1 Introduction

Eneko Agirre[1] and Philip Edmonds[2]

[1]University of the Basque Country
[2]Sharp Laboratories of Europe Limited

1.1 Word sense disambiguation

Anyone who gets the joke when they hear a pun will realize that lexical ambiguity is a fundamental characteristic of language: Words can have more than one distinct meaning. So why is it that text doesn't seem like one long string of puns? After all, lexical ambiguity is pervasive. The 121 most frequent English nouns, which account for about one in five word occurrences in real text, have on average 7.8 meanings each (in the Princeton WordNet (Miller 1990), tabulated by Ng and Lee (1996)). But the potential for ambiguous readings tends to go completely unnoticed in normal text and flowing conversation. The effect is so strong that some people will even miss a pun (a real ambiguity) obvious to others. Words may be polysemous in principle, but in actual text there is very little real ambiguity – to a person.

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 (Ide and Véronis 1998).

In the field of computational linguistics, the problem is generally called word sense disambiguation (WSD), and is defined as the problem of computationally determining which "sense" of a word is activated by the use of the word in a particular context. WSD is essentially a task of classification:

DRAFT of
Agirre, Eneko and Edmonds, Philip. 2006. Introduction. In Agirre and Edmonds (eds.), *Word Sense Disambiguation: Algorithms and Applications*, 1–28. Springer.

word senses are the classes, the context provides the evidence, and each occurrence of a word is assigned to one or more of its possible classes based on the evidence. This is the traditional and common characterization of WSD that sees it as an explicit process of disambiguation with respect to a fixed inventory of word senses. Words are assumed to have a finite and discrete set of senses from a dictionary, a lexical knowledge base, or an ontology (in the latter, senses correspond to concepts that a word lexicalizes). Application-specific inventories can also be used. For instance, in a machine translation (MT) setting, one can treat word translations as word senses, an approach that is becoming increasingly feasible because of the availability of large multi-lingual parallel corpora that can serve as training data. The fixed inventory of traditional WSD reduces the complexity of the problem, making it tractable, but alternatives exist, as we will see below.

WSD has obvious relationships to other fields such as lexical semantics, whose main endeavour is to define, analyze, and ultimately understand the relationships between "word", "meaning", and "context". But even though word meaning is at the heart of the problem, WSD has never really found a home in lexical semantics. It could be that lexical semantics has always been more concerned with representational issues (see, for example, Lyons 1995) and models of word meaning and polysemy so far too complex for WSD (Cruse 1986; Ravin and Leacock 2000). And so, the obvious procedural or computational nature of WSD paired with its early invocation in the context of machine translation (Weaver 1949) has allied it more closely with language technology and thus computational linguistics. In fact, WSD has more in common with modern lexicography, with its intuitive premise that word uses group into coherent semantic units and its empirical corpus-based approaches, than with lexical semantics (Wilks et al. 1993).

The importance of WSD has been widely acknowledged in computational linguistics; some 700 papers in the ACL Anthology mention the term "word sense disambiguation".[1] Of course, WSD is not thought of as an end in itself, but as an enabler for other tasks and applications of computational linguistics and natural language processing (NLP) such as parsing, semantic interpretation, machine translation, information retrieval, text

[1] To compare, "anaphora resolution" occurs in 438 papers; however, such statistics should not be taken too seriously. The ACL Anthology is a digital archive of research papers in computational linguistics, covering conferences and workshops from 1979 to the present, maintained by the Association for Computational Linguistics (www.aclweb.org/anthology). Our statistics were gathered in November 2005.

mining, and (lexical) knowledge acquisition. However, in counterpoint to its theoretical importance, explicit WSD has not always demonstrated benefits in real applications.

A long-standing and central debate is whether WSD should be researched as a generic or as an integrated component. In the generic setting, the WSD component is a black box encompassing an explicit process of WSD that can be dropped into any application, much like a part-of-speech tagger or a syntactic parser. The alternative is to include WSD as a task-specific "component" of a particular application in a specific domain and integrated so completely into a system that it is difficult to separate out. Research into explicit WSD, having received the bulk of effort, has progressed steadily and successfully to a point where some people now question if the upper limit in accuracy (low as it is on fine-grained sense distinctions) has been attained (Section 1.6 gives current performance levels). And yet, explicit WSD has not yet been convincingly demonstrated to have a significant positive effect on any application. Only the integrated approach has been successful, with disambiguation often occurring implicitly by virtue of other operations, for example, in the language and translation models of statistical machine translation. The former conception is easier to define, experiment with, and evaluate, and is thus more amenable to the scientific method; the latter is more applicable and puts the need for explicit WSD into question.

Despite uncertain results on real applications, the effort on explicit WSD has produced a solid legacy of research results, methodology, and insights for computational semantics. For example, local contextual features (i.e., other words near the target word) provide better evidence in general than wider topical features (Yarowsky 2000). Indeed, the role of context in WSD is much better understood: Compared to other classification tasks in NLP (such as part-of-speech tagging), WSD requires a wide range of contextual knowledge to be modeled from fixed patterns of part-of-speech tags around a topic word to syntactic relations to topical and domain associations. Each part-of-speech and even each word relies on different types of knowledge for disambiguation. For instance, nouns benefit from a wide context and local collocations, whereas verbs benefit from syntactic features. Some words can be disambiguated by a single feature in the right position, benefiting from a "discriminative" method; others require an aggregation of many features. Homographs are generally much

easier to disambiguate than polysemous words.[2] An evaluation methodology has been defined by Senseval (Kilgarriff and Palmer 2000) and many resources in several languages are now available. Finally, for a small sample of tested words, that have sufficient training data, the performance of WSD systems is comparable to that of humans (measured as the inter-tagger agreement among two or more humans), as demonstrated by the recent Senseval results (see Sect. 1.6 below).

Two "spin offs" worth mentioning include the development of explicit WSD as a benchmark application for machine learning research, because of the clear problem definition and methodology, the variety of problem spaces (each word is a separate classification task), the high-dimensional feature space, and the skewed nature of word sense distributions. And second, WSD research is helping in the development of popular lexical resources such as WordNet (Fellbaum 1998; Palmer et al. 2001, 2006) and the multilingual lexicons of the MEANING project (Vossen et al. 2006).

To introduce the topic of WSD, we begin with a brief history. Then, in Section 1.3 we discuss the central theoretical issues of "word sense" and the sense inventory. In Sections 1.4–1.6 we summarize several practical aspects including applicability to NLP tasks, the three basic approaches to WSD, and current performance achievements. Finally, Section 1.7 gathers our thoughts on emerging and future research into WSD.

