Polysemy and Category Structure in WordNet: An Evidential Approach

Tony Veale

Department of Computer Science,

University College Dublin, Ireland.

Tony.Veale@UCD.ie

Abstract

While polysemy is a form of ambiguity that can complicate natural language processing, it is also a rich lexical resource that yields useful insights into the mapping between words and concepts. WordNet, a comprehensive lexical knowledge-base of English word meanings, is replete with instances of polysemy, but also contains many instances of homonymy, and fails to distinguish between both kinds of ambiguity. We propose in this paper an alternative to the distributional approach for recognizing polysemous sense-pairs in WordNet. Our approach does not rely on the systematicity of regular polysemy to identify the families of words that instantiate a particular metonymic pattern, but seeks instead local ontological evidence for each word, on a case by case basis.

1. Introduction

Ambiguity is such a vexing problem in natural language processing (Ravin and Leacock, 2000) that it is easy to forget that, like cholesterol, lexical ambiguity comes in both a good form and a bad form. Homonymy, the bad form, is ambiguity arising from historical coincidences of language that do not follow any predictable conceptual patterns, and which generally serve no useful purpose beyond the generation of puns. The study of homonymy may illuminate in some small way the diachronic development of language but sheds no light at all on its conceptual underpinnings. Polysemy, on the other hand, arises when two or more related meanings are shoehorned into the same lexical form for reasons of linguistic economy or creativity. Polysemy is thus a good form of lexical ambiguity. It's presence, if detected, can reveal the workings of a systematic conceptual trend at work (Apresjan, 1974), or a relational similarity between senses that has not been explicitly marked in the lexicon.

The key phrase here is "if detected". WordNet, a comprehensive ontologically-structured lexicon of English (Miller, 1995), is rich in instances of polysemy, but it is also home to many instances of homonymy and does not explicitly differentiate one from the other. This deficit can lead to false rationalizations and silly inferences. The fact that the word "bank" has both a river-side sense and a financial institution sense does not mean that the latter is to be found on the former. However, the fact that "bank" has a building sense and a financial institution sense does mean that the latter is to be found in the former. The building/financial-institution ambiguity of "bank" is an example of polysemy that instantiates a general conceptual tendency to conflate organizations with their locations (think "Whitehouse" and "Wall Street"), while the building/river-side ambiguity is simply a case of homonymy (each use of "bank" having a different etymological origin).

NLP systems must be able to distinguish polysemy from homonymy since each demands a different resolution mechanism. Unlike homonymy, where a single choice of senses must be made, polysemous words can be used in multiple different senses simultaneously (Pustejovsky, 1991; Cruse, 2002). For example, the sentence "the book was badly written but beautifully produced" requires two related senses of "book" to be co-active – "book" as a container of abstract content ("badly written") and as a physical artifact ("beautifully produced"). Since polysemy is also a (semi-)productive phenomenon, it allows an NLP system to dynamically plug the holes in its lexicon. Most dictionaries list "cod" and "haddock" as both a fish and a food, but few would bother to also list "shark" as a food. Yet "shark soup" requires a system to understand that sharks too are edible fish. The basis for this productivity lies in the existing polysemy patterns of language; sharks are fish, but "fish" denotes both a marine animal and the food derived from it.

In this paper we present an ontological basis for detecting polysemy patterns in WordNet. The approach differs from past work by working not at the general level of distributional analysis and word families as a whole, but at the specific level of individual words, seeking local ontological evidence for each instance of polysemy that it hypothesizes. This significantly reduces the possibility of falsely identifying homonymy as polysemy, while in many cases revealing, in relationally specific terms, the conceptual motivation of the polysemy.

2 Past Work

Polysemy is pervasive in the lexicon because it is, to a large extent, a productive phenomenon that arises from the conflation of general categories like Location, Person, Animal, Food, and so on. This productivity means that a particular polysemy pattern may be instantiated by a substantial family of different words. For instance, WordNet contains 344 words that can denote both a type of person and a language, reflecting our tendency to name languages after the peoples that speak them. Furthermore, the same polysemy pattern may be observed at different ontological levels – WordNet contains 158 words that denote both a grouping of American Indians and the language spoken by that grouping. This potential for polysemy to be regular and systematic has been expressed by (Apresjan, 1974) thus: "Polysemy of the word A with the meanings a_i and a_j is called regular if, in the language, there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way".

