Different Sense Granularities for Different Applications

Martha Palmer, Olga Babko-Malaya, Hoa Trang Dang
University of Pennsylvania
{mpalmer/malayao/htd}@linc.cis.upenn.edu

Abstract

This paper describes an hierarchical approach to WordNet sense distinctions that provides different types of automatic Word Sense Disambiguation (WSD) systems, which perform at varying levels of accuracy. For tasks where fine-grained sense distinctions may not be essential, an accurate coarse-grained WSD system may be sufficient. The paper discusses the criteria behind the three different levels of sense granularity, as well as the machine learning approach used by the WSD system.

1 Introduction

The difficulty of finding consistent criteria for making sense distinctions has been thoroughly attested to in the literature (Kilgarriff, '97, Hanks, '00). Difficulties have been found with truth-theoretical criteria, linguistic criteria and definitional criteria (Sparck-Jones, '86, Geeraerts, '93). In spite of the proliferation of dictionaries, there is no methodology by which two lexicographers working independently are guaranteed to derive the same set of distinctions for a given word, with objects and events vying for which is the most difficult to characterize (Cruse, '86, Apresjan, '74, Pustejovsky, '91, '95).

On the other hand, accurate Word Sense Disambiguation (WSD) could significantly improve the precision of Information Retrieval by ensuring that the senses of verbs in the retrieved documents match the sense of the verb in the query. For example, the two queries *What do you call a successful movie?* and *Whom do you call for a successful movie?* submitted to AskJeeves both retrieve the same set of documents, even though they are asking quite different questions, referencing very different senses of call. The documents retrieved are also not very relevant, again because they do not distinguish which matches contain relevant senses and which do not.

Tips on Being a **Successful Movie** Vampire ... I shall **call** the police.

Successful Casting **Call** & Shoot for ``Clash of Empires" ... thank everyone for their participation in the making of yesterday's **movie**.

Demme's casting is also highly entertaining, although I wouldn't go so far as to **call** it **successful**. This **movie**'s resemblance to its predecessor is pretty vague...

VHS **Movies**: **Successful** Cold **Call** Selling: Over 100 New Ideas, Scripts, and Examples from the Nation's Foremost Sales Trainer.

The two senses of *call* in the two queries can be easily distinguished by their differing predicate-argument structures. They are also separate senses in WordNet, but WordNet has an additional 26 senses for *call*, and the current best performance of an automatic Word Sense Disambiguation system this type of polysemous verb is only 60.2% (Dang and Palmer, 2002). Is it possible that sense distinctions that are less fine-grained than WordNet's distinctions could be made more reliably, and could still benefit this type of NLP application?

The idea of underspecification as a solution to WSD has been proposed in Buitelaar 2000 (among others), who pointed out that for some applications, such as document categorization, information retrieval, and information extraction it may be sufficient to know if a given word belongs to a certain class of WordNet senses or underspecified sense. On the other hand, there is evidence that machine translation of languages as diverse as Chinese and English will require all of the fine-grained sense distinctions that WordNet is capable of providing, and even more (Ng, et al 2003, Palmer, et. al., to appear).

An hierarchical approach to verb senses, of the type discussed in this paper, presents obvious advantages for the problem of word sense disambiguation. The human an-

notation task is simplified, since there are fewer choices at each level and clearer distinctions between them. The automated systems can combine training data from closely related senses to overcome the sparse data problem, and both humans and systems can back off to a more coarse-grained choice when fine-grained choices prove too difficult.

The approach to verb senses presented in this paper assumes three different levels of sense distinctions: PropBank Framesets, WordNet groupings, and WordNet senses. In a project for the semantic annotation of predicate-argument structure, PropBank, we have made coarse-grained sense distinctions for the 700 most polysemous verbs in the Penn TreeBank (Kingsbury and Palmer, '02). These distinctions are based primarily on different subcategorization frames that require different argument label annotations. In a separate project, as discussed in Palmer et al 2004, we have grouped SENSEVAL-2 verb senses (which came from WordNet 1.7). These manual groupings were shown to reconcile a substantial portion of the manual and automatic tagging disagreements, showing that many of these disagreements are fairly subtle (Palmer, et.al., '04).

The tree levels of sense distinctions form a continuum of granularity. Our criterion for the Framesets, being primarily syntactic, is also the most clear cut. These distinctions are based primarily on usages of a verb that have different numbers of predicate-arguments, however they also separate verb senses on semantic grounds, if these senses are not closely related. Sense groupings provide an intermediate level of hierarchy, where groups are distinguished by more fine-grained criteria. Both Frameset and grouping distinctions can be made consistently by humans and systems (over 90% accuracy for Framesets and 82% for groupings) and are surprisingly compatible; 95% of our *groups* map directly onto a single PropBank sense.

