideas developed, we refer the reader to Ide and Veronis (1998) and Rapaport (2005)[55]. Then we look at how the different approaches to machine learning are applied to word sense disambiguation. Finally, we give an outline of the ideas and the contribution of this work to word sense disambiguation.

6.2 Manually Developed Resources

The main ideas for manually developed resources in word sense disambiguation has evolved into two broad categories, i.e., artificial intelligence-based and knowledge-based ideas. First we will look at some of their common early ideas following the outline of Ide and Veronis (1998)[56]. We will then turn to machine learning methods3, which are the bulk of the current systems.

6.3 Early Ideas

Many of the ideas which are reiterated and refined today in WSD were put forward already in the early days of computational linguistics during the 1950s. Kaplan (1950) observed that two words on either side of a word was not significantly better or worse than giving the entire sentence to human translators in order for them to do word sense

resolution. Reifler (1955) observed that the syntactic relations between words are often a decisive component when determining a word sense. At the same time, Weaver (1955) pointed out that a given word most often only has one meaning in a particular domain. (Ide and Veronis, 1998) Following the observation that a word generally has only one meaning in one domain, many natural language processing systems still today use domain-specific dictionaries or microglossaries as the main solution for tailoring the process to a specific domain. As Weaver emphasized in his Memorandum (Weaver, 1955): This approach brings into the foreground an aspect of the matter which is absolutely basic—namely, the statistical character of the problem And it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken as a necessary primary step.

This was also pursued by various researchers, and Pimsleur (1957) introduced the notion of levels of depth for a translation corresponding to what today is known as the most frequent sense or baseline tagging. Level 1 uses the most frequent equivalent, producing 80% correct translations, level 2 contains additional meanings, producing 90% correct translations, etc. (Ide and Veronis, 1998[58]) As Ide and Veronis (1998) point out in their survey, many of the fundamental

ideas of word sense disambiguation that have resurfaced several decades later, were originally tested only on a small scale due to severe constraints on the available resources, i.e., computers, corpora and machine-readable dictionaries. Even though the fundamental ideas themselves are not new, it is still interesting to revisit them in light of recent developments in computational linguistics and soft computing, see Baayen (2001) or Manning and Sch¨utze (1999).

6.4 Artificial Intelligence Ideas

The idea of interlingua was also proposed in the 1950s leading to semantic networks by Richens (1958) and Masterman (1962) using a language-independent semantic network of concepts, onto which the words could be mapped. AI methods such as semantic networks were used for disambiguation by trying to find the shortest path through a common concept for two words. One of the problems with the AI methods was the knowledge acquisition bottleneck. (Ide and Veronis, 1998) The knowledge acquisition bottleneck is now being approached in a distributed effort all over the World Wide Web

under the title Semantic Web with the aim to create a compatible set of ontologies[59] or knowledge repositories which can be used for various language understanding and information processing tasks by man or machine. "The Semantic Web is a web of data, in some ways like a global database" as the effort is characterized by its inventor Tim Berners-Lee (1998, 2000). What seemed like a bottle-neck in the 1950s, because the information needed to be manually encoded in a structured way, now seems like it could be achievable due to a world-wide distributed encoding effort and exploited as sample contexts for machine learning methods in a not too distant future.

6.5 Knowledge-based Ideas

In order to remedy the knowledge acquisition bottleneck for natural language processing, several knowledge-based methods using machine-readable dictionaries and thesauruses have been investigated. The main problem with machine-readable dictionaries is that they are designed for humans. Often they are too detailed and not formal enough for machines. The most common way to disambiguate with machine-readable dictionaries is to select the word sense which according to some given criteria maximizes the relatedness among the word senses of the words co-occurring in a passage of text.

One of the most frequently used resources is WordNet (Fellbaum, 1998)[60], an electronic online thesaurus with more than 152,000 words of English. WordNet has served as a basis for creating corresponding WordNets in more than 35 languages (Vossen, 2001; Vossen and Fellbaum, 2004). WordNet may not have all the necessary information, but currently there are few other publicly available resources which could compete with it. Since it is the focus of such interest, it will hold its position and develop further for some time to come.

6.6 Machine Learning Methods

Statistical methods have become a standard paradigm in computational linguistics. They can be grouped roughly into descriptive statistics, generative models (i.e., stochastic finite-state, context-free, and attribute-value grammars), and machine learning methods (i.e., supervised, semi-supervised, and unsupervised methods). (Abney, 2005) Most current methods for word sense disambiguation use machine learning. Supervised methods require annotated corpora. Semi-supervised methods use mainly unannotated

corpora combined with some annotated data. Unsupervised methods get by on unannotated corpora combined with a thesaurus for a sense inventory. The trend is to use all the resources available (Agirre et al., 2000), and consequently some hybrid methods use unannotated corpora, thesauruses and annotated corpora when available. Annotated corpora are costly to develop, so the trend is also toward semi-supervised or unsupervised methods.

