# Sense-Linking in a Machine Readable Dictionary

**Robert Krovetz**
Department of Computer Science
University of Massachusetts, Amherst, MA 01003

## Abstract

Dictionaries contain a rich set of relationships between their senses, but often these relationships are only implicit. We report on our experiments to automatically identify links between the senses in a machine-readable dictionary. In particular, we automatically identify instances of zero-affix morphology, and use that information to find specific linkages between senses. This work has provided insight into the performance of a stochastic tagger.

## 1 Introduction

Machine-readable dictionaries contain a rich set of relationships between their senses, and indicate them in a variety of ways. Sometimes the relationship is provided explicitly, such as with a synonym or antonym reference. More commonly the relationship is only implicit, and needs to be uncovered through outside mechanisms. This paper describes our efforts at identifying these links.

The purpose of the research is to obtain a better understanding of the relationships between word meanings, and to provide data for our work on word-sense disambiguation and information retrieval. Our hypothesis is that retrieving documents on the basis of word senses (instead of words) will result in better performance. Our approach is to treat the information associated with dictionary senses (part of speech, subcategorization, subject area codes, etc.) as multiple sources of evidence (cf. Krovetz [3]). This process is fundamentally a divisive one, and each of the sources of evidence has exceptions (i.e., instances in which senses are *related* in spite of being separated by part of speech, subcategorization, or morphology). Identifying related senses will help us to test the hypothesis that unrelated meanings will be more effective at separating relevant from nonrelevant documents than meanings which are related.

We will first discuss some of the explicit indications of sense relationships as found in usage notes and deictic references. We will then describe our efforts at uncovering the implicit relationships via stochastic tagging and word collocation.

## 2 Explicit Sense Links

The dictionary we are using in our research, the Longman Dictionary of Contemporary English (LDOCE), is a dictionary for learners of English as a second language. As such, it provides a great deal of information about word meanings in the form of example sentences, usage notes, and grammar codes. The Longman dictionary is also unique among learner's dictionaries in that its definitions are generally written using a controlled vocabulary of approximately 2200 words. When exceptions occur they are indicated by means of a different font. For example, consider the definition of the word *gravity*:

- **gravity** n 1b. worrying importance: *He doesn't understand the gravity of his illness* - see GRAVE[2]
- **grave** adj 2. important and needing attention and (often) worrying: *This is grave news — The sick man's condition is grave*

These definitions serve to illustrate how words can be synonymous[1] even though they have different parts of speech. They also indicate how the Longman dictionary not only indicates that a word is a synonym, but sometimes specifies the *sense* of that word (indicated in this example by the superscript following the word 'GRAVE'). This is extremely important because synonymy is not a relation that holds between words, but between the *senses* of words.

Unfortunately these explicit sense indications are not always consistently provided. For example, the definition of 'marbled' provides an explicit indication of the appropriate sense of 'marble' (the stone instead of the child's toy), but this is not done within the definition of 'marbles'.

LDOCE also provides explicit indications of sense relationships via usage notes. For example, the definition for *argument* mentions that it derives from both senses of *argue* - to quarrel (to have an argument), and to reason (to present an argument). The notes also provide advice regarding similar looking variants (e.g., the difference between *distinct* and *distinctive*, or the fact that an *attendant* is not someone who *attends* a play, concert, or religious service). Usage notes can also specify information that is shared among some word meanings, but not others (e.g., the note for *venture* mentions that both verb and noun carry a connotation of risk, but this isn't necessarily true for *adventure*).

Finally, LDOCE provides explicit connections between senses via deictic reference (links created by

[1] We take two words to be synonymous if they have the same or closely related meanings.

