Krovetz, R. (1990). Viewing the dictionary as a classification system. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 87-94. doi: 10.7152/acro.v1i1.12467

Viewing the Dictionary as a Classification System

Robert Krovetz

Computer and Information Science Department
University of Massachusetts, Amherst, MA 01003, USA

INTRODUCTION

Information retrieval is one of the earliest applications of computers. Starting with the speculative work of Vannevar Bush on Memex [Bush 45], to the development of Key Word in Context (KWIC) indexing by H.P. Luhn [Luhn 60] and Boolean retrieval by John Horty [Horty 62], to the statistical techniques for automatic indexing and document retrieval done in the 1960's and continuing to the present [Salton and McGill 83], Information Retrieval has continued to develop and progress. However, there is a growing consensus that current generation statistical techniques have gone about as far as they can go, and that further improvement requires the use of natural language processing and knowledge representation. We believe that the best place to start is by focusing on the lexicon, and to index documents not by words, but by word *senses*.

Why use word senses? Conventional approaches advocate either indexing by the words themselves, or by manual indexing using a controlled vocabulary. Manual indexing offers some of the advantage of word senses, in that the terms are not ambiguous, but it suffers from problems of consistency. In addition, as text data bases continue to grow, it will only be possible to index a fraction of them by hand.

In advocating word senses as indices we are not suggesting that they are the ultimate answer. There is much more to the meaning of a document then the senses of the words it contains; we are just saying that senses are a good start. Any approach to providing a semantic analysis must deal with the problem of word meaning. Existing retrieval systems try to go beyond single words by using a thesaurus,[1] but this has the problem that words are not synonymous in all contexts. The word 'term' may be synonymous with 'word' (as in a vocabulary term), 'sentence' (as in a prison term), or 'condition' (as in 'terms of agreement'). If we expand the query with words from a thesaurus, we must be careful to use the right senses of those words. We not only have to know the sense of the word in the query (in this example, the sense of the word 'term'), but the sense of the word that is being used to augment it (e.g., the appropriate sense of the word 'sentence'). The thesaurus we use should be one in which the senses of words are explicitly indicated [Chodorow et al. 88].

We contend that the best place to obtain word senses is a machine-readable dictionary. Although it is possible that another list of senses might be manually constructed, this strategy might cause some senses to be overlooked, and the task will entail a great degree of effort.

DICTIONARY AS SEMANTIC CLASSIFICATIONS

In our view, the dictionary can be thought of as a classification system. As mentioned above, it has similarities to controlled vocabulary systems such as *MeSH (Medical Subject Headings)*

1. By *thesaurus* we do not mean a controlled vocabulary system, but a resource akin to *Roget's Thesaurus*.

ISSN: 2324-9773

Krovetz, R. (1990). Viewing the dictionary as a classification system. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 87-94. doi: 10.7152/acro.v1i1.12467

and the *West* keynotes system (which is used to index legal text). Both dictionaries and controlled vocabularies consist of an unambiguous collection of 'terms' (word senses in the case of a dictionary) which can be used to represent the content of a document. An important distinction is that the terms in a controlled vocabulary system must be assigned to a document by an indexer. If we use the word senses that make up a document as representations of its content, this problem does not exist. However, these senses are not known *a priori*. We need to determine how to identify the sense of a word given the context in which the word appears. This has been the focus of our work, and we will summarize it in the rest of this paper. First we will provide a bit more detail about dictionaries as classifications.

Dictionary definitions are usually given according to a 'lexicographic tradition' - nouns are defined as noun phrases and verbs are defined as verb phrases. The head of each phrase forms a taxonomic link to a higher level term. For example, a 'car' might be defined as 'a vehicle with 4 wheels usually used for transporting people'. The head of the phrase is the word 'vehicle' and this can be used to create a taxonomic link with the word 'car'. In addition, the definition contains differentia that serve to distinguish the term from the genus. In this example the differentia specifies the number of wheels, and the fact that the purpose of the vehicle is transporting people. Finally, we note that the definition may provide typicality information via the words 'usually' or 'especially'.

The structure of definitions has a great deal in common with frame-based systems used in Artificial Intelligence. Word senses and frames are both used to describe concepts, and they are both organized according to an inheritance hierarchy with specializing information used to distinguish one level from the next. An important distinction is that the dictionary does not indicate the sense of the genus term (e.g., the particular sense of 'vehicle' used in the definition of 'car'). Furthermore, frame-based systems explicitly name the attributes used to distinguish frames in the hierarchy. The differentia in definitions are indicated with prepositional phrases, relative clauses, and adjectival and adverbial modifiers; the particular relations must be identified by external knowledge, such as the fact that 'grey' is a color, or that 'used for' describes a purpose.