This formulation by Aprejan is sufficiently algorithmic to yield a means of detecting polysemy via its systematic effect on the lexicon. This insight is applied to WordNet in (Peters and Peters, 2000), where the potential polysemy of two word senses is categorized in terms of their divergent positions in the WordNet ontology. For instance, "waltz" has a sense that lies under the hypernym {music} and another that lies under {dancing, dance, terpsichore}. Because this divergence is precisely mirrored by other words such as "Samba", "Rumba" and "Tango", it is possible to locate a sizeable family of words (23 in WordNet 1.6) that can denote both a kind of dancing and the kind of music that accompanies it. The larger the family of words that can be found to support a given divergence pattern, the more evidence there is to assume that the pattern captures a systematic tendency and that the ambiguity in each case is the result of polysemy rather than homonymy.

This distributional approach, prefigured by (Apresjan, 1974) and principally elaborated by (Peters and Peters, 2000), provides an excellent means of finding significant tendencies toward polysemy in the lexicon. But in exploiting the systematicity of regular polysemy, the distributional approach is prone to three forms of error. First, while it can reliably identify patterns like Animal/Food that give rise to word families like "turkey", "lamb", "chicken", "hen" and so on, it cannot reliably exclude homonyms from these families. For instance, "mate" the berry-tea drink is not derived from the animal sense of "mate", despite the fact that a significant 193 words instantiate the Animal/Food pattern. Secondly, the distributional approach ignores the fact that polysemy patterns are not transitive. For instance, WordNet defines several animal senses for "hen", the primary sense describing an adult female chicken and another, extended sense describing the female of certain aquatic animals such as lobsters and octopuses. However, the food sense of "hen" is a metonymic extension of the chicken sense only, and should not be used to denote the food obtained from lobsters and octopuses. Thirdly, some of the most interesting polysemy is, or at least appears to be, ad-hoc. WordNet only contains one word that denotes both a place of business and a person, "florist", yet this is a very useful polysemy pattern for an NLP system to comprehend. When one says "I went to the dentist", the intended meaning is more specific than "I went to the location of the dentist", which can allow odd interpretations like "I went to the restaurant where the dentist was eating". Tight metonymic connections like that between places of business and businessmen are worth extracting from WordNet even if they happen to be under-represented and thus appear ad-hoc.

Systematic tendencies toward polysemy can be represented in an ontology like WordNet quite effectively, by connecting the general categories involved with lexical rules. For instance, a connection between {animal} and {food} can be used to imply that any word that has both an animal sense and a food sense is a product of this polysemous tendency. This is the approach employed by the WordNet *cousins* mechanism, so-called because the word-families that are created as a result – such as "turkey", "lamb", and "chicken" for the {animal}/{food} connection – are deemed to be lexical cousins. The problem, of course, is that such rules have many exceptions, and because these exceptions do not obey easy generalizations, they must be exhaustively listed. WordNet is thus forced to list an average of 15 exceptions for every cousin rule. The extent of this list represents a warning for all distributional approaches, suggesting that homonymy can be a very compelling null hypothesis when a word has multiple senses.

3 An Evidential Approach

Polysemy is a form of lexical ambiguity where different senses are psychologically related. So to recognize polysemy between word senses in WordNet without exploiting a distributional analysis, it is worth considering how these specific senses are connected, either ontologically, or via their glosses. We take as an example the word "olive", which has five noun senses in WordNet 1.6 (note: isa+ indicates hypernymy via transitive closure, and superscripts indicate the junctures at which one sense suggests a relationship with another).

