CONSISTENT CRITERIA FOR SENSE DISTINCTIONS

MARTHA PALMER
Department of Computer Science, IRCS
University of Pennsylvania, Phil, PA 19104
mpalmer@.cis.upenn.edu

Abstract.

This paper specifically addresses the question of polysemy with respect to verbs, and whether or not the sense distinctions that are made in on-line lexical resources are appropriate for computational lexicons. The use of sets of related syntactic frames and verb classes are examined as a means of simplifying the task of defining different senses, and the importance of concrete criteria such as different predicate argument structures, selectional restrictions and lexical co-occurrences is emphasized.

1. Introduction

The difficulty of achieving adequate hand-crafted semantic representations has limited the field of natural language processing to applications that can be contained within well-defined subdomains. The only escape from this limitation will be through the use of automated or semi-automated methods of lexical acquisition. However, the field has yet to develop a clear consensus on guidelines for a computational lexicon that could provide a springboard for such methods, in spite of all of the effort on different lexicon development approaches (Mel'cuk, 1988), (Pustejovsky, 1991),(Nirenburg et al., 1992), (Copestake and Sanfilippo, 1993), (Lowe et al., 1997), (Dorr, 1997). One of the most controversial areas has to do with polysemy. What constitutes a clear separation into senses for any one verb, and how can these senses be computationally characterized and distinguished? The answer to this question is the key to breaking the bottleneck of semantic representation that is currently the single greatest limitation on the general application of natural language processing techniques.

In this paper we specifically address the question of polysemy with respect to verbs, and whether or not the sense distinctions that are made in on- line dictionary resources are appropriate for computational lexicons. We examine the use of sets of related syntactic frames and verb classes as a means of simplifying the task of defining different senses, and we focus on the mismatches between distinctions that can readily be made with these tools and some of the distinctions that are made in available on-line lexical resources.

2. Challenges in building large-scale lexicons

Computational lexicons are an integral part of any natural language processing system, and perform many essential tasks. The difficulty in achieving adequate performance of one of the most visible tasks, word sense disambiguation, is currently creating a major roadblock on the information highway.[1] Information retrieval mismatches such as the retrieval of an article on plea bargaining, (**speed***ier trials and* **light***er sentences*), given *speed of light* as a query are caused by inadequate word sense disambiguation. These are clearly not the same senses of *light*, (or even the same parts of speech), but a system would have to distinguish between WordNet light1, involving `visible light`, and WordNet light2, having to do with `quantity or degree` in order to rule out this retrieval. (For a description of WordNet, a widely used available on-line lexical resource, see (Miller, 1990; Miller and Fellbaum, 1991). Failing to properly disambiguate can also lead to erroneous machine translations. For instance, in Korean, there are two different translations for the English verb *lose*, depending on whether it is an object that has been misplaced or a competition that has been lost: lose1, *lose the report* - *ilepeli-ess-ta*, and lose2, *lose the battle* - *ci-ess-ta*, (Palmer et al., 1998a).

These particular sense distinctions can be made by placing semantic class constraints on the objects of *lose*, `+solid object`, and `+competition`, as illustrated. The first constraint corresponds directly to a WordNet hypernym, but the second one does not. The closest correlate in WordNet would be `+abstract activity`, which is the common hypernym for both *hostile military engagement* and *game*, and which may discriminate sufficiently.

Computational lexicons can most readily make sense distinctions based on concrete criteria such as:

- different predicate argument structures
- different selectional restrictions on verb arguments
- different lexical co-occurrences, such as prepositions

[1]See the SIGIR-97 and 98 Proceedings for a comprehensive view of this field.

This seems straightforward enough, but in actuality available on-line resources rarely use this type of criteria for making sense distinctions. Or, even if they do, the criteria are not made explicit. For instance, out of the ten senses that WordNet 1.6, gives for *lose*, we find one, WN2, that corresponds to our lose1 from above, *lose the battle* sense, but two,WN1 and WN5, that correspond to our lose2, *misplace an item.*

– lose1 - WN2. lose – (fail to win; "We lost the battle but we won the war")
– lose2 - WN1. (fail to keep or to maintain; cease to have, either physically or in an abstract sense; fail to keep in mind or in sight; "She lost her purse when she left it unattended on her seat"; "She lost her husband a year ago")
– lose2 - WN5. (miss from one's possessions; lose sight of; "I've lost my glasses again!")

