Applications of Word Sense Disambiguation: A Historical Perspective

Neetu Sharma[1], Prof. S. Niranjan[2]
[1,2]Department of Computer Science &Engineering,
Ganga Institute of Technology and Management,
Kablana, Jhajjar, Haryana, India

***Abstract*— Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a computational manner. WSD is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. We introduce the reader to the motivations for solving the ambiguity of words and provide a description of the task. We overview supervised, unsupervised,**

and knowledge-based approaches. Finally, applications, open problems, and future directions are discussed.

***Keywords*— WSD, NLP, MT, IR, QA**

I. INTRODUCTION

Word sense disambiguation is a core research problem in computational linguistics, which was recognized at the beginning of the scientific interest in machine translation and artificial intelligence a template. Lexical disambiguation in its broadest definition is nothing less than determining the meaning of every word in context, which appears to be a largely unconscious process in people. As a computational problem it is often described as “AI-complete”, that is, a problem whose solution presupposes a solution to complete natural-language understanding or common sense reasoning.

II. BRIEF HISTORY

In order to introduce current WSD research, we provide here a brief review of the history of WSD. WSD was first formulated as a distinct computational task during the early days of machine translation in the late 1940s, making it one of the oldest problems in computational linguistics. Weaver (1949) introduced the problem in his now famous memorandum on machine translation.

In addition to formulating the general methodology still applied today , Weaver acknowledged that context is crucial, and recognized the basic statistical character of the problem in proposing that “statistical semantic studies should be undertaken, as a necessary primary step”. The 1950s then saw much work in estimating the degree of ambiguity in texts and bilingual dictionaries, and applying simple statistical models.

Zipf (1949) published his “Law of Meaning”4 that accounts for the skewed distribution of words by number of senses, that is, that more frequent words have more senses than less frequent words in a power-law relationship; the relationship has been confirmed for the *British National Corpus* (Edmonds 2005). Kaplan (1950) determined that two words of context on either side of an ambiguous word was equivalent to a whole sentence of context in resolving power.

WSD was resurrected in the 1970s within artificial intelligence (AI) account for WSD. The system used selection restrictions and a frame-based developed “preference semantics”, one of the first systems to explicitly account for WSD.

III WHY WSD?

Why *do* so many NLP researchers and developers remain convinced that WSD *should* matter in NLP applications? There seem to be three main species of argument. A belief in the importance of WSD for applications is a part of the canon

in natural language processing. It is passed from teacher to student and easily accepted on intuitive grounds – it just seems obvious that if *bank* can refer to either a financial institution or a riverbank, a search engine query *must* be more likely to pull back the wrong documents, an MT system *must* be more likely to arrive at the wrong translation, and so forth,

unless the intended meaning of the word is picked from an enumerated list of the meanings it can have. Ide and Véronis (1998:1), in their valuable overview of sense disambiguation and its history, begin by saying that WSD is “obviously essential for language understanding applications such as message understanding, man-machine communication, etc.” and “at least helpful, and in some instances required” for applications such as machine translation and information retrieval where deep understanding may not be the goal. Like many firmly held beliefs, this idea is supported by widely quoted scriptural references, most notably Bar-Hillel’s (1960) famous “the box is in the pen” example, where it is taken as self evident that accurate translation of this sentence requires distinguishing among explicit senses of *pen* (‘writing utensil’ versus ‘enclosure where small children play’).

Common as this argument is, it must be viewed with suspicion. If presented without further support, it is nothing more than a variant of what Dawkins (1986) terms the “argument from personal incredulity”, which is

to say, a claim that something must be the case because one cannot imagine it being otherwise.2 The problem of word sense ambiguity has been a central concern since the earliest days of computing – Hutchins (1997) quotes a 1949 letter from Warren Weaver to Norbert Wiener on the question of machine translation, in which he refers to "the semantic difficulties because of multiple meanings." And yet, central though it still is in many people's minds, the facts remain that (a) explicit word sense disambiguation plays very little role in

many current NLP applications, and (b) NLP applications are making fine progress anyway. Those facts suggest that rather than taking the importance of WSD for granted, its role in applications is worth examining.

The path of language technology development over the last two decades presents a tremendous challenge to the traditional breakdown of NLP into problems of morphological analysis, syntactic parsing, word sense disambiguation, logical-form semantic representation, discourse analysis, and so forth. Enabling technologies designed to solve those problems have played very little role in the most visible successes of human language technology for end users, such as the commercial viability of automatic speech recognition, the ubiquity of spell checking, or the incorporation of Web text retrieval into everyday life. On the other hand, the language technology community is discovering that some forms of linguistic depth *can* make a difference, notably syntactic structure, and one could argue by analogy that this bodes well for WSD.

