Word Sense Disambiguation: A Seventy Year Perspective

Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, New York USA
ide@cs.vassar.edu

Abstract. This chapter considers automatic word sense disambiguation (WSD) methodologies since the late 1940s and examines the influences and advances that have been made over the past 70 years. One of the goals of the chapter is to bring current WSD methodology into the perspective of the past 70 years of work on the topic, and consider, in the context of its entire history–both methodological and theoretical–possible next directions. A second goal is to make researchers aware of this long history and the commonalities of methodology that characterize it, both in order to appreciate the intellectual development within WSD research and make the point that many "new" approaches often have their roots in much earlier (and rarely cited) work.

Keywords. Word sense disambiguation, history of WSD.

1 Introduction

Work on automatic word sense disambiguation (WSD) has a history as long as automated language processing generally. Throughout the 1950s and into the 1960s there was continuous work on WSD, implementing a variety of strategies including various statistical analyses, the use of external knowledge from dictionaries and thesauri, the use of semantic and neural networks, and computation of sense probabilities based on data from domain-specific texts. However, without large-scale resources most of these ideas remained untested and to large extent forgotten until several decades later, when, following the roughly 20-year "symbolic era" of NLP, the availability of large-scale resources such as corpora, dictionaries, and thesauri engendered a revival of empirical methods. With sufficiently greater resources and enhanced statistical methods at their disposal, researchers in the 1990s improved on earlier results, but the methods applied were fundamentally those developed thirty years before. Now, twenty years later, researchers have at their disposal a new resource: the internet, which provides not only orders of magnitude more language data for study but also massive inter-linkage among these data. This has sparked most recently a focus on graph-based methods that exploit inter-connections in resources such as Wikipedia as well as inter-linked lexicons such as BabelNet, and the development of more and more linguistic linked data in the Semantic Web promises to provide even more information-rich resources of this kind. Nonetheless, at its roots WSD methodology has remained relatively consistent, even up to this day; the major changes have been due to the availability of more resources and enhanced statistical methods, but in the broadest sense, relatively little in terms of fundamental methodology has changed over nearly 70 years. More to the point, WSD work continues to rely on the notion of the graphic word as the unit of analysis and the existence of (more or less) discrete senses associated with each word.

This chapter considers automatic WSD methodologies since the late 1940s and examines the influences and advances that have been made over the past 70 years. One of the goals of the chapter is to bring current WSD methodology into the perspective of the past 70 years of work on the topic, and consider, in the context of its entire history–both methodological and theoretical–possible next directions. A second goal is to make researchers aware of this long history and the commonalities of methodology that characterize it, both in order to appreciate the intellectual development within WSD research and make the point that many "new" approaches often have their roots in much earlier (and rarely cited) work–i.e., to make the point that "Those Who Ignore History Are Doomed To Repeat It".[1]

[1] A variant of similar observations made by Edmund Burke, George Santayana, and others.

2 History

A comprehensive survey of work on automated WSD through 1998 is given in [40]; coverage of more recent developments in the area is provided in [1] and [59]. Some of the material in this section is derived from those and other sources, but it is structured differently in order to highlight the similarities in approach over the past seventy years as well as developments that have moved the field forward. This section covers WSD work in what is commonly regarded as the "first age of empiricism" in NLP from roughly 1950 to 1970 [14, 13], which was focused on machine translation, and considers its relation to work in the second NLP's second era of empiricism dating from roughly 1990 to the present. Our premise is that despite various methodological tweaks over the years, very little has changed, and the striking fact is the degree to which the fundamental problems and approaches to the problem were foreseen and developed in the earliest days of automatic WSD. However, the shifted methodological focus in the intervening "rationalist" or "symbolic" era between 1970 and 1990 seems to have obliterated much of the memory of these early achievements, dooming the field to repeat much of its history in the 1990s and later.

The history of automatic WSD begins with Warren Weaver, who in 1947 first suggested using computers to translate documents between natural human languages. In 1949 he produced a memorandum entitled "Translation" [98], generally regarded as the single most influential publication in the early days of machine translation (MT), which outlined a series of methods for that task. Given that the problem of multiple meanings poses one of the most immediate obstacles to MT, the first method he proposed suggests a strategy for automatic WSD:

> If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. [...] 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 of the central word. [...]The practical question is: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?"

Thus the relation between a word and its context was established as the determining factor in WSD.[2] This has been true since the 1950s and early 1960s, and remains true today; it has provided the basis for WSD work in MT, content analysis, AI-based disambiguation, dictionary-based WSD, as well as statistical, neural network, and machine learning approaches, all of which rely on information in local context to disambiguate a given word occurrence.

As Weaver notes, the size of the context window is a matter for further consideration. In early experiments, Kaplan [42] and others (e.g., [54, 47]) determined that the answer to this question is "two". Later, working in the field of Humanities Computing, Choueka and Lusignan [12] re-verified Kaplan's finding that 2-contexts are highly reliable for disambiguation and showed that even 1-contexts are reliable in 8 out of 10 cases, but this work was largely unknown in the NLP community. Almost forty years later, in the 1990s, WSD researchers spent considerable effort experimenting with window size, presumably with little awareness of the much earlier work. For example, Yarowsky [107, 108, 110] examined several different windows of local context in an attempt to find the most reliable evidence for disambiguation, eventually (along with other studies such as [49] coming to more or less the same conclusion: for the most part, two- or three-word windows work as well as any other.[3]

Researchers in the MT era quickly refined the notion of context to include only nearby words in certain relations to the target, in particular syntactic relations; Reifler [74] says:

> Grammatical structure can also help disambiguate, as, for instance, the word keep, which can be disambiguated by determining whether its object is gerund (He kept eating), adjectival phrase (He kept calm), or noun phrase (He kept a record).