1.2 A brief history of WSD research

In order to introduce current WSD research, reported in the book, we provide here a brief review of the history of WSD research.[3]

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:

> If one examines the words in a book, one at a time through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine,

[2] For the present purposes, a homograph is a coarse-grained sense distinction between often completely unrelated meanings of the same word string (e.g., *bank* as a financial institution or a river side). Polysemy involves a finer-grained sense distinction in which the senses can be related in different ways (e.g., *bank* as a physical building or as an institution). See Section 1.3 for further details.

[3] See Ide and Véronis (1998) for a more extensive history (up to 1998, of course.)

> one at a time, the meaning of words. "Fast" may mean "rapid"; or it may mean "motionless"; and there is no way of telling which.
>
> But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning …

In addition to formulating the general methodology still applied today (see also Kaplan (1950) and Reifler (1955)), 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.

Some early work set the stage for methods still pursued today. Masterman (1957), for instance, used the headings of the categories in *Roget's International Thesaurus* (Chapman 1977) to represent the different senses of a word, and then chose the heading whose contained words were most prominent in the context. Madhu and Lytle (1965) calculated sense frequencies of words in different domains – observing early on that domain constrains sense – and then applied Bayes formula to choose the most probable sense given a context.

Early researchers well understood the significance and difficulty of WSD. In fact, this difficulty was one of the reasons why most of MT was abandoned in the 1960s due to the unfavorable ALPAC report (1966). For example, Bar-Hillel (1960) argued that "no existing or imaginable program will enable an electronic computer to determine that the word *pen*" is used in its 'enclosure' sense in the passage below, because of the need to model, in general, all world knowledge like, for example, the relative sizes of objects:

[4] Zipf's "Law of Meaning" is different from his well known "Zipf's Law" about the power-law distribution of word frequencies.

Little John was looking for his toy box. Finally he found it. The box was in the *pen*. John was very happy.

Ironically, the very "statistical semantics" that Weaver proposed might have applied in cases such as this: Yarowsky (2000) notes that the trigram *in the pen* is very strongly indicative of the enclosure sense, since one almost never refers to what is in a writing pen, except for ink.

WSD was resurrected in the 1970s within artificial intelligence (AI) research on full natural language understanding. In this spirit, Wilks (1975) developed "preference semantics", one of the first systems to explicitly account for WSD. The system used selectional restrictions and a frame-based lexical semantics to find a consistent set of word senses for the words in a sentence. The idea of individual "word experts" evolved over this time (Rieger and Small 1979). For example, in Hirst's (1987) system, a word was gradually disambiguated as information was passed between the various modules (including a lexicon, parser, and semantic interpreter) in a process he called "Polaroid Words". "Proper" knowledge representation was important in the AI paradigm. Knowledge sources had to be handcrafted, so the ensuing knowledge acquisition bottleneck inevitably led to limited lexical coverage of narrow domains and would not scale.

The 1980s were a turning point for WSD. Large-scale lexical resources and corpora became available so handcrafting could be replaced with knowledge extracted automatically from the resources (Wilks et al. 1990). Lesk's (1986) short but extremely seminal paper used the overlap of word sense definitions in the *Oxford Advanced Learner's Dictionary of Current English* (OALD) to resolve word senses. Given two (or more) target words in a sentence, the pair of senses whose definitions have the greatest lexical overlap are chosen (see Chap. 5 (Sect. 5.2)). Dictionary-based WSD had begun and the relationship of WSD to lexicography became explicit. For example, Guthrie et al. (1991) used the subject codes (e.g., Economics, Engineering, etc.) in the *Longman Dictionary of Contemporary English* (LDOCE) (Procter 1978) on top of Lesk's method. Yarowsky (1992) combined the information in *Roget's International Thesaurus* with co-occurrence data from large corpora in order to learn disambiguation rules for Roget's classes, which could then be applied to words in a manner reminiscent of Masterman (1957) (see Chap. 10 (Sect. 10.2.1)). Although dictionary methods are useful for some cases of word sense ambiguity (such as homographs), they are not robust since dictionaries lack complete coverage of information on sense distinctions.

The 1990s saw three major developments: WordNet became available, the statistical revolution in NLP swept through, and Senseval began.

WordNet (Miller 1990) pushed research forward because it was both computationally accessible and hierarchically organized into word senses called synsets. Today, English WordNet (together with wordnets for other languages) is the most-used general sense inventory in WSD research.

Statistical and machine learning methods have been successfully applied to the sense classification problem. Today, methods that train on manually sense-tagged corpora (i.e., supervised learning methods) have become the mainstream approach to WSD, with the best results in all tasks of the Senseval competitions. Weaver had recognized the statistical nature of the problem as early as 1949 and early corpus-based work by Weiss (1973), Kelley and Stone (1975), and Black (1988) presaged the statistical revolution by demonstrating the potential of empirical methods to extract disambiguation clues from manually-tagged corpora. Brown et al. (1991) were the first to use corpus-based WSD in statistical MT.

Before Senseval, it was extremely difficult to compare and evaluate different systems because of disparities in test words, annotators, sense inventories, and corpora. For instance, Gale et al. (1992:252) noted that "the literature on word sense disambiguation fails to offer a clear model that we might follow in order to quantify the performance of our disambiguation algorithms," and so they introduced lower bounds (choosing the most frequent sense) and upper bounds (the performance of human annotators). However, these could not be used effectively until sufficiently large test corpora were generated. Senseval was first discussed in 1997 (Resnik and Yarowsky 1999; Kilgarriff and Palmer 2000) and now after hosting three evaluation exercises has grown into the primary forum for researchers to discuss and advance the field. Its main contribution was to establish a framework for WSD evaluation that includes standardized task descriptions and an evaluation methodology. It has also focused research, enabled scientific rigor, produced benchmarks, and generated substantial resources in many languages (e.g., sense-annotated corpora), thus enabling research in languages other than English.

Recently, at the Senseval-3 workshop (Mihalcea and Edmonds 2004) there was a general consensus (and a sense of unease) that the traditional explicit WSD task, so effective at driving research, had reached a plateau and was not likely to lead to fundamentally new research. This could indicate the need to look for new research directions in the field, some of which may already be emerging, for instance the use of parallel bilingual corpora. Section 1.7 explores the emerging research, but let's first review the issue at the center of it all: word senses.

1.3 What is a word sense?

Word meaning is in principle infinitely variable and context sensitive. It does not divide up easily into distinct sub-meanings or senses. Lexicographers frequently discover in corpus data loose and overlapping word meanings, and standard or conventional meanings extended, modulated, and exploited in a bewildering variety of ways (Kilgarriff 1997; Hanks 2000; also Chap. 2). In lexical semantics, this phenomenon is often addressed in theories that model sense extension and semantic vagueness, but such theories are at a very early stage in explaining the complexities of word meaning (e.g., Cruse 1986; Tuggy 1993; Lyons 1995).