When we try to establish concrete criteria for distinguishing between WN1 (*lost her purse*) and WN5 (*lost my glasses*), we realize that these two WordNet senses are not distinguished because of anything to do with the verb arguments (an animate Agent and a solid object possessed by the Agent in both cases), but rather are distinguished by possible future events - namely the likelihood of the object being found. Are these two usages really different senses of the verb, or are they simply two different instances of the same verb, and the same basic event, but in very different world contexts which will therefore lead to different outcomes? In either event, it is not reasonable to expect a computational lexicon to, in addition to meticulously listing changes in syntactic context, in argument structure, and in selectional restrictions on those arguments, to also characterize all possible worlds in which an event can take place, and distinguish between all possible outcomes.

We are not denying that a computational lexicon should include particular changes in the state of the world that are entailed by specific actions, quite the contrary, (Palmer, 1990). However, the characterizations of these changes should be generally applicable, and cannot be so dependent on a single world context that they change with every new situation.

This does not invalidate the sense distinctions being made by WordNet (or other dictionaries). It is perfectly legitimate for people to use their knowledge of world context to illustrate subtle shifts in possible outcomes. However, we need to separate that from the task we are setting the computer, and the demands being placed on it. It may be appropriate someday for a complete natural language understanding system to produce representations of the sentences in the definitions presented here that capture exactly those differences in outcomes that are occasioned by the different situations. But that burden should not be bourne by the lexicon alone, but rather by the lexicon interacting with world knowledge and discourse context.

2.1. REGULARIZING CLASSES OF SENSE DISTINCTIONS

Other areas of difference between computational lexicons and more traditional lexical resources have to do with the flexibility of the representation. Computational lexicons are particularly well suited to capturing hierarchical relationships and regular sense extensions based on verb class membership. For instance, the following two senses are in and among the 63 sense distinctions WordNet listed for *break*.

- break - WN2. break, separate, split up, fall apart, come apart – (become separated into pieces or fragments; "The figurine broke"; "The freshly baked loaf fell apart")
- break - WN5. (destroy the integrity of; usually by force; cause to separate into pieces or fragments; "He broke the glass plate"; "She broke the match")

They are shown as being related senses in WordNet 1.6, but the relationship is not made explicit. It is a simple task for a computational lexicon to specify the type of relationship, i.e., the transitive frame is the `+causative` form of the ergative, and has explicit inclusion of an Agent as an additional argument. In the XTAG English lexicon, (Joshi et al., 1975), (Schabes et al., 1988; Schabes, 1990),(Palmer et al., 1998b), this is currently handled by associating both the ergative and transitive tree families [2] with the same syntactic database entry for *break*. In the transitive form the NP1 (Patient) becomes the Object and an NP0 (Agent) is added as the Subject, as well as a **+causative** semantic feature, as illustrated in Fig 1. All ergative verbs are given the same type of entry.

In addition to regular extensions in meaning that derive from systematic changes in subcategorization frames, there are also regular extensions occasioned by the adjunction of optional prepositions, adverbials and prepositional phrases. For example, the basic meaning of *push*, *He pushed the next boy*, can be extended to explicitly indicate accompanied motion by the adjunction of a path prepositional phrase, as in *He pushed the boxes across the room* (Palmer et al., 1997; Dang et al., 1998), which corresponds to WN1 below. The possibility of motion of the object can be explicitly denied through the use of the conative, as in *He pushed at the box*, which is captured by WN5. Finally, the basic sense can also be extended to indicate

[2] A tree family contains all of the syntactic realizations associated with a particular subcategorization frame, such as subject and object extraction and passive, (XTAG-Group, 1995).

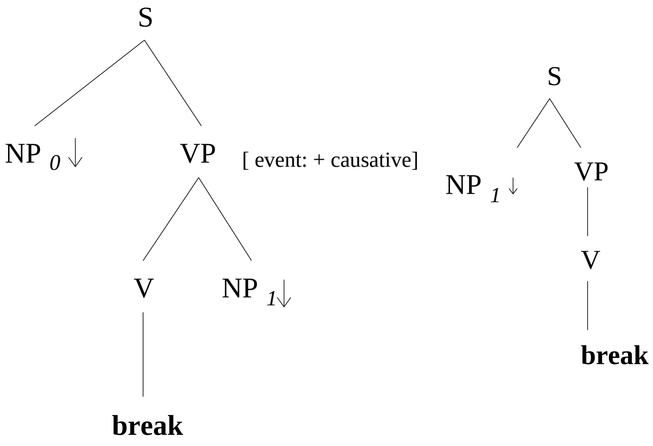

Figure 1. An ergative verb and its causative sense extension

a change of state of the object by the adjunction of *apart*, as in *He pushed the boxes apart.* There is no WordNet sense that corresponds to this, nor should there be. What is important is for the lexicon to provide the capability of recognizing and generating these usages where appropriate. If they are general enough to apply to entire classes of verbs, then they can be captured through regular adjunctions rather than being listed explicitly, (for more details, see (Bleam et al., 1998)).