6.6.1 Supervised

Supervised machine learning[61] of classifiers uses a set of annotated training data from which a classifier is induced by an algorithm. The training algorithm is called supervised if it uses the annotated training data for improving the capacity of an classifier to reproduce the annotation. For an introduction to supervised machine learning, see Manning and Sch¨utze (1999) or Christianini and Shawe-Taylor (2000). For a description and evaluation of a number of supervised methods applied to word sense disambiguation, see Kilgarriff and Palmer (2000), Edmonds and Kilgarriff (2002) and Mihalcea and Edmonds (2004[62]).

6.6.2 Semi-Supervised

The semi-supervised or minimally supervised methods are gaining popularity because of their ability to get by with only a small amount of annotated reference data while often outperforming totally unsupervised methods on large data sets. There are a host of diverse methods and approaches, which learn important characteristics from auxiliary data and cluster or annotate data in light of the acquired information. The auxiliary data may provide seeds for labeling the primary data (Yarowsky, 1995; Blum and Mitchell, 1998; Nigam et al., 2000; Banko and Brill, 2001; Abney, 2002)[63], or it may provide some structure of its own, which guides the clustering of the primary data (Tishby et al., 1999; Kaski et al., 2005b,a).

6.6.3 Unsupervised

Typically, unsupervised methods use unannotated data sets, which mainly enable clustering of the data. In order to perform annotation they need to learn some annotation for the clusters. As long as they adopt the annotation without modifying the clustering, the algorithms are considered unsupervised. For an account of the unsupervised systems used in word sense disambiguation, see Kilgarriff and Palmer (2000), Edmonds and Kilgarriff (2002) and Mihalcea and Edmonds
(2004).

Following the observations of Kilgarriff that the most dominant sense is usually very dominant, cf. Section 2.4.1, it may be interesting to determine only the predominant sense of a word and then tag each word with this word sense. Mc- Carthy et al. (2004b) create a corpus-based similarity list for a target word. They use the list for ranking theWordNet senses of the target word, and identify the predominant sense. For some words the predominant word sense may vary according to domain as Ide and Veronis demonstrated, cf. Section 3.2, so the predominant word sense should be learned from a domain-specific corpus.

6.6.4 Combining classifiers

As classifiers usually are strong at different aspects of a task, it is often beneficial to combine them into ensembles. For different methods of combining classifiers for word sense disambiguation[64], see Florian et al. (2002) and Florian and Yarowsky (2002). Classifiers in an ensemble need to be different from each other, otherwise there is no gain in combining them. Unfortunately, there is no single best way for measuring classifier diversity (Brown et al., 2004).

6.6.5 Degrees of Supervision

We need to dwell on the difference between supervised and unsupervised training algorithms in the case where some classifier output is needed. If we transfer information from unannotated data to an existing classifier, this counts as unsupervised training. However, if we transfer information in the other direction using a classifier to improve the clustering of the data, the situation needs clarification. Many would agree that preprocessing for an unsupervised system is allowed if it does not add the annotation we

are later trying to discover. Post processing is also allowed, as long as it does not modify the clusters provided by the unsupervised learning. However, it is at least a mild form of supervision if an external classifier is used for determining or improving the clustering in order to better reproduce the annotation.

An external classifier may appear in many disguises: it may arrive in the form of a hand-made precompiled data structure with an elaborate access function, or it may be a list of annotated samples with a similarity function. In word sense disambiguation we have examples of both: a hand-made data structure with an elaborate access function is, e.g., the WordNet with the Lesk algorithm for semantic distance, whereas lists of annotated samples are provided by the SENSEVAL data.

SENSEVAL mainly uses the word senses provided by WordNet. All classifiers involved in word sense disambiguation must learn how to annotate an instance of a word in its context. The association can be learned from the WordNet hierarchies, from annotated WordNet synset glosses, or from a set of annotated samples. An unsupervised method may learn the annotation of a word belonging to a context cluster from any of these three sources. The SENSEVAL organizing committee divided the algorithms of the SENSEVAL-1 and SENSEVAL-2 Lexical Sample task into supervised and unsupervised, solely based on whether they use SENSEVAL training data or not. This may have been a clear-cut and easy-to-implement rule, but it put the unsupervised classifiers in an awkward position, because they did not even have access to unannotated data from the same domain as they are being evaluated on. From what we now know about domain influence on word sense distribution, this was an overly harsh constraint giving unfair advantage to the supervised methods. However, this changed with SENSEVAL-3, where the best unsupervised system (Ramakrishnan et al., 2004) in SENSEVAL-3 was allowed to use the full training data set by defining it as an extended set of WordNet glosses.