"Polysemy" means to have multiple meanings. It is an intrinsic property of words (in isolation from text), whereas "ambiguity" is a property of text. Whenever there is uncertainty as to the meaning that a speaker or writer intends, there is ambiguity. So, polysemy indicates only potential ambiguity, and context works to remove ambiguity.

At a coarse grain a word often has a small number of senses that are clearly different and probably completely unrelated to each other, usually called *homographs*. Such senses are just "accidentally" collected under the same word string. As one moves to finer-grained distinctions the coarse-grained senses break up into a complex structure of interrelated senses, involving phenomena such as general polysemy, regular polysemy, and metaphorical extension. Thus, most sense distinctions are not as clear as the distinction between *bank* as 'financial institution' and *bank* as 'river side'. For example, *bank* as financial institution splits into the following cloud of related senses: the company or institution, the building itself, the counter where money is exchanged, a fund or reserve of money, a money box (*piggy bank*), the funds in a gambling house, the dealer in a gambling house, and a supply of something held in reserve (*blood bank*) (WordNet 2.1).

Even rare and seemingly innocuous words such as *quoin* offer a rich structure of meanings. *The American Heritage Dictionary of the English Language* lists three related noun-senses: the outer angle or corner of a wall, a brick forming such an angle (a cornerstone), and a wedge-shaped block. As a verb, it can mean to build a corner with distinctive blocks, or, in the printing domain, to secure metal type with a quoin.

Given the range of sense distinctions in examples such as these, which represent the norm, one might start to wonder if the very idea of word-sense is suspect. Some argue that task-independent senses simply cannot be enumerated in a list (Kilgarriff 1997; others that words are monose-

mous, having a have only a single, abstract meaning (Ruhl 1989). And perhaps the only tenable position is that a word must have a different meaning in each distinct context in which it occurs. But a strong word-in-context position ignores the intuition that word usages seem to cluster together into coherent sets, which could be called senses, even if the sets cannot be satisfactorily described or labeled. The work on sense discovery or induction gives some empirical evidence for this intuition, however such "senses" are more aptly called "word uses" (see Chap. 6 (Sect. 6.3)).

Concerns about the theoretical, linguistic, or psychological reality of word senses notwithstanding, the field of WSD has successfully established itself by largely ignoring them, much as lexicographers do in order to produce dictionaries. Except, Kilgarriff (Chap. 2) suggests that it *is* time to take notice.

In practice, the need for a sense inventory has driven WSD research. In the common conception, a sense inventory is an exhaustive and fixed list of the senses of every word of concern in an application. The nature of the sense inventory depends on the application, and the nature of the disambiguation task depends on the inventory. The three Cs of sense inventories are: clarity, consistency, and complete coverage of the range of meaning distinctions that matter. Sense granularity is actually a key consideration: too coarse and some critical senses may be missed, too fine and unnecessary errors may occur. For example, the ambiguity of *mouse* (animal or device) is not relevant in English-Basque machine translation, where *sagu* is the only translation, but is relevant in (English and Basque) information retrieval. The opposite is true of *sister*, which is translated differently into Basque depending on the gender of the other sibling: *ahizpa* for 'sister of a girl' and *arreba* for 'sister of a boy'. In fact, Ide and Wilks (Chap. 3) argue that coarse-level distinctions are the only ones that humans and machines can reliably discriminate (and that they are *the* distinctions of concern to applications). There is evidence (see Chap. 4) that if senses are too fine or unclear, human annotators also have difficulty assigning them.

The "sense inventory" has been the most contentious issue in the WSD community, and it surfaced during the formation of Senseval, which required agreement on a common standard. The main inventories used in English research have included LDOCE, *Roget's International Thesaurus*, Hector, and WordNet. For other languages a variety of dictionaries have been used, together with local WordNet versions. Each resource has its pros and cons, which will become clear throughout the book (especially Chaps. 2, 3, and 4). For example, Hector (Atkins 1991) is lexicographically sound and detailed, but lacks coverage; LDOCE has subject codes

and a structure such that homographs are part-of-speech-homogeneous, but is not freely available; WordNet is an open and very popular resource, but is too fine-grained in many cases. Senseval eventually settled on WordNet, mainly because of its availability and coverage. Of course, this choice sidesteps the greater debate of explicit versus implicit WSD, which brings the challenge that entirely different kinds of inventory would be required for applications such as MT (translation equivalences) and IR (induced clusters of usages).

1.4 Applications of WSD

Machine translation is the original and most obvious application for WSD but disambiguation has been considered in almost every NLP application, and is becoming increasingly important in recent areas such as bioinformatics and the Semantic Web..

Machine translation (MT). WSD is required for lexical choice in MT for words that have different translations for different senses and that are potentially ambiguous within a given domain (since non-domain senses could be removed during lexicon development). For example, in an English-French financial news translator, the English noun *change* could translate to either *changement* ('transformation') or *monnaie* ('pocket money'). In MT, the senses are often represented directly as words in the target language. However, most MT models do not use explicit WSD. Either the lexicon is pre-disambiguated for a given domain, hand-crafted rules are devised, or WSD is folded into a statistical translation model (Brown et al. 1991).

Information retrieval (IR). Ambiguity has to be resolved in some queries. For instance, given the query "*depression*" should the system return documents about illness, weather systems, or economics? A similar problem arises for proper nouns such as *Raleigh* (bicycle, person, city, etc.). Current IR systems do not use explicit WSD, and rely on the user typing enough context in the query to only retrieve documents relevant to the intended sense (e.g., "*tropical depression*"). Early experiments suggested that reliable IR would require at least 90% disambiguation accuracy for explicit WSD to be of benefit (Sanderson 1994). More recently, WSD has been shown to improve cross-lingual IR and document classification (Vossen et al. 2006; Bloehdorn and Hotho 2004; Clough and Stevenson 2004). Besides document classification and cross-lingual IR, related appli-

cations include news recommendation and alerting, topic tracking, and automatic advertisement placement.

Information extraction (IE) and text mining. WSD is required for the accurate analysis of text in many applications. For instance, an intelligence gathering system might require the flagging of, say, all the references to illegal *drugs*, rather than medical *drugs*. Bioinformatics research requires the relationships between genes and gene products to be catalogued from the vast scientific literature; however, genes and their proteins often have the same name. More generally, the Semantic Web requires automatic annotation of documents according to a reference ontology: all textual references must be resolved to the right concepts and event structures in the ontology (Dill et al. 2003). Named-entity classification, co-reference determination, and acronym expansion (*MG* as *magnesium* or *milligram*) can also be cast as WSD problems for proper names. WSD is only beginning to be applied in these areas.