WSD work in the second empirical era follows this practice, utilizing information from shallow or partial parsing [33], relative distance, order, and syntactic relationship, and in some cases, domain or topic [108, 109,

[2] Encapsulated a few years later in J. R. Firth's infamous quote, "You shall know a word by the company it keeps" [24]

[3] Optimum window size for WSD has recently received some renewed attention in the biomedical field (e.g., [57, 96]).

7, 110, 49]. The common wisdom is that different kinds of disambiguation procedures are needed depending on syntactic category and other (unidentified) characteristics of the target word; however, to date there has been little systematic study of the contribution of different information types for different types of target words, possibly because to do so requires dealing with stickier semantic issues that are less amenable to empirical methods. Context is also frequently defined as all words or characters falling within some window of the target, with no regard for distance, syntactic, or other relations, starting with early corpus-based work (e.g., [99]) and continuing to the present day.

Weaver's observation about the role of local context in automatic WSD was not the only fundamental insight into WSD methodology provided in his memorandum. He also discusses the role of the domain in sense disambiguation:

> In mathematics, to take what is probably the easiest example, one can very nearly say that each word, within the general context of a mathematical article, has one and only one meaning.

Forty years later, Gale and others produced their well known paper entitled"One Sense per Discourse" [26], which is now routinely cited as the seminal article on domain-driven disambiguation. However, back in the 1950s this idea played a central role in WSD for MT, as evidenced by the extensive development and use of specialized dictionaries or "micro-glossaries" ([65, 20, 63, 64] to name just a few) containing only the meaning of a given word relevant to a particular domain of discourse. Building on this work, Madhu and Lytle [52] calculated sense frequency for texts in different domains and applied a Bayesian formula to determine the probability of each sense in a given context–although the terminology is different, the use of a metric to provide a "default" sense is very similar to the notion of baseline tagging used in modern work (see, e.g., Gale et al., 1992b). This work also provides an early example of "domain-tuning" for WSD[4] (recent examples include [11, 28, 5, 69]). Interestingly, domain tuning effectively identifies a sense subset applicable to a particular domain, thus effectively coming full circle back to the micro-glossaries of the MT era.

Perhaps the most ambitious work in the early years of NLP, and the one that expanded WSD methodologies more than any other, is that of Masterman, who constructed the first machine-implemented knowledge base from *Roget's Thesaurus* [53] and applied it to WSD for machine translation.[5]. *Roget's* was used heavily in subsequent WSD work in the 1970s and 80s [8, 9, 85, 86] in studies seemingly unknown to the NLP community; the current most well known use of *Roget's* for WSD is Yarowksy's [107]. To translate from a given source language to English, Masterman looked up, for each word stem, the translation in a bi-lingual dictionary. She then looked up each English equivalent in the word-to-head index of *Roget's*, thus associating each word stem in the source text with a list of Roget head numbers related to its English equivalents. The head numbers for all words appearing in the same sentence were examined for overlaps; any of the English equivalents appearing under one or more of the overlapping head categories were chosen for the translation. Masterman's methodology is strikingly similar to that underlying much of the knowledge-based WSD undertaken in the 1990s, but again, this work is rarely cited in this more recent work that echoes her methodology.

Masterman's early work also foreshadowed graph-based approaches to WSD that are increasingly common today. She created a semantic network from which she derived the representation of sentences in an "interlingua" [78] comprised of fundamental, language-independent language concepts and used it to solve the disambiguation problem [53]. To do this, she built a 15,000 entry concept dictionary, where concept types are organized in a lattice with inheritance of properties from super-concepts to sub-concepts. Sense distinctions were implicitly made by choosing representations that reflect groups of closely related nodes in the network, anticipating the clustering approaches used in much later work [81, 82, 83, 67, 84]. This and similar work [78] laid the groundwork for Quillian's semantic network [71, 73, 72], the earliest implementation of a spreading activation ("neural") network used for word sense disambiguation. Quillian's network, which was created from dictionary definitions enhanced by hand-encoded knowledge, directly informed later dictionary-based neural networks applied to WSD [15, 38] as well as dictionary-based approaches to WSD that were popular in the early 1990s in general.

[4] See [10] for an overview.

[5] This belies the claim that "The first knowledge-based approaches to WSD date back to the 1970s and 1980s when experiments were conducted on extremely limited domains" [59]

Masterman's methodology is also closely related to the subsequently developed notion of *lexical chains* [30], which in turn informs recent graph-based approaches. In these studies, graphs are built from different resources, including cooccurrence data [100, 2], lexicons such as WordNet [27, 62], or lexicons and other knowledge sources [88]. As in earlier work, nodes in the graph are words or concepts, and edges are typically labeled by grammatical or semantic relations or are unlabeled and used to compute degree of association via proximity; various algorithms (e.g. PageRank, graph-traversal algorithms) are then applied to identify the most closely related nodes, à la Masterman.