First, note how three senses of "olive", 2,3 and 5, directly reference the concept {olive tree} in their glosses. We can assume then that these senses are all derivations of sense 4, which specifies {olive tree} as a hypernym. Now consider how some senses seem to relate to another sense via some kind of ontological *frame:slot* filling. Sense 5 is a {fruit} and sense 4 is a {fruit tree}, so it seems likely that sense 5 denotes the fruit of the tree denoted by sense 4. Sense 3 denotes a kind of {wood}, while sense 4 denotes a kind of {woody plant}, so again, it seems likely that sense 3 denotes the wood derived from the tree denoted by sense 4 (simple morphology is needed to make this connection). However, sense 1, the color sense, seems problematic until we realize that it can be considered a hyponym of the chromatic color {yellow, yellowness}. WordNet does not state this connection, but it can be inferred since sense 1 is also a chromatic color that mentions “yellow” in its gloss. Thus, sense 3, the wood sense, can be seen to refer to olive.1 in its gloss by its use of the term “yellow”.

olive.1 *gloss:* a yellow^{3} green color of low brightness and saturation
isa: {chromatic_color, spectral_color}

olive.2 *gloss:* one-seeded fruit^{5} of the European olive tree^{4} usually pickled …
isa: {relish}

olive.3 *gloss:* hard yellow^{1} variegated wood of an olive tree^{4}
isa: {wood^{4}} *which isa* {plant^{4} material}

olive.4 *gloss:* evergreen tree cultivated in the Mediterranean region …
isa: {olive tree} *which isa* {fruit^{5} tree}
which isa^{+} {wood^{3}y plant}

olive.5 *gloss:* small ovoid fruit of the European olive tree^{4}
isa: {fruit^{4}} *which isa*$^{+}$ {plant^{4} organ}

In effect then, the information needed to recognize the five noun senses of “olive” as mutually polysemous can be found in the definitions of these senses themselves. It is not necessary to look outside the senses to find distributional evidence elsewhere for what is already stated, in a somewhat implicit form, as part of the senses themselves. This is the essence of an evidence-based approach to polysemy detection: a variety of ontologically-motivated connection strategies are used to identify the implicit relationships between senses to support the hypothesis that these senses form a polysemous bond. When no evidence is found, we err on the side of caution and assume homonymy. In the case of “olive” above, the sense-pairings were produced by applying the following two strategies:

Explicit Ontological Bridging: a sense pair $<\omega_1, \omega_2>$ for a word ω can be bridged if ω_1 has a hypernym that can be lexicalized as M-H and ω_2 has a hypernym that can be lexicalized as M, the rationale being that ω_2 is the M of ω_1 and ω_1 is the H of ω_2. E.g., the word “basketball” has two WordNet senses, one a transitive hyponym of {game}, the other a hyponym of {game equipment}. In this case then, M = *game* and H = *equipment*. The second sense thus denotes the equipment used in the activity of the first sense.

Cross-Reference: if $<\omega_1, \omega_2>$ is a sense pair for a word ω and the WordNet gloss for ω_2 explicitly mentions a hypernym of ω_1, then ω_2 can be seen as a conceptual extension of ω_1. For instance, WordNet contains several senses of the word “charcoal”, one of which is a hyponym of {drawing}, another of which refers to “drawing” in its gloss. The latter sense, a hyponym of {writing implement}, can thus be seen as making a reference to the former. Cross-reference is a powerful connection strategy, and is all the more powerful for considering the glosses of sense hypernyms as well. For instance, WordNet defines “angler” as both a {fisherman, fisher} and as a type of acanthopterygian {fish} that lures other fish as its prey. Since the gloss for {fisherman, fisher} makes reference to “fish”, the polysemous link between both senses of angler can be detected.

No one strategy is powerful enough to recognize all inter-sense relationships. We see this approach not as an essentialist account of polysemy, but as an engineering approach to detecting the tell-tale connections that exist between sense descriptions in a hand-built ontology like WordNet. We thus engage a wide variety of different strategies, each capturing a different intuition about sense definitions and the way they reflect linguistic knowledge. Here are two more:

Hierarchical Reinforcement: if $<\alpha_1, \alpha_2>$ and $<\beta_1, \beta_2>$ are sense pairs for two words α and β where α_1 is a hypernym of β_1 and α_2 is a hypernym of β_2, then $<\alpha_1, \alpha_2>$ reinforces the belief that $<\beta_1, \beta_2>$ is polysemous, and vice versa. For example, “herb” can denote either a plant or a foodstuff in WordNet, while the words “sage”, “dill”, “coriander”, “cilantro” and twenty others can denote a subclass of either of these senses. If words like “herb” were truly homonymous, we would not expect their ambiguity, essentially an accident of language, to be mirrored at their subclass level. Hierarchical reinforcement is essentially a special case of the distributional approach, applying Apresjan’s intuition about systematicity to the local context of a word. Nonetheless, it is also an evidential strategy, seeking word-specific evidence before polysemy is hypothesized.