Lexicography. Modern lexicography is corpus-based, thus WSD and lexicography can work in a loop, with WSD providing rough empirical sense groupings and statistically significant contextual indicators of sense to lexicographers, who provide better sense inventories and sense-annotated corpora to WSD. Furthermore, intelligent dictionaries and thesauri might one day provide us with a semantically-cross-referenced dictionary as well as better contextual look-up facilities.

Despite this range of applications where WSD shows a great potential to be useful, WSD has not yet been shown to make a decisive difference in any application. There are various isolated results that show minor improvements, but just as often WSD can hurt performance, as is the case in one experiment on information retrieval (Sanderson 1994). There are several possible reasons for this. First, the domain of an application often constrains the number of senses a word can have (e.g., one would not expect to see the 'river side' sense of *bank* in a financial application), and so lexicons can be constructed accordingly. Second, WSD might not be accurate enough yet to show an effect. Third, treating WSD as an explicit component, as the majority of research does, means that it cannot be properly integrated into a particular application or appropriately trained on the domain. Most applications, such as MT, do not have a place for a WSD module (but see Carpuat and Wu (2005)), so either the application or the WSD would have to be redesigned. Research is just beginning on domain-specific WSD (see Chap. 10).

Nevertheless, it's clear that applications do require WSD in some form – perhaps through an *implicit* encoding of the same contextual models used in explicit WSD. For example in IR, a two-word query can disambiguate itself, implicitly, since both words are often used in text together in the senses intended by the user (e.g., *tropical depression*, above), and we've already mentioned the modeling of WSD in MT. The work on explicit WSD can serve to explore and highlight the particular features that provide the best evidence for accurate disambiguation, implicit or explicit.

1.5 Basic approaches to WSD

Approaches to WSD are often classified according to the main source of knowledge used in sense differentiation. Methods that rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence, are termed *dictionary-based* or *knowledge-based*. Methods that eschew (almost) completely external information and work directly from raw unannotated corpora are termed *unsupervised* methods (adopting terminology from machine learning). Included in this category are methods that use word-aligned corpora to gather cross-linguistic evidence for sense discrimination. Finally, *supervised* and *semi-supervised* WSD make use of annotated corpora to train from, or as seed data in a bootstrapping process.

Almost every approach to supervised learning has now been applied to WSD, including aggregative and discriminative algorithms and associated techniques such as feature selection, parameter optimization, and ensemble learning (see Chap. 7).

Unsupervised learning methods have the potential to overcome the new knowledge acquisition bottleneck (manual sense-tagging) and have achieved good results (Schütze 1998). These methods are able to induce word senses from training text by clustering word occurrences, and then classifying new occurrences into the induced clusters/senses (see Chap. 6).

The knowledge-based proposals of the 1970s and 80s are still a matter of current research. The main techniques use selectional restrictions, the overlap of definition text, and semantic similarity measures (see Chap. 5). Ultimately, the goal is to do general semantic inference using knowledge bases, with WSD as a by-product.

Table 1.1 is our attempt to be systematic in covering the main approaches to WSD in this book, but it was not always easy. For instance, Chapters 9 and 10 cover some techniques that did not fit very well in other chapters. Indeed, drawing a line between current systems is difficult, not

Table 1.1. A variety of approaches to word sense disambiguation are discussed in this book

Approach	Technique	Chapter
Knowledge-based	Hand-crafted disambiguation rules	Not covered
	Selectional restrictions (or preferences), used to filter out inconsistent senses	5
	Comparing dictionary definitions to the context (Lesk's method)	5
	The sense most similar to its context, using semantic similarity measures	5
	"One-sense-per-discourse" and other heuristics	5
Unsupervised corpus-based	Unsupervised methods that cluster word occurrences or contexts, thus inducing senses	6
	Using an aligned parallel corpus to infer cross-language sense distinctions	6, 9, 11
Supervised corpus-based	Supervised machine learning, trained on a manually-tagged corpus	7
	Bootstrapping from seed data (semi-supervised)	7
Combinations	Unsupervised clustering techniques combined with knowledge base similarities	6
	Using knowledge bases to search for examples for training in supervised WSD	9
	Using an aligned parallel corpus, combined with knowledge-based methods	9
	Using domain knowledge and subject codes	10

least because recent research is exploring novel combinations of already existing techniques. For instance, cross-linguistic evidence gathered from word-aligned corpora can be used to train supervised systems, and then be combined with knowledge bases; unsupervised clustering techniques can be combined with knowledge-base similarities to produce sense preferences; and the information in knowledge-bases can be used to search for training examples which are then fed into supervised WSD.

Regardless of the approach, all WSD systems extract contextual features of a target word (in text) and compare them against the sense differentiation information stored for that word. A natural classification problem, WSD is characterized by its very high-dimensional feature space. Almost every type of local and topical feature has been shown to be useful including part-of-speech, word (as written and lemma), collocation, semantic class, subject or domain code, and syntactic dependency (see Chap. 8).

1.6 State-of-the-art performance

We will briefly summarize the performance achieved by state-of-the-art WSD systems. First, homographs are often considered to be a solved problem. Accuracy above 95% is routinely achieved using very little input knowledge: for example, Yarowsky (1995) used a semi-supervised approach evaluated on 12 words (96.5%), and Stevenson and Wilks (2001) used part-of-speech data (and other knowledge sources) on all words using LDOCE (94.7%).

Accurate WSD on general polysemy has been more difficult to achieve, but has improved over time. In 1997, Senseval-1 (Kilgarriff and Palmer 2000) found accuracy of 77% on the English lexical sample task,[5] just below the 80% level of human performance (estimated by inter-tagger agreement; however, human replicability was estimated at 95%; see Chap. 4). In 2001, scores at Senseval-2 (Edmonds and Cotton 2001) appeared to be lower, but the task was more difficult, as it was based on the finer-grained senses of WordNet. The best accuracy on the English lexical sample task at Senseval-2 was 64% (to an inter-tagger agreement of 86%). Table 1.2 gives the results for all evaluated languages. Previous to Senseval-2, there was debate over whether a knowledge-based or machine learning approach was better, but Senseval-2 showed that supervised approaches had the best overall performance. However, the best unsupervised system on the English lexical sample task performed at 40%, well below the most-frequent-sense baseline of 48%, but better than the random baseline of 16%.