3 The rational era

Warren emphasizes the importance of statistics for language analysis in his memorandum:

> This approach brings into the foreground an aspect of the matter that probably 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.

As shown above, empiricism played an important role in MT work (see also e.g. Richards, 1953; Yngve, 1955; Parker-Rhodes, 1958), but this characterization obscures the fact that Weaver and the body of research that followed from his memorandum also engendered methodologies that are typically associated with the "rational" or "symbolic" era of NLP, which is said to have emerged after the ALPAC Report [68] led to the eventual withdrawal of large-scale funding for MT and continued for twenty years. This era was dominated by the field of Artificial Intelligence (AI), which began to attack the problem of WSD in the context of larger systems intended for full language understanding. In the spirit of the times, such systems were almost always grounded in some theory of human language understanding that they attempted to model, and often involved the use of detailed knowledge about syntax and semantics to perform their task. Not surprisingly, this knowledge was exploited for WSD.

Another proposal in Weaver's memorandum was that translation could be addressed as a problem of formal logic, by deducing "conclusions" in the target language from "premises" in the source language. His hypothesis derives from McCulloch and Pitts' 1943 article entitled "A Logical Calculus of Ideas Immanent in Nervous Activity" [56], which described an early type of neural nets. In the AI era, various approaches following from this hypothesis were applied to WSD: Kelley and Stone's decision trees for WSD [43] were an early example, followed by Small and Reiger's Word Expert Parser [89], which attempted to provide a procedure for each word that is "cognizant of all possible contextual interpretations of the word it represents" (p. 9).[6] Dahlgren's reasoning module [19] traverses an ontology to find common ancestors for words in context, thus anticipating Resnik's [76] results by determining that ontological similarity involving a common ancestor in the ontology is a powerful disambiguator. In the 1980s and early 90s, following on earlier work by Quillian, neural nets were applied to WSD: for example, Cottrell and Small [15] employed them to represent words as nodes, which activate the concepts to which they are semantically related and vice versa; and Ide and Véronis [39] built a neural network that connects each word to its senses and, in turn, each sense to the words occurring in its definition.[7]

In an interesting development, neural networks designed by Waltz and Pollack [97] and Bookman [4] were based not on words, but rather on "microfeatures" corresponding to fundamental semantic distinctions (animate/inanimate, edible/inedible, threatening/safe, etc.), characteristic duration of events (second, minute, hour, day, etc.), locations (city, country, continent, etc.), and other similar distinctions. This approach, as well as Masterman's much earlier "interlingua", speaks to another of Weaver's proposals in his memorandum: that there may be linguistic universals underlying all human languages which could be exploited to make the problem of translation more straightforward. This view is not unrelated to Wilks' Preference Semantics

[6] The Word Expert Parser, although procedural rather than descriptive, is in some ways a precursor of Kilgarriff's "word sketch", which similarly attempts to encapsulate all contextual interpretations of a given word.

[7] Recently, Tsatsaronis et al. [94] successfully extended this approach to include all related senses linked by semantic relations in WordNet.

approach to meaning representation [105], which uses Masterman's primitives and was the first case-based approach specifically intended for sense disambiguation. The introduction of frame semantics [23] provoked a major leap in WSD and NLP generally, and as a result, frames were widely used in combination with semantic or neural networks for WSD during this period (e.g., [32, 34]).

AI era approaches to WSD suffered from many of the same problems as generalized AI-based language understanding systems: the need for extremely large and complex rule systems, the difficulty of hand-crafting knowledge sources and training data, and the need for massive computing facilities for processing large networks and rule systems. As a result, these strategies were restricted to "toy" implementations handling only a tiny fraction of the language and/or testing on only a very small test set in a limited context (most often, a single sentence), making it impossible to determine their effectiveness on real texts. [8] This problem persisted into the 1990s: for example, although neural networks have been shown to perform well compared to other supervised methods [50, 93, 58], due to lack of facilities, experiments are often performed on a small number of words and are therefore inconclusive.

4 What now?

The 1990s ushered in the so-called "second era of empiricism" in NLP, which continues through the present. In 1998, Ide and Véronis [40] observed:

> In a sense WSD work effectively came full circle, returning to empirical methods and corpus-based analyses that characterize some of the earliest attempts to solve the problem. With sufficiently greater resources and enhanced statistical methods at their disposal, researchers in this period have obviously improved on earlier results, but it appears that we may have reached near the limit of what can be achieved in the current framework.

Eight years later, Ide and Wilks [41] made a similar remark:

> At present, WSD work is at a crossroads: systems have hit what is seemingly a ceiling of 70%+ accuracy...the source and kinds of sense inventories that should be used in WSD work is an issue of continued debate, and the usefulness of stand-alone WSD systems for current NLP applications is questionable.

With little or no modification, these statements still apply. This is not to say that nothing has been achieved over the past twenty years–WSD work during this period has, at least, served as a hothouse of techniques, and recent WSD research has had a boost from the availability of web resources and large-scale networks created from them that combine encyclopedic and lexicographic knowledge (e.g., BabelNet [61]). However, taken from a broader perspective, methodologically the field has not progressed much beyond the first empirical era: consider that current machine learning techniques for WSD rely almost entirely on feature sets consisting of context words, domain and topic words, and syntactic features [59]. The availability of reasonably-sized bodies of sense-tagged data has provided what could be seen as new features for supervised machine learning that have a semantic basis, but at the same time, more sophisticated semantic features based on selectional restrictions, frame semantics, etc. that figured in work in the 1980s are rarely used, despite the development of resources such as FrameNet [3]. Results hover around the same 80-85% level, occasionally inching upward, but no "breakthrough" methods or results are apparent now or on the horizon.