The differentia in dictionaries have not yet received much study. Some work has been done by Alshawi for the definitions in the *Longman Dictionary of Contemporary English (LDOCE)*. A partial analysis was made of the differentia, and these were extracted into a simple semantic structure [Alshawi 87]. Amsler analyzed the differentia for a subset of the verbs of motion in the *Merriam-Webster Pocket Dictionary* as part of his doctoral dissertation research [Amsler 80]. Amsler also studied the dictionary's taxonomic organization. Each genus term in the noun and verb hierarchy was manually disambiguated, and the overall structure stored in a database system. Amsler studied the entire taxonomy of the word 'vehicle' (which was extremely well organized), and 'group', which was found to form a 'tangled hierarchy' (a taxonomy in which a node can have more than one parent). Dictionary taxonomies have also been explored by Chodorow, who used automatic methods for identifying the genus term and semi-automatic methods for producing the taxonomies [Chodorow et al. 85].

As with any classification system, dictionaries need to be updated. Sometimes a word will acquire a new sense in a technical field, and new words and senses are always being added to the

ISSN: 2324-9773

Krovetz, R. (1990). Viewing the dictionary as a classification system. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 87-94. doi: 10.7152/acro.v1i1.12467

general vocabulary. Various strategies can be used to identify new word senses, but we do not have the space to describe them in any detail. The central notion is that dictionaries provide a framework for determining what information needs to be acquired; they contain information about the word's senses such as part-of-speech, subcategorization,[2] semantic restrictions, subject area, and other words that co-occur. If a word in context is being used in a different sense, that fact can be detected by a deviation from the information associated with the senses that are already in the dictionary. New word senses can also be detected if the word occurs with a frequency radically different from what is expected. For more information see [Krovetz 91].

The main problem with using dictionaries as classification systems is that the words in the documents need to be disambiguated. We also need to show that word senses are useful categories - that they provide good discrimination between relevant and non-relevant documents. This has been the focus of our efforts. Our hypothesis is that when a word sense in a query does not match the sense of the word in the document, then that document is not likely to be relevant with respect to that word. The results reported in [Krovetz and Croft 89] provide a preliminary indication that this hypothesis is correct.

The rest of this paper will provide a brief review of our work on lexical ambiguity and information retrieval.

OVERVIEW OF PREVIOUS WORK

Lexical ambiguity is a pervasive problem in natural language processing. Although work has been done on disambiguation, most of it has focused on a small number of words and a restricted set of senses. In general, the lexicons used in natural language systems have been very small. Winograd's lexicon for his blocks world program, SHRDLU, contained only 217 words (Winograd, personal communication). The lexicon for MARGIE, one of the early systems developed by Roger Schank, contained only 60 verbs and a smattering of other words to go with them (Riesbeck, personal communication). An informal poll held at a recent workshop [Boguraev, 1986 Alvey/ICL Workshop on Linguistic Theory and Computer Applications] revealed that the average lexicon size used was 1500 words, and only 250 words once a large machine translation system was excluded [Ritchie 87]. The small size of the lexicons used in natural language processing is one of the main reasons for their failure to be robust. Lack of vocabulary was cited as one of the main causes of failure in FRUMP, a system designed for sketchy parsing of real world text [DeJong 82].

Within the field of Information Retrieval the lexicon has been virtually ignored. Sometimes it is expressed as a list of synonyms and phrases. Often it is only treated in a negative sense via a 'stop list' (a list of words not considered useful for indexing).

Our initial work focused on determining the degree of ambiguity found in information retrieval test collections. This was done with respect to a machine-readable version of *LDOCE*. We found that the words in the test collections have an average of 5 senses. Our experiments were done with two test collections, one consisting of titles and abstracts from *Communications of the ACM (CACM)*, and the other consisting of short articles from *TIME* magazine. Because

2. Distinctions within a grammatical category, such as the difference between transitive and intransitive verbs.

ISSN: 2324-9773

Krovetz, R. (1990). Viewing the dictionary as a classification system. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 87-94. doi: 10.7152/acro.v1i1.12467

CACM is technical text, we expected it to contain many words that were domain-specific. Instead we found that both collections contained about the same percentage of words in the dictionary (40 percent without considering inflectional morphology, and 57 to 65 percent afterwards). This is due in part to a higher percentage of proper nouns in the *TIME* collection, and partially to the fact that some of the words in the *CACM* collection are in the dictionary but used with a domain-specific sense. We then conducted an experiment involving weighting of words by the number of senses they have. This was done because the number of senses a word has is highly correlated with its frequency, and it is known that weighting by frequency produces better results. We found that sense weighting did as well as weighting by frequency in one of the test collections we used (*TIME*), but generated worse results for the other collection (*CACM*). An analysis of these results showed that the *CACM* collection contained a number of low frequency words that were highly ambiguous. One of the reasons for this was an overlap between technical and general word meanings, which led to the observation that an anomalous distribution can be useful in detecting domain-specific word senses.