By 2004, the top systems on the English lexical sample task at Senseval-3 (Mihalcea and Edmonds 2004) were performing at human levels according to inter-tagger agreement (see Table 1.3). The ten top systems, all supervised, made between 71.8% and 72.9% correct disambiguations compared to an inter-tagger agreement of 67%.[6] The best unsupervised system overcame the most-frequent-sense baseline achieving 66% accuracy. The

[5] A "lexical sample" task involves tagging a few occurrences of a sample of words for which hand-annotated training data is provided. An "all-words" task involves tagging all words occurring in running text. See Chapter 4.

[6] This low agreement is perhaps explained because the annotators in this case were non-experts at the task – they were merely self-selected participants in the Open Mind Word Expert project (Chlovski & Mihalcea 2002) – rather than linguistically trained lexicographers and students as employed previously. Systems can beat human ITA because adjudication for the gold standard occurs after inter-

Table 1.2. Performance of WSD systems in the Senseval-2 evaluation (Edmonds and Kilgarriff 2002)

Language	Task[a]	Systems	Lemmas	Instances	ITA[b]	Baseline[d]	Best score
English	AW	21	1,082	2,473	75%	57%/–[e]	69%/55%
Estonian	AW	2	4,608	11,504	72	85	67
Basque	LS	3	40	5,284	75	65	76
English	LS	26	73	12,939	86[c]	48/16	64/40
Italian	LS	2	83	3,900	21	–	39
Japanese	LS	7	100	10,000	86	72	78
Korean	LS	2	11	1,733	–	71	74
Spanish	LS	12	39	6,705	64	48	65
Swedish	LS	8	40	10,241	95	–	70
Japanese	TM	9	40	1,200	81	37	79

[a] *AW* all-words, *LS* lexical sample, *TM* translation memory.

[b] ITA is inter-tagger agreement, which is deemed as upper bound for the task.

[c] The ITA for English nouns and adjectives is reported. Verbs had an ITA of 71%.

[d] The baseline is most-frequent sense.

[e] Scores separated by a slash are supervised/unsupervised methods; supervised when there is no slash.

score on the all-words task was lower than for Senseval-2, probably because of a more difficult text. Senseval-3 also brought the complete domination of supervised approaches over pure knowledge-based approaches.

1.7 Promising directions

Martin Kay, in his acceptance speech for the 2005 ACL Lifetime Achievement Award, made a distinction between "computational linguistics" (CL), the use of computers to investigate and further linguistic theory, and "natural language processing" (NLP), engineering technologies for speech and text processing. Although much of the recent work in computational WSD falls squarely in the latter, solving the WSD problem is actually a prototypical endeavor for the former.

tagger agreement is calculated (see Chap. 4). This means that the systems could be performing more like linguistically trained individuals, having learned from the adjudicated corpus. Notice that other languages had higher agreements.

Table 1.3. Performance of WSD systems in the Senseval-3 evaluation (Mihalcea and Edmonds 2004)

Language	Task[a]	Systems	Lemmas	Instances	ITA[b]	Baseline[c]	Best score
English	AW	26	–	2,081	62%	62%/–[d]	65%/58%
Basque	LS	8	40	7,362	78	59	70
Catalan	LS	7	27	6,721	93	66	85
English	LS	47	57	–	67	55/–	73/66
Italian	LS	6	45	7,584	89	18	53
Romanian	LS	7	39	11,532	–	58	73
Spanish	LS	9	46	12,625	83–90	67	84
Hindi	TM	8	41	11,984	–	56	67
English	GL	10	–	42,491	–	–	68

[a]*AW* all-words, *LS* lexical sample, *TM* translation memory, *GL* gloss task.

[b]ITA is inter-tagger agreement.

[c]The baseline is most-frequent sense.

[d]Scores separated by a slash are supervised/unsupervised methods; supervised when there is no slash.

Thus, the field finds itself in a strange position. The problem of resolving lexical ambiguity itself is one of the oldest problems in CL/NLP and MT research, acknowledged as both difficult and necessary. So difficult that it was partially responsible for the cessation of funding to MT research in the 1960s following the ALPAC report. Nevertheless, researchers have made great strides in solving one constrained version of the problem: the traditional conception as an explicit task of resolving fine-grained and coarse-grained ambiguity to a fixed inventory of senses. The three evaluation exercises run by Senseval show that over a variety of word types, word frequencies, and sense distributions, explicit WSD systems are achieving consistent and respectable accuracy levels. And yet, this success has not translated into better performance or utility in real applications. Ironically, research into WSD has become separate from research into NLP applications, despite several efforts to investigate and demonstrate utility.

As we mentioned in Section 1.2, there is a growing feeling in the community that change is necessary. The route taken to reach the state-of-the-art systems – explicit WSD solved by supervised learning approaches – may not lead to future performance increases or to fundamentally new research results.

We believe that there are two complementary routes forward. The first is to become more theoretical, to return to computational linguistics, to work on WSD embracing more realistic models of word sense (including non-discreteness, vagueness, and analogy), thus drawing on and feeding theories of word meaning and context from (computational) lexical semantics and lexicography. While not obviously immediately applicable, this research has defensible goals. Can we look to WSD research to provide a practical computational lexical semantics?

The second route is to focus on making WSD applicable whatever it takes. Can any of the results to date be applied in real applications? Why doesn't explicit WSD work in applications when other generic NLP components do? Does WSD have to be more accurate? Are homographs the best level of granularity? Is domain-based WSD the answer?

Both routes could lead to better applications and a better understanding of meaning and language – surely the two main goals of NLP and computational linguistics.

It is worth revisiting the three main open problems of 1998, as put forth by Ide and Véronis (1998), and to add a few more.

The role of context. Ide and Véronis said the "relative role and importance of information from the different contexts and their inter-relations are not well understood." (p. 18) Although there is still more work to be done in isolating the contribution of different knowledge sources, much is now understood about the role of context, such as the diversity of feature types that can be used as evidence, and the types of features most useful for a few classes of words (see Chap. 8). Perhaps a goal of future WSD research should be to understand how contextual information comes to bear on semantic processing in different applications such as MT and IR and to choose the approach and knowledge sources that best fit the applications.

Sense division. How to divide senses still remains one of the main open problems of WSD. As discussed in this chapter and throughout the book (see especially Chaps. 2, 3, and 4), semantic granularity is not well understood, and the relation to specific applications is unexplored territory. Given the state of the art, coarse-grained differences could allow for performance closer to an application's needs.