Morphosemantic Linking: a sense pair $<\omega_1, \omega_2>$ for a word ω can be related by this strategy if ω_1 has a hypernym that can be lexicalized as H_1, and ω_2 has a hypernym that can be lexicalized as H_2, and H_1 is morphologically derived from H_2 or vice versa. For instance, WordNet attributes a sense to “gossip” that is a hyponym of {communicator}, and another that is a hyponym of {communication} via {chat, confab, confabulation}. This suggests that both senses engage in a *communicator:communication* relationship. Morphosemantic linkages between synsets like {communication} and {communicator} are now provided as standard as part of WordNet 2.0, obviating the need for morphology rules to achieve this linkage.

4 Evaluating the Approach

The set of specific polysemy predictions made by the WordNet cousin rules (WordNet documentation, 2003) make an ideal basis of comparison for this approach.

Note however that the cousin rules provide only a partial account of the polysemy inherent in WordNet, connecting senses in just 20% of the ambiguous nouns in WordNet 1.6. In contrast, the 25 evidential strategies we have implemented so far (section 3 describes the most significant ones) succeed in connecting at least one pair of senses in 70% of all of WordNet's multi-sense nouns. Though the specific breakdown between polysemous and homonymous words in WordNet is unknown, we can nevertheless estimate the coverage of these strategies by calculating the percentage of cousin predictions that each strategy manages to independently replicate. Furthermore, because the cousin rules have a manually constructed exception list, this allows us to also estimate the precision of each strategy, by calculating the percentage of hypothesized instances of polysemy that do not correspond to cousin exceptions. The results of these estimations are shown in Table 1, reported both on a per-strategy basis and for all 25 strategies combined.

Strategy	Coverage	Precision
Ontological Bridging	15%	95%
Cross-Reference	76%	89%
Hierarchical Reinforcement	11%	76%
Morphosemantic Linking	3%	93%
All 25 Strategies combined	98%	85%

Table 1. Estimations of coverage and precision as measured against WN cousins.

The results suggest that WordNet contains enough information in its ontological structure and in its glosses to allow a non-distributional, word-by-word analysis of polysemy to succeed.

5 Concluding Remarks

Polysemy is not a weakness of WordNet but a strength, one that can open a richly illuminating window on lexico-conceptual structure if approached from the right perspective. The evidence-based approach described here offers an essentially bottom-up perspective on polysemy detection, while the distributional approach offers what is essentially a top-down perspective. However, both perspectives can complement each other, to help mitigate the weaknesses of both. The evidence-based approach can be stymied by opaque glosses and counter-intuitive ontological definitions, while the distributional approach is weak in dealing with ad-hoc polysemy and individual instances of homonymy that happen to only coincidently conform to systematic patterns of polysemy.

In an ideal lexical ontology, polysemy detection strategies would not rely on a shallow knowledge source like flat text glosses (as in *cross-reference*), but would derive all of their guidance from the logical structure of the ontology. It is instructive to imagine what such an ontology might look like, if the conceptual motivation for each instance of polysemy in a given language were to be somehow made explicit in its underlying taxonomic structure. As we look for stronger strategies to replace *cross-reference*, it is perhaps worth considering how we might change WordNet itself, both through automatic and manual activities, so that it may truly explain, rather than merely report, polysemy.

References

J. D. Apresjan (1974). Regular Polysemy. ***Linguistics***, 142:5-32.

D. A. Cruse. (2000). Aspects of the Micro-structure of Word Meanings. ***Polysemy: Theoretical and Computational Approaches,*** eds. Ravin, Y, Leacock, C. Oxford University Press, Oxford.

G. A. Miller (1995). WordNet: A Lexical Database for English. ***Communications of the ACM***, Vol. 38 No. 11.

W. Peters, I. Peters. (2000). Lexicalized Systematic Polysemy in WordNet. ***Proceedings of the 2nd international conference on Language Resources and Evaluation.*** Athens.