If the history of NLP methodologies over the past sixty or seventy years can be characterized as a pendulum swinging from extreme empiricism to extreme rationalism [13], we are due for a return to the rational approach. This time the pendulum swings back, it may do us well to consider work from the previous era as a starting point–especially possibilities for semantic representations such as Waltz and Pollack's microfeatures, or the semantic primitives in Richens' MT-era interlingua [78] and those that served as a center for Wilks' work for decades (see, for example, [101, 102, 104]). Resources like FrameNet [3] go in the right direction by organizing lexical entries into semantic frames and defining frame slots with case primitives, but as

[8] In addition to these problems, it is interesting to note that many of the AI-based disambiguation results involve highly ambiguous words and fine sense distinctions (e.g., ask, idea, hand, move, use, work, etc.) and unlikely test sentences (The astronomer married the star), which make the results even less easy to evaluate in the light of the now-known difficulties of discriminating even gross sense distinctions.

noted above, this kind of information is rarely used for WSD. Instead, WSD has relied almost exclusively on enumerated sense lists in dictionaries, thesauri, and–in particular–lexicons such as WordNet to provide a representation of meaning, despite widespread skepticism about the appropriateness of sense lists from dictionaries and similar resources for WSD [103, 22, 66, 41].

Attempts to avoid pre-defined sense inventories, starting from the earliest days of WSD, have used the same kinds of information used in most recent WSD work (context, syntactic role, etc.) to identify groups of occurrences that can be seen as representing a distinct sense (e.g., [53, 90, 84, 21, 75]), or have been based on cross-lingual or multi-lingual correspondences [6, 18, 77, 35, 36, 37, 51, 60]. These data-driven approaches are philosophically in the good company of major 20th century linguists (Halliday, Sinclair, Harris) and lexicographers (Atkins, Kilgarriff). Famously, Adam Kilgarriff's "I Don't Believe in Word Senses" [44] argued that senses are abstractions from clusters of corpus citations rather than isolatable entities that can be definitively itemized, and that distinct senses thus exist only relative to a task. Coming from someone with a foot in the field of computational linguistics, Kilgarriff's view had significant impact on WSD methodology in the years to follow. For example, in recent work WSD is re-cast as a *lexical substitution* task [17, 55], in which algorithms select a substitute word that preserves as much as possible the meaning of the target word, in order to avoid the use of a pre-defined inventory (e.g., [92, 91]). However, automatic WSD necessarily requires distinguishable meaning representations, and to serve as such, groups of corpus citations must be somehow labeled and distinguished in ways similar to a traditional dictionary. As a result, despite the claim to move away from the use of sense inventories, these studies typically invoke the sense distinctions in lexical and other knowledge resources, either explicitly or implicitly, by relying on information (e.g. synonyms, ontological information, paraphrases) associated with a given word, and resembling the same types of information that have characterized WSD work for decades.

A different approach to representing word meaning, coupled with neural networks as in the 1980s work of Waltz and Pollack [97] and Bookman [4], re-opens additional possibilities for WSD. In the 1990s, in what felt like a revolt against the newly-named field of cognitive science, it was common to assert that NLP work focused on enabling computers to understand language without concern for how humans might actually do it–i.e., that a computer might work very differently from the human brain, and human-like language understanding by a computer might therefore be accomplished using entirely different mechanisms. Hence, apart from the persistence of simple neural models in machine learning algorithms, approaches that paid any homage to neurological models went out of fashion. However, now, the trend toward "deep learning" offers the possibility to explore complex meaning representations that more closely resemble neurological models of the human brain. The potential to learn rather than pre-determine features could lead to semantic representations that are not simply word-based and engender a return to approaches reminiscent of Waltz and Pollacks microfeatures or Wilks semantic primitives. The availability of far greater computing resources and, eventually, "neuromorphic" chips that mimic the neurons and synapses of biological brains suggests an inevitable direction for NLP that is, in fact, brain-based. The use of this kind of model for meaning representation could also serve to better integrate WSD with other levels of processing, rather than attacking the task as if disambiguation is performed in isolation from other linguistic processes.

Work in other fields can also inform an approach to WSD. For example, as discussed in [41], there have been numerous studies in psycholinguistics dealing with the distinctness of word senses in terms of their representations in the mental lexicon, and the degree to which different senses of a polysemous word correspond to a unified taxonomic, thematic, or *ad hoc* category (e.g., [45, 46, 79, 80]). Kilgarriff [44] discusses a psycholinguistic study that examines the effects of "semantic priming" [106] and proposes that meaning representations in the mental lexicon are hierarchical, with some senses of a given word more prominent or accessible than others. Later psycholinguistic studies argue that readers and listeners retrieve an underspecified meaning ("radical underspecification") when presented with a given word, which is later fleshed out based on contextual information, as opposed to retrieving one or more specific senses [25, 31, 87]. Some of this work resonates with the "sense chaining" and "generative" approaches to sense determination proposed by linguists and computational linguists (e.g., [16, 48, 70]) in which meanings extend in systematic ways from core senses; in contrast, other psycholinguistic studies suggest that the types of relationships among senses are not at all systematic, but rather more or less random and unpredictable [46]. Kilgarriff noted twenty years ago that "priming experiments do show potential for providing a theoretical grounding for

distinguishing ambiguity and generality" ([44], p. 99). Nonetheless, today there is little awareness of work in psycholinguistics in the field of WSD, despite its obvious relevance and the possibility that consideration of meaning representation from this very different perspective could inform our own.