Evaluation. The first Senseval was held at about the time Ide and Véronis (1998) was published. As mentioned above, Senseval's common evaluation framework has focused research, enabled scientific rigor, and generated substantial resources. But, to date, it has worked with only *in vitro* evaluation of generic WSD, separating the task from application. *In vivo*

evaluation, or application-specific evaluation, has not yet been approached, but it is precisely this kind of evaluation that could prove the utility of WSD. (See Chapter 4.)

Additional open problems include (following a survey of this book's contributors):

Domain- and application-based WSD. We discussed the need for application-specific research above as one major route forward for the field, but this will entail a change in the conception of the task. Knowing the domain of a text can often disambiguate its words, but this assumes a specialized domain lexicon or a general lexicon expanded and tuned with domain-specific information. All-words WSD would be required and *in vivo* evaluation would support the effort. (See Chapters 10 and 11.)

Unsupervised WSD and cross-lingual approaches. Tagging with no, or very little, hand-annotated training data still holds the promise of great riches. Recent work by McCarthy et al. (2004) on tagging with the predominant sense has reinvigorated this direction, and techniques that exploit alignments in parallel or comparable corpora are gaining momentum (Diab 2003; Ng et al. 2003; Bhattacharya et al. 2004; Li and Li 2004; Tufiş et al. 2004). The knowledge acquisition bottleneck is a serious impediment to supervised all-words WSD, but this could be alleviated by advances in robust methods for acquiring large sets of training examples (for all languages) with a minimum of human annotation effort. (See Chapters 6, 9, and 11.)

WSD as an optimization problem. Current WSD systems disambiguate texts one word at a time, treating each word in isolation. It is clear though that meanings are interdependent and the disambiguation of a word can affect others in its context. This was clear in earlier systems (e.g., Lesk (1986) and Cowie et al. (1992)). The interdependencies among senses in the context could be modeled and treated as an optimization problem (in contrast to the classification model of WSD).

Applying deeper linguistic knowledge. Significant advances in the performance of current supervised WSD systems could rely on enriched feature representations based on deeper linguistic knowledge, rather than better learning algorithms. We refer, for instance, to sub-categorization frames, syntactic structure, selectional preferences, semantic roles, domain information, and other semantics, which are becoming available in wide-coverage lexical knowledge bases like WordNet, VerbNet (Kipper et al. 2000), and FrameNet (Baker et al. 2003). The recent trend to rediscover semantic interpretation and entailment includes WSD and semantic role

labeling as component technologies (Gildea and Jurafsky 2002; Dagan et al. 2005). Coupling these techniques with the currently available resources, we are seeing a shift back to knowledge-based methods, but this time coupled with corpus-based methods.

Sense discovery. A sense inventory that a priori lists all relevant senses will never be able to cope with borrowed words, new words, new usages, or just rare or spurious usages. In practical terms, this makes it very difficult to move a system into a new domain. Sense discovery was a major component of Schütze's (1998) work (see Chap. 6 (Sect. 6.3)), but little work has been done since, except Véronis (2004). Even identifying which words are being used in a novel (previously unknown) way, either with a completely new meaning or an existing meaning, would be useful in many applications. Senses can also be mined from parallel corpora and the Web (see Chap. 9).

1.8 Overview of this book

This is the first book that covers the entire topic of word sense disambiguation (WSD) including: all the major algorithms, techniques, performance measures, philosophical issues, applications, and future trends. Leading researchers in the field have contributed chapters that synthesize and overview past and state-of-the-art research in their respective areas of expertise. For researchers, lecturers, students, and developers, we intend the book to answer (or begin answering) questions such as How well does WSD work? What are the main approaches and algorithms? Which technique is best for my application? How do I build it and evaluate it? What performance can I expect? What are the open problems? What is the nature of the relationship between WSD and other language processing components? What *is* a word sense? Is WSD a good topic for my PhD? Where is the field heading?

We hope that the chapters you have in your hands are helpful in this direction.

Chapter 2. Word senses. Adam Kilgarriff explores various conceptions of "word sense", including views from lexicographers to philosophers. He argues that any attempt to pin down an inventory of word senses for WSD will be problematic by considering limiting cases of metaphor, quotation, and reasoning from general knowledge.

Chapter 3. Making sense about sense. Nancy Ide and Yorick Wilks suggest that the standard fine-grained division of senses by a lexicographer for use by a human reader may not be an appropriate goal for the computational WSD task. Giving an overview of the literature on the psycholinguistic basis of sense in the mental lexicon, they argue that the level of sense-discrimination that NLP needs corresponds roughly to homographs, which are often lexicalized cross-linguistically. Thus, they propose to reorient WSD to what it can actually perform at high accuracy.

Chapter 4. Evaluation of WSD systems. Martha Palmer, Hwee Tou Ng, and Hoa Trang Dang discuss the methodology for the evaluation of WSD systems, developed through Senseval. They give an overview of previous evaluation exercises and investigate sources of human inter-tagger disagreements. Many errors are at least partially reconciled by a more coarse-grained partition of the senses. Well-defined sense groups can be of value in improving sense tagging consistency for both humans and machines.

Chapter 5. Knowledge-based methods for WSD. Rada Mihalcea reviews current research on knowledge-intensive methods, including those using overlap of dictionary definitions, similarity measures over semantic networks, selectional preferences for arguments, and several heuristics, such as “one-sense-per-discourse”.

Chapter 6. Unsupervised corpus-based methods for WSD. Ted Pedersen focuses on knowledge-lean methods that do not rely on external sources of evidence other than the untagged corpus itself. These methods do not assign sense tags to words, but rather discriminate between word uses or induce word-use clusters. The chapter reviews both distributional approaches relying on monolingual corpora and methods based on translational equivalences as found in word-aligned parallel corpora.

Chapter 7. Supervised corpus-based methods for WSD. Lluís Màrquez, Gerard Escudero, David Martínez, and German Rigau present methods that automatically induce classification models or rules from manually annotated examples, currently the mainstream approach. This chapter presents a detailed review of the literature, descriptions of five of the key machine learning algorithms including Naïve Bayes and Support Vector Machines, and a discussion of central issues such as learning paradigms, corpora used, sense repositories, and feature representation.

Chapter 8. Knowledge sources for WSD. Eneko Agirre and Mark Stevenson explore the different sources of linguistic knowledge that can be used by WSD systems. An analysis of actual WSD systems reveals that the

best results are often obtained by combining knowledge sources and the chapter concludes by analyzing experiments on the effect of different knowledge sources.

Chapter 9. Automatic acquisition of lexical information and examples. Julio Gonzalo and Felisa Verdejo consider the knowledge acquisition bottleneck faced by supervised corpus-based methods. The chapter reviews current research to remedy the lack of sufficient hand-tagged examples, by using, for example, techniques that mine large corpora for examples of word senses or coupling parallel corpora with knowledge-based methods.