'this', 'these', 'that', 'those', 'its', 'itself', and 'such a/an'). That is, some of the senses use these words to refer to a previous sense (e.g., 'the fruit of this tree', or 'a plant bearing these seeds'). These relationships are important because they allow us to get a better understanding of the nature of polysemy (related word meanings). Most of the literature on polysemy only provides anecdotal examples; it usually does not provide information about how to determine whether word meanings are related, what kind of relationships there are, or how frequently they occur. The grouping of senses in a dictionary is generally based on part of speech and etymology, but part of speech is orthogonal to a semantic relationship (cf. Krovetz [3]), and word senses can be related etymologically, but be perceived as distinct at the present time (e.g., the 'cardinal' of a church and 'cardinal' numbers are etymologically related). By examining deictic reference we gain a better understanding of senses that are truly related, and it also helps us to understand how language can be used creatively (i.e., how senses can be productively extended). Deictic references are also important in the design of an algorithm for word-sense disambiguation (e.g., exceptions to subcategorization).

The primary relations we have identified so far are: substance/product (tree:fruit or wood, plant:flower or seeds), substance/color (jade, amber, rust), object/shape (pyramid, globe, lozenge), animal/food (chicken, lamb, tuna), count-noun/mass-noun,[2] language/people (English, Spanish, Dutch), animal/skin or fur (crocodile, beaver, rabbit), and music/dance (waltz, conga, tango).[3]

## 3 Zero-Affix Morphology

Deictic reference provides us with different types of relationships within the same part of speech. We can also get related senses that differ in part of speech, and these are referred to as instances of zero-affix morphology or functional shift. The Longman dictionary explicitly indicates some of these relationships by homographs that have more than one part of speech. It usually provides an indication of the relationship by a leading parenthesized expression. For example, the word *bay* is defined as N,ADJ, and the definition reads '(a horse whose color is) reddish-brown'. However, out of the 41122 homographs defined, there are only 695 that have more than one part of speech. Another way in which LDOCE provides these links is by an explicit sense reference for a word outside the controlled vocabulary; the definition of *anchor* (v) reads: 'to lower an **anchor**[1] (1) to keep (a ship) from moving'. This indicates a reference to sense 1 of the first homograph.

Zero-affix morphology is also present implicitly, and we conducted an experiment to try to identify instances of it using a probabilistic tagger [2]. The hypothesis is that if the word that's being defined (the definiendum) occurs within the text of its own definition, but occurs with a different part of speech, then it will be an instance of zero-affix morphology. The question is: How do we tell whether or not we have an instance of zero-affix morphology when there is no explicit indication of a suffix? Part of the answer is to rely on subjective judgment, but we can also support these judgments by making an analogy with derivational morphology. For example, the word *wad* is defined as 'to make a **wad** of'. That is, the noun bears the semantic relation of *formation* to the verb that defines it. This is similar to the effect that the morpheme *-ize* has on the noun *union* in order to make the verb *unionize* (cf. Marchand [5]).

The experiment not only gives us insight into semantic relatedness across part of speech, it also enabled us to determine the effectiveness of tagging. We initially examined the results of the tagger on all words starting with the letter 'W'; this letter was chosen because it provided a sufficient number of words for examination, but wasn't so small as to be trivial. There were a total of 1141 words that were processed, which amounted to 1309 homographs and 2471 word senses; of these senses, 209 were identified by the tagger as containing the definiendum with a different part of speech. We analyzed these instances and the result was that only 51 of the 209 instances were found to be correct (i.e., actual zero-morphs).

The instances that are indicated as correct are currently based on our subjective judgment; we are in the process of examining them to identify the type of semantic relation and any analog to a derivational suffix. The instances that were not found to be correct (76 percent of the total) were due to incorrect tagging; that is, we had a large number of false positives because the tagger did not correctly identify the part of speech. We were surprised that the number of incorrect tags was so high given the performance figures cited in the literature (more than a 90 percent accuracy rate). However, the figures reported in the literature were based on word tokens, and 60 percent of all word tokens have only one part of speech to begin with. We feel that the performance figures should be supplemented with the tagger's performance on word types as well. Most word types are rare, and the stochastic methods do not perform as well on them because they do not have sufficient information. Church has plans for improving the smoothing algorithms used in his tagger, and this would help on these low frequency words. In addition, we conducted a failure analysis and it indicated that 91% the errors occurred in idiomatic

[2]These may or may not be related; consider 'computer vision' vs. 'visions of computers'. The related senses are usually indicated by the defining formula: 'an example of this'.