As one good example, Chelba and Jelinek (1998) managed to demonstrate (after decades of experience to the contrary) that the use of syntactic structure in stochastic language models can lead to reductions in perplexity and word error rate compared to standard trigram modeling, an advance with potential repercussions for applications such as speech recognition, statistical MT, and optical character recognition. Other examples include Kumar and Byrne (2002), who showed that a syntactic measure of lexical distance can be used to improve the performance of stochastic word-alignment models for machine translation; Microsoft Word, which has for some years incorporated grammar checking based on syntactic parsing; and recent success applying synchronous context-free parsing in machine translation (Chiang 2005).

Ultimately, the value of WSD in applications comes down to a question of specifics: in which applications does it help, and why? Although no application can be cited as an unequivocal success for WSD, there is certainly one widely noted failure: monolingual information retrieval. Given how starkly the results contradict the intuition that word sense disambiguation *should* make a difference, monolingual IR is the flagship example for pessimism about WSD's practical potential. WSD is more broadly a focal point for examination of all of NLP's practical utility in IR. Voorhees (1999), for instance, uses negative results involving WSD in IR to illustrate broader insights into problems of exploiting deeper NLP techniques

for IR more generally. Voorhees (1999:23) correctly observes that (monolingual) text retrieval "can be viewed as a great success story for natural language processing ... a major industry has been built around the automatic manipulation of unstructured natural language text", but, contrary to Ide and Véronis (1998), this world changing phenomenon has taken place without the use of explicit disambiguation.

IV. APPLICATIONS OF WSD

Traditional WSD in Applications

The most enduring conception of word senses in NLP comes from the lexicographic tradition, where the meanings of words are explicitly enumerated, sometimes being organized into a hierarchical structure, and they are considered to be properties of those words independent of any particular application. In a computational setting, most traditional natural language processing approaches adopt this sort of characterization of word senses as an *a priori* enumeration of discrete (albeit possibly overlapping) meanings, and what I term "traditional WSD" is the problem of selecting one of those meanings.

One might therefore extend Kilgarriff's synopsis of traditional WSD with the following sentence:

Successful natural language processing applications, therefore, must also pick the correct meaning of a word from that range of possibilities

A. Information Retrieval

The dominant paradigm in IR is based on "bag-of-words" representations: a piece of text is characterized as an unordered collection of terms, and the assessment of a document's relevance in response to a query depends primarily on the terms they have in common. Most intuitively, the terms are the words themselves. In practice, common uninformative words are excluded as terms, and multiple forms of words are mapped down to a single form via stemming – for example, *connect*, *connects*, *connecting*, and *connection* would all be stemmed as *connect*. As a result, a query about "connecting my camera" and a document containing "connection of a digital camera" would have terms *connect* and *camera* in common Three reasons for this have been widely noted. First, if queries are short, there is extremely limited context available for context-based disambiguation of query terms, which makes WSD difficult. Second, even for words with multiple senses, the most frequent sense often heavily dominates the frequency distribution of the text collection under consideration; in such

cases using the word itself is likely to be just as good as correct disambiguation. Third, most document retrieval models exhibit a tendency toward implicit disambiguation of multi-word queries, which helps bag-of words IR perform well even in the absence of explicit word senses, particularly for longer queries.

1) *Cross-language IR*

Cross-language information retrieval (CLIR) is an application developing at a rapid pace, thanks to the increasingly global nature of information seeking on the Web, global commerce, and the needs of the intelligence community. WSD may well have greater potential in CLIR than in IR, owing to the interaction of sense ambiguity with translation ambiguity. Consider, for example, a situation where word x in an English query can translate into any of $x1$, $x2$, and $x3$ in, say, Chinese, and where word y in

the English query can translate into any of $y1$, $y2$, $y3$, and $y4$. Furthermore, suppose that $x1$ and $y1$ are the correct

Chinese translations. As one would hope, Chinese documents containing both *x*1 and *y*1 will score higher than documents containing only one or the other, yielding the implicit disambiguation effect. But notice that documents containing both of *x*1 and *y*2 will *also* score higher, even though *y*2 is the wrong translation for *y* in this context, and the same holds for *x*2 and *y*3, as well as all the other combinations. In practice, for at least some of the eleven unintended *xi*, *yj* combinations,
there will be at least some documents containing both terms, and those documents will be rewarded in their scores even though they may not contain concepts corresponding to the original query terms *x* and *y*.