Chapter 10. Domain-specific WSD. Paul Buitelaar, Bernardo Magnini, Carlo Strapparava, and Piek Vossen describe approaches to WSD that take the subject, domain, or topic of words into account. They discuss the use of subject codes, the extraction of topic signatures through a combined use of a semantic resource and domain-specific corpora, and domain-specific tuning of semantic resources.

Chapter 11. WSD in NLP applications. Philip Resnik considers applications of WSD in language technology, looking at established and emerging applications and at more and less traditional conceptions of the task.

1.9 Further reading

Visit the book website, www.wsdbook.org, for the latest information and updates.

Ide and Véronis's (1998) survey of WSD is an excellent starting point for a thorough analysis and history of WSD. It forms the introduction to the special issue of *Computational Linguistics 24(1)* on WSD. A special issue of *Computer, Speech, and Language 18(4)* (edited by Preiss and Stevenson, 2004) contains more recent contributions.

The article "Disambiguation, lexical" in the *Elsevier Encyclopedia of Language and Linguistics, 2nd ed.* (Edmonds 2005) gives an accessible overview of WSD.

Recent technical surveys are to be found in *Foundations of Statistical Natural Language Processing* (Manning and Schütze 1999), *Speech and Language Processing* (Jurafsky and Martin 2000), and the *Handbook of Natural Language Processing* (Dale et al. 2000). The first introduces WSD in the statistical framework (including the three main approaches) with detailed algorithms of a few selected systems. The second frames the problem in the context of semantic representation and analysis, and includes a

discussion of selectional preferences as well as a brief overview of the machine learning focus. The third article, by David Yarowsky, gives a good overview of the characteristics of the WSD problem, and then focuses primarily on machine learning and related solutions. An older survey in Allen's (1995) *Natural Language Understanding* treats WSD as a component in semantic interpretation. Finally, several chapters in *Electric Words* (Wilks et al. 1996) take a lexicographic perspective on WSD and discuss how LDOCE can be used.

A few books focus squarely on WSD. Lexical Ambiguity Resolution (Small et al. 1988) is a collection of papers from a cognitive science perspective. Hirst's (1987) *Semantic Interpretation and the Resolution of Ambiguity* discusses his semantic interpretation system and "Polaroid Words". And Stevenson's (2003) *Word Sense Disambiguation* is based on his PhD dissertation on the benefits of combining knowledge sources.

Evaluation is discussed in two journal special issues: *Computers in the Humanities 34(1–2)* (special issue on Senseval, edited by Kilgarriff and Palmer, 2000) and *Natural Language Engineering 8(4)* (special issue on evaluating word sense disambiguation systems, edited by Edmonds and Kilgarriff, 2002).

The main venues for research papers in WSD are the journals *Computational Linguistics* and *Natural Language Engineering*, and the conference proceedings of the Association for Computational Linguistics (ACL), the International Conference on Computational Linguistics (COLING), and their associated organizations, special interest groups (SIGs), and workshops.

Polysemy is of course discussed frequently in the lexical semantics literature. Cruse's (1986) *Lexical Semantics* gives a solid overview of polysemy, and acts as a good starting point for further reading. Lyons' (1995) *Linguistic Semantics* is worth consulting. Ravin and Leacock's (2000) *Polysemy: Theoretical and Computational Approaches* is a recent summary of activity, with three chapters about computational approaches.

References

Allen, James. 1995. *Natural Language Understanding*. Redwood City, California: Benjamin Cummings.

ALPAC. 1966. *Language and Machine: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee,

Division of Behavioral Sciences, National Research Council. Washington, D.C.: National Academy of Sciences.

Atkins, Sue. 1991. Tools for computer-aided corpus lexicography: The Hector project. *Acta Linguistica Hungarica*, 41:5–72.

Baker, Collin F., Charles J. Fillmore & Beau Cronin. 2003. The structure of the FrameNet database. *International Journal of Lexicography*, 16(3):281–296.

Bar-Hillel, Yehoshua. 1960. The present status of automatic translation of languages. *Advances in Computers*, ed. by Franz Alt et al., 91–163. New York: Academic Press.

Bhattacharya, Indrajit, Lise Getoor & Yoshua Bengio. 2004. Unsupervised word sense disambiguation using bilingual probabilistic models. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, 288–295.

Black, Ezra. 1988. An experiment in computational discrimination of English word senses. *IBM Journal of Research and Development*, 32(2):185–194.

Bloehdorn, Stephan & Andreas Hotho. 2004. Text classification by boosting weak learners based on terms and concepts. *Proceedings of the Fourth IEEE International Conference on Data Mining*, 331–334.

Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra & Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berkeley, California, 264–270.

Carpuat, Marine & Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. Proceedings of the *43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan, 387–394.

Chapman, Robert. 1977. *Roget's International Thesaurus (Fourth Edition)*. New York: Harper and Row.

Chlovski, Timothy & Rada Mihalcea. 2002. Building a sense tagged corpus with Open Mind Word Expert. *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, PA, USA, 116–122.

Clough, Paul & Mark Stevenson. 2004. Cross-language information retrieval using EuroWordNet and word sense disambiguation. *Advances in Information Retrieval, 26th European Conference on IR Research (ECIR)*, Sunderland, UK, 327–337.

Cowie, Jim, Joe A. Guthrie & Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, Nantes, France, 359–365.

Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Dagan, Ido, Oren Glickman & Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Dale, Robert, Hermann Moisl & Harold Somers, eds. 2000. *Handbook of Natural Language Processing*. New York: Marcel Dekker.

Diab, Mona. 2003. *Word Sense Disambiguation within a Multilingual Framework*. Ph.D. Thesis, Department of Linguistics, University of Maryland, College Park, Maryland.

Dill, Stephen, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin & Jason Y. Zien. 2003. SemTag and Seeker: Bootstrapping the Semantic Web via automated semantic annotation. *Proceedings of the Twelfth International Conference on World Wide Web (WWW-2003)*, Budapest, Hungary, 178–186.

Edmonds, Philip & Scott Cotton. 2001. Senseval-2: Overview. *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, 1–5.

Edmonds, Philip & Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4):279–291.

Edmonds, Philip. 2005. Lexical disambiguation. *The Elsevier Encyclopedia of Language and Linguistics, 2nd Ed.*, ed. by Keith Brown, 607–23. Oxford: Elsevier.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Gale, William, Kenneth Church & David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL)*, Newark, Delaware, 249–256.