[3]The related senses are sometimes merged into one; for example, the definition of *foxtrot* is '(a piece of music for) a type of formal dance...'

expressions (45 instances) or example sentences (98 instances). We therefore eliminated these from further processing and tagged the rest of the dictionary. We are still in the process of analyzing these results.

## 4 Derivational Morphology

Word collocation is one method that has been proposed as a means for identifying word meanings. The basic idea is to take two words in context, and find the definitions that have the most words in common. This strategy was tried by Lesk using the Oxford Advanced Learner's Dictionary [4]. For example, the word 'pine' can have two senses: a tree, or sadness (as in 'pine away'), and the word 'cone' may be a geometric structure, or a fruit of a tree. Lesk's program computes the overlap between the senses of 'pine' and 'cone', and finds that the senses meaning 'tree' and 'fruit of a tree' have the most words in common. Lesk gives a success rate of fifty to seventy percent in disambiguating the words over a small collection of text. Later work by Becker on the New OED indicated that Lesk's algorithm did not perform as well as expected [1].

The difficulty with the word overlap approach is that a wide range of vocabulary can be used in defining a word's meaning. It is possible that we will be more likely to have an overlap in a dictionary with a restricted defining vocabulary. When the senses to be matched are further restricted to be morphological variants, the approach seems to work very well. For example, consider the definitions of the word 'appreciate' and 'appreciation':

- **appreciate**
  1. to be thankful or grateful for
  2. to understand and enjoy the good qualities of
  3. to understand fully
  4. to understand the high worth of
  5. (of property, possessions, etc.) to increase in value
- **appreciation**
  1. judgment, as of the quality, worth, or facts of something
  2. a written account of the worth of something
  3. understanding of the qualities or worth of something
  4. grateful feelings
  5. rise in value, esp. of land or possessions

The word overlap approach pairs up sense 1 with sense 4 (grateful), sense 2 with sense 3 (understand; qualities), sense 3 with sense 3 (understand), sense 4 with sense 1 (worth), and sense 5 with sense 5 (value; possessions). The matcher we are using ignores closed class words, and makes use of a simple morphological analyzer (for inflectional morphology). It ignores words found in example sentences (preliminary experiments indicated that this didn't help and sometimes made matches worse), and it also ignores typographical codes and usage labels (formal/informal, poetic, literary, etc.). It also doesn't try to make matches between word senses that are idiomatic (these are identified by font codes). We are currently in the process of determining the effectiveness of the approach. The experiment involves comparing the morphological variations for a set of queries used in an information retrieval test collection. We have manually identified all variations of the words in the queries as well as the root forms. Those variants that appear in LDOCE will be compared against all root forms and the result will be examined to see how well the overlap method was able to identify the correct sense of the variant with the correct sense of the root.

## 5 Conclusion

The purpose of this work is to gain a better understanding of the relationships between word meanings, and to help in development of an algorithm for word sense disambiguation. Our approach is based on treating the information associated with dictionary senses (part of speech, subcategorization, subject area codes, etc.) as multiple sources of evidence (cf. Krovetz [3]). This process is fundamentally a divisive one, and each of the sources of evidence has exceptions (i.e., instances in which senses are *related* in spite of being separated by part of speech, subcategorization, or morphology). Identifying the relationships we have described will help us to determine these exceptions.

## References


[1] Becker B., "Sense Disambiguation using the *New Oxford English Dictionary*", Masters Thesis, University of Waterloo, 1989.

[2] 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.

[3] Krovetz R., "Lexical Acquisition and Information Retrieval", in *Lexical Acquisition: Building the Lexicon Using On-Line Resources*, U. Zernik (ed), pp. 45–64, 1991.

[4] Lesk M., "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a Pine Cone from an Ice Cream Cone", *Proceedings of SIGDOC*, pp. 24–26, 1986.

[5] Marchand H, "On a Question of Contrary Analysis with Derivational Connected but Morphologically Uncharacterized Words", *English Studies*, 44, pp. 176-187, 1963