- WN1. push, force – (move with force, "He pushed the table into a corner"; "She pushed her chin out")
- WN5. push – (press against forcefully without being able to move)

3. Conclusion

It has been suggested that WordNet sense distinctions are too fine-grained and coarser senses are needed to drive the word sense disambiguation task. For instance, in defining *cut*, WordNet distinguishes between WN1, *separating into pieces of a concrete object*, WN29, *cutting grain*, WN30, *cutting trees*, and WN33, *cutting hair*. For many purposes, the three more specialized senses, WN29, WN30 and WN33, which all involve *separation into pieces of concrete objects* could be collapsed into the more coarse-grained WN1. However, when searching for articles on recent changes in hair styles, the more fine-grained WN33 would still be useful. Computational lexicons actually lend themselves readily to moving back and forth between elements of an hierarchical representation based on concrete criteria, and this type of structuring should become more prevalent. The point is that they operate most effectively in the realm of concrete criteria for sense distinctions, such as changes in argument structure, changes in sets of syntactic frames and/or changes in selectional restrictions, and lexical co-occurrences. Distinctions that are based on world knowledge, no matter how diverse, are much more problematic. We must bear this in mind in order to design a word sense disambiguation task that will encourage rational, incremental development of computational lexicons.

References

Bleam, T., Palmer, M., and Shanker, V. (1998). Motion verbs and semantic features in tag. In *In the Proceedings of the TAG+-98 Workshop*, Philadelphia, PA.

Copestake, A. and Sanfilippo, A. (1993). Multilingual lexical representation. In *Proceedings of the AAAI Spring Symposium: Building Lexicons for Machine Translation*, Stanford, California.

Dang, H. T., Kipper, K., Palmer, M., and Rosenzweig, J. (1998). Investigating regular sense extensions based on intersective levin classes. In *Proceedings of Coling-ACL98*, Montreal, CA.

Dorr, B. J. (1997). Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12:1–55.

Joshi, A. K., Levy, L., and Takahashi, M. (1975). Tree Adjunct Grammars. *Journal of Computer and System Sciences.*

Lowe, J., Baker, C., and Fillmore, C. (1997). A frame-semantic approach to semantic annotation. In *Proceedings 1997 Siglex Workshop/ANLP97*, Washington, D.C.

Mel'cuk, I. A. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, pages 165–188.

Miller, G. A. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–312.

Miller, G. A. and Fellbaum, C. (1991). Semantic networks of english. *Lexical and Conceptual Semantics, Cognition Special Issue*, pages 197–229.

Nirenburg, S., Carbonell, J., Tomita, M., and Goodman, K. (1992). *Machine translation: a knowledge-based approach.* Morgan Kaufmann, San Mateo, California, USA.

Palmer, M. (1990). Customizing verb definitions for specific semantic domains. *Machine Translation*, 5.

Palmer, M., Han, C., Xia, F., Egedi, D., and Rosenzweig, J. (1998a). Constraining lexical selection across languages using tags. In Abeille, A. and Rambow, O., editors, *Tree Adjoining Grammars.* CSLI, Palo Alto, CA.

Palmer, M., Rosenzweig, J., Dang, H., and Xia, F. (1997). Capturing syntactic/semantic generalizations in a lexicalized grammar. presentation in working session of Semantic Tagging Workshop, ANLP-97.

Palmer, M., Rosenzweig, J., and Schuler, W. (1998b). Capturing motion verb generalizations with synchronous tag. In Dizier, P. S., editor, *Predicative Forms in NLP: Text, Speech and Language Technology Series.* Kluwer Press.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4).

Schabes, Y. (1990). *Mathematical and Computational Aspects of Lexicalized Grammars.* PhD thesis, Computer Science Department, University of Pennsylvania.

Schabes, Y., Abeillé, A., and Joshi, A. K. (1988). Parsing strategies with 'lexicalized' grammars: Application to Tree Adjoining Grammars. In *Proceedings of the* 12^{th} *International Conference on Computational Linguistics (COLING'88)*, Budapest, Hungary.

XTAG-Group, T. (1995). A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS 95-03, University of Pennsylvania.