2) *Question Answering*

In some respects, question answering (QA) is one of the oldest NLP applications: natural language interfaces to databases and knowledge bases date back at least as far as the LUNAR system for answering questions about rock samples that were brought back by the Apollo expeditions (Woods
and Kaplan 1971). In its most recent incarnation, the aim of QA is to find answers in open-domain natural language text. Rather than querying an IR system with "*Edison light bulb patent*" – receiving, say, full articles on Edison and the light bulb in response – the goal is to ask a QA system specific questions such as "*When did Edison patent the light bulb?*" and receive back a concise answer rather than a set of relevant documents Another important consideration in QA is answering the right question. In order to find the right types of answers for a given question type (*who* versus *where* versus *when*, for example), many approaches to QA rely on some variant of named entity tagging: identifying phrases that can be categorized into high-level semantic categories such as Person, Organization, Location, Date, and so forth.

3) *Document Classification*

Most work on classifying texts into predetermined categories (text classification, categorization, or routing) is based on the same bag-of-words representations that are typical in information retrieval. Some attempts to enrich text representations with word sense information have not yielded improvements in performance (e.g., Kehagias et al. 2003, Moschitti and Basili 2004) for reasons similar to those discussed in Section 11.3.1. However, Vossen et al. (2006) present a study using the Reuters news collection in which they obtain improvements for document classification using a WSD technique that emphasizes the importance of topical domains. In similar work, Bloehdorn and Hotho (2004) show that the integration of features into the document representations for text document classification improves the result, achieving highly competitive results on the Reuters- 21578 and OHSUMED datasets (for IR and information extraction research). Some improvements can be attributed to the detection of multiword expressions and to the conflation of synonyms.
Further improvements can be achieved by generalizing concepts. In closely related work, Hotho et al. (2003) report improved results in document clustering.

B. *MACHINE TRANSLATION*

Discussions of MT conventionally make a distinction between interlingua systems and transfer systems.12 In interlingua systems, traditional WSD is necessary in order to identify the correct interlingua or "meaning" representation
for a concept expressed in the source language. Translating *The news of the attack broke at 6am*, one might select the communication sense of *broke* rather than the destruction sense (cf. *The glass in the window broke at 6am*). This monolingual analysis task produces an interlingua representation; monolingual generation for the target language then maps from that representation to a surface realization.
However, commenting on the effectiveness of explicit WSD in traditional MT systems is difficult for a number of reasons. First, sense ambiguity is only one of a large variety of forms of ambiguity that challenge MT systems, and perhaps for that reason WSD does not tend to be discussed as a separate component. Second, standardized, community-wide MT evaluations are a fairly recent phenomenon. Explicit WSD does not appear to have played a visible role in any of the systems that participated in the ARPA evaluations of the early 1990s (White and O'Connell 1994), and most participants in more recent comparative evaluations have been either statistical MT systems (discussed below) or commercial systems, for which system details are often kept confidential.

C. *SENSE AMBIGUITY IN STATISTICAL MACHINE TRANSLATION*

Explicit WSD generally does not appear to play a role in statistical MT systems. However, lexical choice in statistical models involves many of the same issues as sense disambiguation and it is informative to look at how they address the problem. the past several years, phrase-based models (Och 2002), Koehn et al. 2003) have emerged as the dominant model in statistical MT. Statistical phrase-based translation is similar in spirit to example-based machine translation (EBMT) (Nagao 1984, Brown 1996), in which a target sentence is constructed by covering the source sentence with "chunks" stored in a translation memory. The statistical version employs a probabilistic model of the mapping between "phrases" in the source and target language.17 For example, it might capture the fact that the English sequence *we have* appears frequently as the translation of French sequence *nous avons*. The translation process can be viewed as segmenting the source sentence into phrases, reordering them, and translating the phrases as units based on the mappings. As in many other statistical approaches, "decoding" is the process of searching for the optimal way to accomplish this, where "optimal" is defined by models learned from training data (Koehn et al. 2003, Koehn 2004).

D. OTHER EMERGING APPLICATIONS

A number of other emerging applications share a need to identify the semantic categories of entities. These include the extraction and mining of information in text, and the acquisition of semantic knowledge. In information extraction, the goal is to take a natural language text as input and fill in a “template” describing the basic relations that hold, for a particular, domain-specific set of template relations; in text data mining, the goal is to discover patterns of relationships that occur in large bodies of text. The bioinformatics domain provides a nice illustration. A vast molecular biology literature discusses the relationships between genes, proteins, and enzymatic functions, and enormous databases are under construction tabulating such relationships, but there is a gap between the free text data in articles and the structured data in the databases.