Another source to consult for ideas about meaning is linguistic theory itself, which often seems to have been eclipsed in this second empirical age. In the AI era, those working on language processing were computational *linguists*, in the sense that they grounded their efforts to get computers to understand language (typically with a particular goal, such as translation) in linguistic theories of meaning, and they were often active participants in debates among philosophers and linguistics about how humans process and understand language. Given the computing and resource constraints of the time, this inevitably led to complex but useless implementations, on a practical level, and eventual abandonment. From a modern perspective, where the focus is on finding something that works without much concern for why it works, or whether it has anything to do with actual human processing, theory seems like unnecessary baggage; but even statistical approaches to language processing reflect a theoretical position about the base upon which they build, even if it is not acknowledged. At the same time, if not directly in the NLP world, the Semantic Web and ontologies communities are reviving theoretical questions about meaning, such as the relation between domain-independent and domain-dependent semantic structures, the relation between linguistic and world knowledge, etc.; and some recent work in NLP has addressed similar questions (e.g., [95, 29]). WSD work, in particular, has not yet gone in this direction, but inevitably will in the near future, especially as "linked data" becomes more central to NLP.

5 Conclusion

This chapter provides an abridged and incomplete accounting of the early days of WSD and its influence on later research, and ignores some important WSD work over the past 25 years. However, the goal is not to be comprehensive, but rather to demonstrate that sometimes seemingly "new" methods have their roots in much earlier work.

To move into the next era of NLP, we could do worse than to revisit some of the theories and approaches that have been set aside over the past 30 years, as well as broaden the perspective to encompass different approaches to a representation of meaning. It is worth noting that Adam Kilgarriff pushed for exactly this kind of different and broader view of word senses, most notably in his seminal "I Don't Believe in Word Senses" in 1997, but also throughout his career in his wide-ranging work on lexicography and his development of the "word sketch", a summary of a word's grammatical and collocational characteristics automatically derived from corpora. Hopefully this volume will not only pay homage to Adam's substantial contributions to the fields of WSD and lexicography, but also inspire a move toward the more comprehensive and informed approach to disambiguating language that he championed.

6 Acknowledgements

My thanks to Roberto Navigli, who provided valuable comments and input.

References

1. Agirre, E., Edmonds, P. (eds.): Word Sense Disambiguation: Algorithms and Applications, 1st edn. Springer Publishing Company, Incorporated (2007)
2. Agirre, E., Martínez, D., de Lacalle, O.L., Soroa, A.: Two graph-based algorithms for state-of-the-art wsd. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, pp. 585–593. Association for Computational Linguistics, Stroudsburg, PA, USA (2006)

3. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, ACL '98, pp. 86–90. Association for Computational Linguistics, Stroudsburg, PA, USA (1998)
4. Bookman, L.A.: A microfeature-based scheme for modelling semantics. In: Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 2, pp. 611–614. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1987)
5. Boyd-Graber, J.L., Blei, D.M., Zhu, X.: A topic model for word sense disambiguation. In: EMNLP-CoNLL, pp. 1024–1033. ACL (2007)
6. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Comput. Linguist. **16**(2), 79–85 (1990)
7. Bruce, R., Wiebe, J.: Word-sense disambiguation using decomposable models. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 139–145. Las Cruces, New Mexico (1994)
8. Bryan, R.M.: Abstract thesauri and graph theory applications to thesaurus research. In: S.Y. Sedelow (ed.) Automated Language Analysis, 1972-3, pp. 45–89. University of Kansas Press, Lawrence, Kansas (1973)
9. Bryan, R.M.: Modelling in thesaurus research. In: S.Y. Sedelow (ed.) Automated Language Analysis, 1973-4, pp. 44–59. University of Kansas Press, Lawrence, Kansas (1974)
10. Buitelaar, P., Magnini, B., Strapparava, C., Vossen, P.: Domain-specific wsd. In: E. Agirre, P. Edmonds (eds.) Word Sense Disambiguation: Algorithms and Applications, *Text, Speech and Language Technology*, vol. 33, pp. 275–298. Springer, Dordrecht, The Netherlands (2006)
11. Buitelaar, P., Sacaleanu, B.: Ranking and selecting synsets by domain relevance. In: Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations (2001)
12. Choueka, Y., Lusignan, S.: Disambiguation by short contexts. Computers and the Humanities **19**(3), 147–157 (1985)
13. Church, K.: A pendulum swung too far. In: Linguistic Issues in Language Technology 6(5 (2011)
14. Church, K.W., Mercer, R.L.: Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics **19**(1), 1–24 (1993)
15. Cottrell, G.W., Small, S.L.: A connectionist scheme for modelling word sense disambiguation. Cognition and Brain Theory **6**, 89–120 (1983)
16. Cruse, D.: Lexical Semantics. Cambridge University Press, Cambridge, UK (1986)
17. Dagan, I., Glickman, O., Gliozzo, A.M., Marmorshtein, E., Strapparava, C.: Direct word sense matching for lexical substitution. In: ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (2006)
18. Dagan, I., Itai, A.: Word sense disambiguation using a second language monolingual corpus. Comput. Linguist. **20**(4), 563–596 (1994)
19. Dahlgren, K.: Naive semantics for natural language understanding. Kluwer Academic Publishers, Boston (1988)
20. Dostert, L.E.: The georgetown-I.B.M. experiment. In: W.N. Locke, A.D. Booth (eds.) Machine translation of languages, pp. 124–135. John Wiley & Sons, New York (1955)
21. Erk, K., Padó, S.: A structured vector space model for word meaning in context. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, pp. 897–906. Association for Computational Linguistics, Stroudsburg, PA, USA (2008)
22. Fellbaum, C., Palmer, M., Dang, H.T., Delfs, L., Wolf, S.: Manual and automatic semantic annotation with wordnet. In: Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, pp. 3–10. Pittsburgh, Pa (2001)
23. Fillmore, C.J.: The case for case. In: E. Bach, R.T. Harms (eds.) Universals in Linguistic Theory, pp. 0–88. Holt, Rinehart and Winston, New York (1968)
24. Firth, J.R.: A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis (special volume of the Philological Society) **1952-59**, 1–32 (1957)
25. Forakera, S., Murphy, G.L.: Polysemy in sentence comprehension: Effects of meaning dominance. Journal of Memory and Language **67**(4), 407–425 (2012)
26. Gale, W.A., Church, K.W., Yarowsky, D.: One sense per discourse. In: Proceedings of the Workshop on Speech and Natural Language, HLT '91, pp. 233–237. Association for Computational Linguistics, Stroudsburg, PA, USA (1992)
27. Galley, M., Mckeown, K.: Improving word sense disambiguation in lexical chaining. In: Pro- ceedings of the 18th International Joint Conference on Artificial Intelligence, pp. 1486–1488 (2003)
28. Gliozzo, A.M., Magnini, B., Strapparava, C.: Unsupervised domain relevance estimation for word sense disambiguation. In: Proceedings of EMNLP, pp. 380–387. ACL (2004)
29. Goldwasser, D., Roth, D.: Leveraging domain-independent information in semantic parsing. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 462–466. The Association for Computational Linguistics (2013)
30. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman (1976)
31. Hargreaves, I.S., Pexman, P.M., Pittman, D.J., Goodyear, B.G.: Tolerating ambiguity ambiguous words recruit the left inferior frontal gyrus in absence of a behavioral effect. Experimental Psychology **58**, 19–30 (2011)
32. Hayes, P.J.: On semantic nets, frames and associations. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence, pp. 99–107. Cambridge, Massachusetts (1977)