Following the weighting experiment, we conducted an experiment to directly investigate our hypothesis that word meanings can be useful in separating relevant from non-relevant documents. We first manually identified the meanings of the words used in the queries for the test collections. We then examined the words from the queries that occurred in the top ten ranked documents for each query and determined how often the senses matched. The results of this comparison were that sense mismatches are far more likely to occur in a non-relevant document than in a relevant one. However, word senses matched about ninety percent of the time (i.e., only 10 percent of the time did the word in a query have a different sense from the word in the document). Thus it isn't clear whether word sense disambiguation will directly make a significant improvement in performance. Since the top ranked documents have the most words in common with the query, we did some examination of the documents that only shared one word in common in the hope that mismatches would occur more often. These results were also inconclusive (a discussion of them is given in [Krovetz and Croft 90]).

We are in the process of examining the distribution of the senses for the words in the queries and of the senses for those words in the document collections. The preliminary results are that the high percentage of matches was caused by very uneven sense distribution and the sublanguage used in both collections (computer science and politics). The senses in *LDOCE* are generally ordered by frequency, and for the *CACM* collection the first sense of a word from the queries occurs about forty-five percent of the time; fourteen percent of the words only have one sense, so this gives empirical evidence that the first sense is more frequent. Preliminary analysis of the document collection indicates that many words are used in only one or two senses.

Although the words in the test collections are potentially very ambiguous with respect to the dictionary, they appear to be used with relatively few senses in practice. Nevertheless, there are several reasons why we believe that disambiguation is worthwhile. First, the test collections we used are both on particular subject areas; we expect that with other text databases, such as patents or dissertation abstracts, ambiguity will be more of a problem. Second, the words in the queries were matched against the words in the text via a process called "stemming" (essentially truncation of the word endings). This process does not capture all of the variants a word can have, and thus some documents will not be retrieved due to a failure to match on a variant (for

ISSN: 2324-9773

Krovetz, R. (1990). Viewing the dictionary as a classification system. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 87-94. doi: 10.7152/acro.v1i1.12467

example, 'actor' will not match 'act' or 'actress'). Such variants are based not on the word, but on the *sense* of the word. Third, a query often does not contain all of the words that might be used to find relevant documents. Word senses could serve as a basis for inferring related words that would then be added to the query. For example, 'order' might be used to infer 'arrangement', but not if it's used in the sense of ordering food (we also want to include only the right sense of 'arrangement'). Finally, senses of words is only one factor affecting relevance. The *relationships* that those words have to one another is also important. Determining these relationships is likely to require the use of a natural language parser, and knowing the senses in which the words are used serves as an important constraint on the parse. Although the words may only be used with a small number of senses (relative to the number they have in the dictionary), we do not know in advance which *particular* senses will be used within a given collection of text. Word sense disambiguation is also important in other areas of natural language processing such as machine translation and text critiquing.

Our approach to disambiguation is based on treating the information associated with dictionary senses (part-of-speech, subcategorization, semantic restrictions, etc.) as sources of evidence. We will be investigating how well each source discriminates senses, how well it can be identified with a word in context, and how much improvement it makes in retrieval system performance. We will first investigate these sources in isolation, and then see how much improvement can be gained by consensus. We don't consider the dictionary a panacea for using word senses; there will certainly be words and word senses that won't be found in the dictionary, and the dictionary will sometimes make distinctions that are too fine-grained. These problems and more details about our approach are discussed further in [Krovetz 91].

At the moment it isn't clear what kind of information will prove most useful for disambiguation. In particular it isn't clear what kinds of knowledge will be required that are not contained in the dictionary. In the sentence 'John left a tip', the word 'tip' might mean a gratuity or a piece of advice. Cullingford and Pazzani cite this an example in which scripts[3] are needed for disambiguation [Cullingford and Pazzani 84]. However, it isn't clear how often such a case occurs, how many scripts would be involved, or how much effort is required to construct them. In addition, we might be able to do just as well via the use of word co-occurrences (the gratuity sense of tip is likely to occur in the same context as 'restaurant', 'waiter', 'menu', etc.). In other words, we might be able to use the words that could trigger a script without actually making use of one. This also raises the question of how much context is required. Part of speech can be effectively determined with a context of only two words [Church 88]. A more global approach is taken by Slator, who disambiguates word senses based on subject codes associated with them in *LDOCE* [Slator 89]. This is done for all words in the text being disambiguated. However it isn't clear how the size of the text influences the effectiveness of disambiguation, or how well it works compared with other kinds of information contained in the dictionary.

It is essential to consider the costs involved in constructing sophisticated representation systems. *LDOCE* provides the most information associated with word senses of any machine-readable dictionary; we believe it is worthwhile to see how far this information goes toward

3. A script is a data structure used in Artificial Intelligence that indicates the stereotypical participants in an event and the temporal ordering of the actions of those participants.