J. Pustejovsky. (1991). The Generative Lexicon. ***Computational Linguistics*** 17(4).

Y. Ravin, C. Leacock. (2000). Polysemy: An Overview. ***Polysemy: Theoretical and Computational Approaches,*** eds. Ravin, Y, Leacock, C. Oxford University Press, Oxford.

WordNet (2003). WordNet documentation. ***www.princeton.edu/~wn/***

Making an XML-based Japanese-Slovene Learners' Dictionary

Tomaž Erjavec

Department of Knowledge Technologies
Jožef Stefan Institute
Jamova 31, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Kristina Hmeljak Sangawa

Department of Asian and African Studies
Faculty of Arts
University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia
kristina.hmeljak@guest.arnes.si

Irena Srdanović

Department of Comparative and General Linguistics
Faculty of Arts
University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia
irena_srdanovic@hotmail.com

Anton ml. Vahčič

Faculty of Computer Science and Informatics
University of Ljubljana
Tržaška cesta 25, Ljubljana, Slovenia
vahcica@hotmail.com

Abstract

In this paper we present a hypertext dictionary of Japanese lexical units for Slovene students of Japanese at the Faculty of Arts of Ljubljana University. The dictionary is planned as a long-term project in which a simple dictionary is to be gradually enlarged and enhanced, taking into account the needs of the students. Initially, the dictionary was encoded in a tabular format, in a mixture of encodings, and subsequently rendered in HTML. The paper first discusses the conversion of the dictionary into XML, into an encoding that complies with the Text Encoding Initiative (TEI) Guidelines. The conversion into such an encoding validates, enriches, explicates and standardises the structure of the dictionary, thus making it more usable for further development and linguistically oriented research. We also present the current Web implementation of the dictionary, which offers full text search and a tool for practising inflected parts of speech. The paper gives an overview of related research, i.e. other XML oriented Web dictionaries of Slovene and East Asian languages and presents planned developments, i.e. the inclusion of the dictionary into the Reading Tutor program.

1. Introduction

The establishment of a new Department of Asian and African studies at the University of Ljubljana and a course of Japanese studies within it in 1995 brought forward the need for Japanese language teaching materials and dictionaries for Slovene speaking students. However, due to the limited number of potential users, probably not much more than the current 180 students of Japanese at our department, the compilation of such materials and dictionaries is not a particularly profitable project that could interest a publishing house. The teachers at our department therefore decided to create it with the help of our students, the final users of the dictionary (Hmeljak Sangawa, 2002).

The compilation of a dictionary that would satisfy the needs of Japanese language students both in terms of macrostructure and of microstructure, i.e. with enough lemmas and a detailed enough description for each lemma to cover users' needs, both for passive and for active use, is going to last for many years. However, adopting the "dictionary-making process with 'simultaneous feedback' from the target users to the compilers" which has been proposed by De Schryver and Prinsloo (2000) can help us turn the drawback of having few users into an asset: we can have direct contact and feedback from most of the users at all stages of compilation.

Initially, the dictionary was conceived in a tabular format, suitable for editing in a spreadsheet program, and from which it was possible to directly derive an HTML format. However, it became apparent that this structure exhibited various drawbacks; in particular, it was difficult to extend to accommodate a more complex dictionary structure, as well as being difficult to validate and exchange.

This paper describes the conversion of the dictionary format into XML (eXtensible Markup Language) (W3C, 2000), using a document type definition that complies with the TEI (Text Encoding Initiative) Guidelines (Sperberg-McQueen and Burnard, 2002). This approach takes into account international standards in the field and focuses on describing text properties, i.e. what a particular part of the text means. It brings a number of advantages, such as better documentation, ability to validate the structure of a document, simpler processing, better integration, interchange and longevity, as well as easier usage of data for linguistically oriented research. This format also enabled Web deployment of the dictionary, which offers a full-text search facility, as well as grouping the entries into "learning blocks", ordered by lessons and part-of-speech.

2. The Dictionary Model

Ideally, a dictionary should contain all items its users might ever want to look up. However, striving to cover all vocabulary our students might possibly encounter during their undergraduate study would be unrealistic in our situation. We therefore decided to cover only the core vocabulary encountered up to an intermediate level of language study, and not to include the more specialized or rare vocabulary. Such a vocabulary is presumably encountered at a time when the students' knowledge of Japanese enables them to use the wealth of existing