33. Hearst, M.A.: Noun homograph disambiguation using local context in large text corpora. In: Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research, pp. 1–19. Oxford, UK (1991)
34. Hirst, G.: Semantic interpretation and the resolution of ambiguity. Studies in natural language processing. Cambridge University Press (1987)
35. Ide, N.: Cross-lingual sense determination: Can it work? Computers and the Humanities **34**(1-2), 223–234 (2000)
36. Ide, N., Erjavec, T., Tufis, D.: Automatic sense tagging using parallel corpora. In: Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, pp. 83–90. Tokyo, Japan (2001)
37. Ide, N., Erjavec, T., Tufis, D.: Sense discrimination with parallel corpora. In: Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions - Volume 8, pp. 61–66. Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
38. Ide, N., Véronis, J.: Very large neural networks for word sense disambiguation. In: Proceedings of the European Conference on Artificial Intelligence, pp. 366–368 (1990)
39. Ide, N., Véronis, J.: Very large neural networks for word sense disambiguation. In: Proceedings of the 9th European Conference on Artificial Intelligence, pp. 366–368 (1990)
40. Ide, N., Véronis, J.: Introduction to the special issue on word sense disambiguation: The state of the art. Computational Linguistics **24**(1), 2–40 (1998)
41. Ide, N., Wilks, Y.: Making sense about sense. In: E. Agirre, P. Edmonds (eds.) Word Sense Disambiguation: Algorithms and Applications, *Text, Speech and Language Technology*, vol. 33, pp. 47–74. Springer, Dordrecht, The Netherlands (2006)
42. Kaplan, A.: An experimental study of ambiguity and context. Mechanical Translation **2**, 39–46 (1955)
43. Kelly, E.F., Stone, P.J.: Computer Recognition of English Word Senses. North-Holland, Amsterdam (1975)
44. Kilgarriff, A.: I Don't Believe in Word Senses. Computers and the Humanities **31**(2), 91–113 (1997)
45. Klein, D.E., Murphy, G.L.: The representation of polysemous words. Journal of Memory and Language **45**, 259–282 (2001)
46. Klein, D.E., Murphy, G.L.: Paper has been my ruin: conceptual relations of polysemous senses. Journal of Memory and Language **47**(4), 548–570 (2002)
47. Koutsoudas, A.K., Korfhage, R.: M.T. and the problem of multiple meaning. Mechanical Translation **2**(2), 46–51 (1956)
48. Lakoff, G.: Women, Fire and Dangerous Things: What Categories Reveal About the Mind. University of Chicago Press, Chicago (1987)
49. Leacock, C., Miller, G.A., Chodorow, M.: Using corpus statistics and wordnet relations for sense identification. Computational Linguistics **24**(1), 147–165 (1998)
50. Leacock, C., Towell, G., Voorhees, E.: Corpus-based statistical sense resolution. In: Proceedings of the Workshop on Human Language Technology, pp. 260–265. Association for Computational Linguistics, Stroudsburg, PA, USA (1993)
51. Lefever, E., Hoste, V.: Semeval-2010 task 3: Cross-lingual word sense disambiguation. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 15–20. Association for Computational Linguistics, Uppsala, Sweden (2010)
52. Madhu, S., Lytle, D.W.: A figure of merit technique for the resolution of non-grammatical ambiguity. Mechanical translation **8**(2), 9–13 (1965)
53. Masterman, M.: The thesaurus in syntax and semantics. Mechanical Translation **4**(1-2) (1957)
54. Masterman, M.: Semantic message detection for machine translation using an interlingua. In: Proceedings of the International Conference on Machine Translation, pp. 438–475. Her Majesty's Stationery Office, London (1961)
55. McCarthy, D., Navigli, R.: Semeval-2007 task 10: English lexical substitution task. In: Proceedings of the 4^{th} International Workshop on Semantic Evaluations (SemEval-2007), pp. 48–53. Prague, Czech Republic (2007)
56. Mcculloch, W., Pitts, W.: A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics **5**, 127–147 (1943)
57. Moon S Pakhomov S, M.G.: Automated disambiguation of acronyms and abbreviations in clinical texts: Window and training size considerations. In: AMIA Annual Symposium Proceedings, pp. 1310–1319 (2012)
58. Mooney, R.J.: Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-96), pp. 82–91. Philadelphia, PA (1996)
59. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. **41**(2), 10:1–10:69 (2009)
60. Navigli, R., Jurgens, D., Vannella, D.: Semeval-2013 task 12: Multilingual word sense disambiguation. In: Proceedings of the 7^{th} International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013), pp. 222–231. Atlanta, USA (2013)
61. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012)
62. Navigli, R., Velardi, P.: Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. IEEE Transactions on Pattern Analysis for Machine Intelligence **27**(7), 1075–1086 (2005)
63. Oettinger, A.G.: The design of an automatic russian-english technical dictionary. In: W.N. Locke, A.D. Booth (eds.) Machine translation of languages, pp. 47–65. John Wiley & Sons, New York (1955)
64. Oswald, V.A.J.: The rationale of the idioglossary technique. In: L.E. Dostert (ed.) Research in Machine Translation, pp. 63–69. Georgetown University Press, Washington, D.C. (1957)
65. Oswald Victor A. Jr.Oswald, V.A.J., Lawson, R.H.: An idioglossary for mechanical translation. Modern Language Forum **38**(3/4), 1–11 (1953)