ISSN: 2324-9773

Krovetz, R. (1990). Viewing the dictionary as a classification system. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 87-94. doi: 10.7152/acro.v1i1.12467

disambiguation. An assessment of the cases in which this information fails to distinguish senses will shed light on the additional information that is required. It is also important to consider the use which will be made of the senses. A machine translation system might require finer-grained distinctions than an information retrieval system; distinguishing these senses could entail additional information.

CONCLUSION

This paper has reviewed the work we have done so far on lexical ambiguity and information retrieval. We believe documents should be indexed not by words, but by word senses. At the moment it isn't clear how much direct improvement can be expected by such indexing. However, word senses are an essential component of better natural language analysis. They are needed for more accurate parsing and inference generation, and offer promise with respect to a sense-disambiguated thesaurus. We also expect that our work will help in other areas of natural language processing by means of sense-based morphological analysis and by providing frequency information about sense distribution. We are currently redoing our sense weighting experiment in light of the information obtained from the matching experiment (the distribution of senses for words in the queries). Final results about the outcome of indexing *per se* will have to wait until a disambiguation system is implemented.

ACKNOWLEDGMENTS

This work has been supported by the Office of Naval Research under University Research Initiative Grant N00014-86-K-0746, by the Rome Air Development Center and the Air Force Office of Scientific Research, under contract F30602-85-C-008, and by NSF Grant IRI-8814790. We wish to thank Longman Group, Ltd. for making the machine-readable version of the *Longman Dictionary* available to us. The work is being done as part of the author's dissertation research under the direction of Professor Bruce Croft.

REFERENCES

[Alshawi 87] Alshawi H., "Processing Dictionary Definitions with Phrasal Pattern Hierarchies", *Computational Linguistics*, Vol. 13, No. 3-4, pp. 195-202, 1987.

[Amsler 80] Amsler R., "The Structure of the Merriam Webster Pocket Dictionary", PhD Dissertation, University of Texas at Austin, 1980.

[Bush 45] Bush V., "As We May Think", *Atlantic Monthly*, Vol. 175, pp. 101-108, July, 1945.

[Chodorow et al. 85] Chodorow M., Byrd R., and Heidorn G., "Extracting Semantic Hierarchies from a Large On-Line Dictionary , *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pp. 299-304, 1985.

ISSN: 2324-9773

Krovetz, R. (1990). Viewing the dictionary as a classification system. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 87-94. doi: 10.7152/acro.v1i1.12467

[Chodorow Chodorow M., Ravin Y., and Sachar H., Tool for Investigating the Synonymy Relation in a Sense Disambiguated Thesaurus", *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pp. 144-151, 1988.

[Church 88] Church K., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", in *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pp. 136-143, 1988.

[Cullingford and Pazzani 84] Cullingford R. and Pazzani M., "Word-Meaning Selection in Multiprocess Language Understanding Programs", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 4, pp. 493-509, 1984.

[DeJong 82] DeJong G., "An Overview of the FRUMP System", in *Strategies for Natural Language Processing*, Lehnert and Ringle (eds), Lawrence Erlbaum Press, pp. 149-176, 1982.

[Horty 62] Horty J., "The 'Key Word in Combination' Approach", in *Modern Uses of Logic in Law*, No. 54, 1962.

[Luhn 60] Luhn H.P. "Keyword-in-Context Index for Technical Literature", *American Documentation*, Vol. 11, No. 4, pp. 299-295, 1960.

[Krovetz 91] Krovetz R., "Lexical Acquisition and Information Retrieval", in *Lexical Acquisition: Building the Lexicon Using On-Line Resources*, U. Zernik (ed), LEA Press, 1991.

[Krovetz and Croft 89] Krovetz R. and Croft W.B., "Word Sense Disambiguation Using Machine Readable Dictionaries", in *Proceedings of the Conference on Research and Development in Information Retrieval*, pp. 127-136, 1989.

[Krovetz and Croft 90] Krovetz R. and Croft W.B. "Lexical Ambiguity and Information Retrieval", *Information Retrieval Lab memo IRL-90-5*, Department of Computer Science, University of Massachusetts at Amherst.

[Ritchie 87] Ritchie G., "Discussion Session on the Lexicon", in *Linguistic Theory and Computer Applications*, Whitelock et al. (eds), Academic Press, 1987.

[Salton and McGill 83] Salton G. and McGill M., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[Slator 89] Slator B., "Using Context for Sense Preference", in *Proceedings of the First International Workshop on Lexical Acquisition*, 1989.

ISSN: 2324-9773

Krovetz, R. (1990). Viewing the dictionary as a classification system. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 87-94. doi: 10.7152/acro.v1i1.12467

ISSN: 2324-9773