Gildea, Daniel & Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Guthrie, Joe A., Louise Guthrie, Yorick Wilks & Homa Aidinejad. 1991. Subject dependent co-occurrence and word sense disambiguation. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berkeley, California, 146–152.

Hanks, Patrick. 2000. Do word meanings exist? *Computers in the Humanities*, 34(1–2):205–215.

Hirst, Graeme. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge, UK: Cambridge University Press.

Ide, Nancy & Jean Véronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Jurafsky, Daniel & James H. Martin. 2000. *Speech and Language Processing*. New Jersey, USA: Prentice Hall.

Kaplan, Abraham. 1950. An experimental study of ambiguity and context. Mimeographed, 18pp, November 1950. Published as: Kaplan, Abraham. 1955. An experimental study of ambiguity and context. *Mechanical Translation*, 2(2):39–46.

Kelly, Edward F. & Philip J. Stone. 1975. *Computer Recognition of English Word Senses*. Amsterdam: North-Holland.

Kilgarriff, Adam. 1997. "I don't believe in word senses". *Computers in the Humanities*, 31(2):91–113.

Kilgarriff, Adam & Martha Palmer. 2000. Introduction to the special issue on Senseval. *Computers and the Humanities*, 34(1–2):1–13.

Karin Kipper, Hoa Trang Dang & Martha Palmer. 2000. Class-based construction of a verb lexicon. *Proceedings of the Seventh National Conference on Artificial Intelligence* (AAAI-2000), Austin, Texas.

Lesk, Michael. 1986. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 1986 ACM SIGDOC Conference*, Toronto, Canada, 24–26.

Li, Hang & Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 30(1):1–22.

Lyons, John. 1995. *Linguistic Semantics: An Introduction*. Cambridge, UK: Cambridge University Press.

Madhu, Swaminathan & Dean W. Lytle. 1965. A figure of merit technique for the resolution of non-grammatical ambiguity. *Mechanical translation*, 8(2):9–13.

Maedche, Alexander & Steffen Staab. 2001. Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2):72–79.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Masterman, Margaret. 1957. The thesaurus in syntax and semantics. *Mechanical Translation*, 4(1–2):35–43.

McCarthy, Diana, Rob Koeling, Julie Weeds & John Carroll. 2004. Finding predominant senses in untagged text. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Barcelona, Spain, 280–287.

Mihalcea, Rada, Timothy Chlovski & Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, 25–28.

Mihalcea, Rada & Philip Edmonds, eds. 2004. *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

Miller, George A., ed. 1990. Special Issue, WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4).

Ng, Hwee Tou & Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, 40–47.

Ng, Hwee Tou, Bin Wang & Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, 455–462.

Palmer, Martha, Christiane Fellbaum, Scott Cotton, Lauren Delfs & Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, 21–24.

Palmer, Martha, Christiane Fellbaum & Hoa Trang Dang. 2006. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 12(3).

Preiss, Judita & Mark Stevenson, eds. 2004. *Computer, Speech, and Language*, 18(4). (Special issue on word sense disambiguation)

Procter, Paul, ed. 1978. *Longman Dictionary of Contemporary English.* London: Longman Group.

Quillian, M. Ross. 1968. Semantic memory. *Semantic Information Processing*, ed. by Marvin Minsky, 227–270. Cambridge, MA: MIT Press.

Ravin, Yael & Claudia Leacock. 2000. *Polysemy: Theoretical and Computational Approaches*. Oxford University Press.

Reifler, Edwin. 1955. The mechanical determination of meaning. *Machine Translation of Languages*, ed. William Locke & Donald A. Booth, 136–164. New York: John Wiley & Sons.

Resnik, Philip & David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Rieger, Chuck & Steven Small. 1979. Word expert parsing. *Proceedings of the 6th International Joint Conference on Artificial Intelligence (IJCAI)*, 723–728.

Ruhl, Charles. 1989. *On Monosemy: A Study in Linguistic Semantics*. Albany: State University of New York Press.

Sanderson, Mark. 1994. Word sense disambiguation and information retrieval. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 142–151.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Small, Steven, Garrison Cottrell & Michael Tanenhaus, eds. 1988. *Lexical Ambiguity Resolution: Perspectives from Artificial Intelligence, Psychology and Neurolinguistics.* San Mateo: Morgan Kaufman.

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

Stevenson, Mark. 2003. *Word Sense Disambiguation: The Case for Combination of Knowledge Sources*. Stanford, USA: CSLI Publications.

Tufiş, Dan, Radu Ion & Nancy Ide. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering, and aligned wordnets. *Proceedings of the Twentieth International Conference on Computational Linguistics (COLING)*, Geneva, 1312–1318.

Tuggy, David H. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4:273–90.

Véronis, Jean. 2004. Hyperlex: Lexical cartography for information retrieval. *Computer, Speech and Language*, 18(3):223–252.

Voorhees, Ellen M. 1993. Using WordNet to disambiguate word senses for text retrieval. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, Pennsylvania, 171–180.

Vossen, Piek, German Rigau, Iñaki Alegria, Eneko Agirre, David Farwell & Manuel Fuentes. 2006. Meaningful results for information retrieval in the MEANING project. *Proceedings of the 3rd Global Wordnet Conference*, Jeju Island, Korea.

Weaver, Warren. 1949. Translation. Mimeographed, 12 pp. Reprinted in William N. Locke & Donald A. Booth, eds. 1955. *Machine Translation of Languages*, 15–23. New York: John Wiley & Sons.

Weiss, Stephen. 1973. Learning to disambiguate. *Information Storage and Retrieval*, 9:33–41.

Wilks, Yorick. 1975. Preference semantics. *Formal Semantics of Natural Language*, ed. by E. L. Keenan, III, 329–348. Cambridge, UK: Cambridge University Press.

Wilks, Yorick, Dan Fass, Cheng-Ming Guo, James E. MacDonald, Tony Plate & Brian A. Slator. 1990. Providing machine tractable dictionary tools. *Semantics and the Lexicon*, ed. by James Pustejovsky, 341–401. Dordrecht: Kluwer Academic Publishers.

Wilks, Yorick, Louise Guthrie & Brian Slator. 1996. *Electric Words*. Cambridge, MA: MIT Press.

Yarowsky, David. 1992. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, Nantes, France, 454–460.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Cambridge, MA, 189–196.

Yarowsky, David. 2000. Word-sense disambiguation. *Handbook of Natural Language Processing*, ed. by Dale et al., 629–654. New York: Marcel Dekker.

Zipf, George Kingsley. 1949. *Human Behaviour and the Principle of Least Effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley. Reprinted by New York: Hafner, 1972.