66. Palmer, M., Babko-Malaya, O., Dang, H.T.: Different sense granularities for different applications. In: HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding, pp. 49–56. Association for Computational Linguistics, Boston, Massachusetts, USA (2004)
67. Pedersen, T., Bruce, R.F.: Distinguishing word senses in untagged text. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). Providence, RI (1997)
68. Pierce, J., Carroll, J., Hamp, E., Hays, D., Hockett, C., Oettinger, A., Perlis, A.: Language and machines: Computers in translation and linguistics (1966)
69. Preiss, J., Stevenson, M.: Unsupervised domain tuning to improve word sense disambiguation. In: Proceedings of NAACL-HLT 2013, pp. 680–684 (2013)
70. Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge, MA, USA (1995)
71. Quillian, M.R.: A revised design for an understanding machine. Mechanical Translation **7**(1), 17–29 (1962)
72. Quillian, M.R.: The teachable language comprehender: A simulation program and theory of language. Communications of the ACM **12**(8), 459–476 (1969)
73. Quillian, R.: Semantic memory. In: M. Minsky (ed.) Semantic Information Processing, pp. 216–270. MIT Press (1968)
74. Reifler, E.: The mechanical determination of meaning. In: W.N. Locke, A.D. Booth (eds.) Machine translation of languages, pp. 136–164. John Wiley & Sons, New York (1955)
75. Reisinger, J., Mooney, R.J.: Multi-prototype vector-space models of word meaning. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, pp. 109–117. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
76. Resnik, P.: Semantic classes and syntactic ambiguity. In: Proceedings of the Workshop on Human Language Technology, pp. 278–283. Association for Computational Linguistics, Stroudsburg, PA, USA (1993)
77. Resnik, P., Yarowsky, D.: Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. Natural Language Engineering **5**(2), 113–133 (1999)
78. Richens, R.H.: Interlingual machine translation. Computer Journal **1**(3), 144–47 (1958)
79. Rodd, J., Gaskell, G.: Making sense of semantic ambiguity: Semantic competition in lexical access. Journal of Memory and Language pp. 245–266 (2002)
80. Rodd, J.M., Gaskell, G.M., Marslen-Wilson, W.D.: Modelling the effects of semantic ambiguity in word recognition. Cognitive Science: A Multidisciplinary Journal **28**(1), 89–104 (2004)
81. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
82. Schütze, H.: Dimensions of meaning. In: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing, Supercomputing '92, pp. 787–796. IEEE Computer Society Press, Los Alamitos, CA, USA (1992)
83. Schütze, H.: Word space. In: L.C. Giles, S.J. Hanson, J.D. Cowan (eds.) Advances in Neural Information Processing Systems 5, pp. 895–902. San Francisco, CA: Morgan Kaufmann (1993)
84. Schütze, H.: Automatic word sense discrimination. Computational Linguistics **24**(1), 97–123 (1998)
85. Sedelow, S.Y., Sedelow, W.A.J.: Thesaural knowledge representation. In: Proceedings of the University of Waterloo Conference on Lexicology, pp. 29–43. Waterloo, Ontario (1986)
86. Sedelow, S.Y., Sedelow, W.A.J.: Recent model-based and model-related studies of a large scale lexical resource [roget's thesaurusl]. In: Proceedings of the 14th Conference on Computational Linguistics - Volume 4, COLING '92, pp. 1223–1227. Association for Computational Linguistics, Stroudsburg, PA, USA (1992)
87. Simona, D.A., Lewisa, G., Marantzab, A.: Disambiguating form and lexical frequency effects in meg responses using homonyms. Language and Cognitive Processes **27**(2), 275–287 (2012)
88. Sinha, R.S., Mihalcea, R.F.: Using centrality algorithms on directed graphs for synonym expansion. In: R.C. Murray, P.M. McCarthy (eds.) FLAIRS Conference. AAAI Press (2011)
89. Small, S., Rieger, C.: Parsing and comprehending with word experts (a theory and its realization). In: W. Lehnert, M.H. Ringle (eds.) Strategies for Natural Language Processing, pp. 89–147. L. Erlbaum, Hillsdale, NJ (1982)
90. Sparck Jones, K.: Synonymy and Semantic Classification. Edinburgh University Press, Edinburgh (1964/1986)
91. Szarvas, G., Biemann, C., Gurevych, I.: Supervised all-words lexical substitution using delexicalized features. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1131–1141. Association for Computational Linguistics, Atlanta, Georgia (2013)
92. Thater, S., Fürstenau, H., Pinkal, M.: Word meaning in context: A simple and effective vector model. In: Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 1134–1143. Asian Federation of Natural Language Processing (2011)
93. Towell, G., Voorhees, E.M.: Disambiguating highly ambiguous words. Computational Linguistics **24**(1), 125–145 (1998)
94. Tsatsaronis, G., Vazirgiannis, M., Androutsopoulos, I.: Word sense disambiguation with spreading activation networks generated from thesauri. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1725–1730 (2007)
95. Uematsu, S., Kim, J.D., Tsujii, J.: Bridging the gap between domain-oriented and linguistically-oriented semantics. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp. 162–170. Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
96. Viveros-Jimnez, F., Gelbukh, A., Sidorov, G.: Simple window selection strategies for the simplified lesk algorithm for word sense disambiguation. In: F. Castro, A. Gelbukh, M. Gonzlez (eds.) Advances in Artificial Intelligence and Its Applications, *Lecture Notes in Computer Science*, vol. 8265, pp. 217–227. Springer Berlin Heidelberg (2013)
97. Waltz, D.L., Pollack, J.B.: Massively parallel parsing: A strongly interactive model of natural language interpretation. Cognitive Science **9**, 51–74 (1985)

98. Weaver, W.: Translation. In: W.N. Locke, A.D. Boothe (eds.) Machine Translation of Languages, pp. 15–23. MIT Press, Cambridge, MA (1949; rpt. 1955). Reprinted from a memorandum written by Weaver in 1949.
99. Weiss, S.: Learning to disambiguate. Information Storage and Retrieval **9**(1), 33–41 (1973)
100. Widdows, D., Dorow, B.: A graph model for unsupervised lexical acquisition. In: Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, pp. 1–7. Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
101. Wilks, Y.: Grammar, Meaning and the Machine Analysis of Language. Routledge and Kegan Paul, London (1972)
102. Wilks, Y.: Primitives and words. In: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing, TINLAP '75, pp. 38–41. Association for Computational Linguistics, Stroudsburg, PA, USA (1975)
103. Wilks, Y., Slator, B.: Towards semantic structures from dictionary entries. In: Proceedings of the 2nd Annual Rocky Mountain Conference on Artificial Intelligence, pp. 85–96. Boulder, Colorado (1989)
104. Wilks, Y., Slator, B., Guthrie, L.: Electric Words: Dictionaries, Computers, and Meanings. MIT Press (1996)
105. Wilks, Y.A.: Preference semantics. In: E.L.I. Keenan (ed.) Formal Semantics of Natural Language, pp. 329–348. Cambridge University Press (1975)
106. Williams, J.N.: Processing polysemous words in context: Evidence for interrelated meanings. Journal of Psycholinguistic Research **21**(3), 193–218 (1992)
107. Yarowsky, D.: Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In: Proceedings of the 14th Conference on Computational Linguistics - Volume 2, COLING '92, pp. 454–460. Association for Computational Linguistics, Stroudsburg, PA, USA (1992)
108. Yarowsky, D.: One sense per collocation. In: Proceedings of the Workshop on Human Language Technology, HLT '93, pp. 266–271. Association for Computational Linguistics, Stroudsburg, PA, USA (1993)
109. Yarowsky, D.: Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, ACL '94, pp. 88–95. Association for Computational Linguistics, Stroudsburg, PA, USA (1994)
110. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, ACL '95, pp. 189–196. Association for Computational Linguistics, Stroudsburg, PA, USA (1995)