Weeber et al. (2001) discuss ambiguity resolution in medical NLP more generally, mentioning such applications as medical decision support, indexing, and literature-based discovery. Other problems of ambiguity include abbreviations, e.g., whether *MG* refers to *milligram* or *magnesium* (Yu et al. 2002, Yu et al. 2004) and the interpretation of acronyms, for example, whether or not *COLD* should be interpreted as *chronic obstructive pulmonary disease*. Finally, there is one more class of emerging application for which explicit WSD may have particular value: tasks where the goal is to place terms or phrases into an explicit knowledge structure. These include the development of better user interfaces – Hearst (2000) argues for task oriented search interfaces that more thoroughly integrate metadata, such as topical categories, into the user’s experience. Yee et al. (2003) and Stoica and Hearst (2004) illustrate these ideas in an interface searching a collection of fine arts images, creating categories for the collection automatically from image captions using WordNet; however, when faced with sense ambiguity they were forced to either ignore ambiguous terms or choose the first WordNet sense dating the Semantic Web, there is, of course, a long tradition of work on building ontologies in support of computational applications. In some cases WSD is an explicit part of the process. For example, Dorr and Jones (2000) employ WSD to improve the creation of large-scale semantic

lexicons; Rigau et al. (2002) describe a bootstrapping process including WSD and knowledge acquisition in a multilingual setting; and Basili et al. (2004) discuss the creation of multilingual knowledge sources in the context of ontology-based QA

V. CONCLUSIONS

In this paper, we have discussed the relationship between WSD research and NLP applications. Traditional WSD is characterized as the selection of one meaning for a word from a range of possibilities. So conceived, the role of WSD with respect to an explicit sense inventory appears to be questionable in two of the most heavily researched language technology applications, monolingual information retrieval and machine translation. However, there is evidence that traditional WSD and directly analogous techniques are useful in emerging applications such as question answering and biomedical information extraction.

In addition, there appears to be a promising argument for the utility of WSD techniques in disambiguating specialized terms such as person and place names, abbreviations, and acronyms. Finally, wediscussed several application areas worth watching, including the creation and navigation of metadata, computer-assisted lexicography, and ontology-driven frameworks such as the Semantic Web.

REFERENCES

1. Kilgarriff, A., Evans, M. (eds): Special issue on SENSEVAL. Computer and the Humanities, 34(1-2) (2000)
2. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Five Papers on WordNet. Special Issue of International Journal of Lexicography, 3(4), (1990)
3. Miller, G. A., Leacock, C., Tengi, R., Bunker, R.T.: A Semantic Concordance. Proceedings of the ARPA Workshop on Human Language Technology (1993)
4. Hirst, G.: Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press. Cambridge, England (1987)
5. McRoy, S.: Using Multiple Knowledge Sources for Word Sense Discrimination. Computational Linguistics, 18(1) (1992)
6. Ide N., Veronis J.: Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. Computational Linguistics, 24(1) (1998)
7. Bruce, R., Wilks, Y., Guthrie, L., Slator, B., Dunning, T.: NounSense - A Disambiguated Noun Taxonomy with a Sense of Humour. Research Report MCCS-92-246. Computing Research Laboratory, New Mexico State University (1992)
8. Rigau, G., Atserias, J., Agirre, E.: Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. Proceedings of ACL-EACL, Madrid, Spain. (1997)
9. Resnik, P.: Selection and Information: A Class-Based Approach to Lexical Relationships. Ph.D. University of Pennsylvania (1993)
10. Agirre, E.: Formalization of concept-relatedness using ontologies: Conceptual Density. Ph.D. thesis. University of the Basque Country (1999)
11. Yarowsky, D.: One Sense per Collocation. Proc. of the 5th DARPA Speech and Natural Language Workshop (1993)
12. Ng, H. T., Lee, H. B.: Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-based Approach. Proceedings of the ACL (1996)
13. Leacock, C., Chodorow, M., Miller, G. A.: Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics, 24(1) (1998)
14. Agirre, E., Martinez, D.: Exploring automatic word sense disambiguation with decision lists and the Web. Proceedings of the COLING Workshop on Semantic Annotation and Intelligent Content. Saarbrücken, Germany (2000)
15. Mihalcea, R., Moldovan, D.: Word Sense Disambiguation based on Semantic Density. Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems. Montreal, Canada (1998)
16. Yarowsky, D.: Word-Sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora. Proceedings of COLING. Nantes, France (1992)
17. Yarowsky, D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Proceedings of the ACL. Cambridge, USA (1995)
18. Agirre, E., Martinez, D.: Learning class-to-class selectional preferences. Proceedings of the ACL CONLL Workshop. Toulouse, France (2001)
19. Agirre, E., Ansa, O., Martinez, D., Hovy, E.: Enriching WordNet concepts with topic signatures. Proceedings of the NAACL worshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. Pittsburg, USA (2001)
20. Agirre, E., Rigau, G.: Word Sense Disambiguation using Conceptual Density. Proceedings of COLING. Copenhagen, Denmark (1996)
21. Martinez, D., Agirre, E.: Word Sense Disambiguation using syntactic cues. Internal Report. University of the Basque Country (forthcoming)
22. Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. Proceedings of the ACL (1994)