2 Background

2.1 Propbank

PropBank [Kingsbury & Palmer, 2002] is an annotation of the Wall Street Journal portion of the Penn Treebank II [Marcus, 1994] with dependency structures (or `predicate-argument' structures), using sense tags for highly polysemous words and semantic role labels for each dependency. An important goal is to provide consistent semantic role labels across different syntactic realizations of the same verb, as in the window in *[$_{ARG0}$ John] broke [$_{ARG1}$ the window]* and *[$_{ARG1}$ The window] broke*. PropBank can provide frequency counts for (statistical) analysis or generation components in a machine translation system, but provides only a shallow semantic analysis in that the annotation is close to the syntactic structure and each verb is its own predicate.

In addition to the annotated corpus, PropBank provides a lexicon that lists, for each broad meaning of each annotated verb, its *Frameset*, i.e., the possible arguments in the predicate and their labels and all possible syntactic realizations. The notion of ``meaning" used is fairly coarse-grained, and it is typically motivated from differing syntactic behavior. The Frameset also includes a ``descriptor" field for each role which is intended for use during annotation and as documentation, but which does not have any theoretical standing. The collection of Frameset entries for a verb is referred to as the verb's *frame*. As an example of a PropBank entry, we give the frame for the verb *leave* below. Currently, there are frames for over 3,000 verbs, with a total of just over 4,300 Framesets described. Of these 3,000 verb frames, only a small percentage 21.8 % (700) have more than one Frameset, with less than 100 verbs with 4 or more. The process of sense-tagging the PropBank corpus with the Frameset tags has just been completed.

The criteria used for the Framesets are primarily syntactic and clear cut. The guiding principle is that two verb meanings are distinguished as different framesets if they have distinct subcategorization frames. For example, the verb 'leave' has 2 framesets with the following frames, illustrated by the examples in (1) and (2):

Frameset 1: move away from
Arg0:entity leaving
Arg1:place left

Frameset 2: *give*
Arg0:giver / leaver
Arg1:thing given
Arg2:benefactive / given-to

(1) John left the room.
(2) Mary left her daughter-in-law her pearls in her will

2.2 WordNet Sense Groupings

In a separate project, as part of Senseval tagging exercises, we have developed a lexicon with another level of coarse-grained distinctions, as described below.

The Senseval-1 workshop (Kilgarriff and Palmer, 2000) provided convincing evidence that supervised automatic systems can perform word sense disambiguation (WSD) satisfactorily, given clear, consistent sense distinctions and suitable training data. However, the Hector lexicon that was used as the sense inventory was very small and under proprietary constraints, and the question remained

whether it was possible to have a publicly available, broad-coverage lexical resource for English and other languages, with the requisite clear, consistent sense distinctions.

Subsequently, the Senseval-2 (Edmonds and Cotton, 2001) exercise was run, which included WSD tasks for 10 languages. A concerted effort was made to use existing WordNets as sense inventories because of their widespread popularity and availability. Each language had a choice between the lexical sample task and the all-words task. The most polysemous words in the English Lexical Sample task are the 29 verbs, with an average polysemy of 16.28 senses using the pre-release version of WordNet 1.7. Double blind annotation by two linguistically trained annotators was performed on corpus instances, with a third linguist adjudicating between inter-annotator differences to create the "Gold Standard." The average inter-annotator agreement rate was only 71%, which is comparable to the 73% agreement for all words in SemCor, with a much lower average polysemy. However, a comparison of system performance on words of similar polysemy in Senseval-1 and Senseval-2 showed very little difference in accuracy (Palmer et al., submitted). In spite of the lower inter-annotator agreement figures for Senseval-2, the double blind annotation and adjudication provided a reliable enough filter to ensure consistently tagged data with WordNet senses. Even so, the high polysemy of the WordNet 1.7 entries on average poses a challenge for automatic word sense disambiguation. In addition, WordNet only gives a flat listing of alternative senses, unlike most standard dictionaries which are more structured and often provide hierarchical entries. To address this lack, the verbs were grouped by two or more people, with differences being reconciled, and the sense groups were used for coarse-grained scoring of the systems.

The criteria used for groupings included syntactic and semantic ones. Syntactic structure performed two distinct functions in our groupings. Recognizable alternations with similar corresponding predicate-argument structures were often a factor in choosing to group senses together, as in the Levin classes and PropBank, whereas distinct subcategorization frames were also often a factor in putting senses in separate groups. Furthermore, senses were grouped together if they were more specialized versions of a general sense. The semantic criteria for grouping senses separately included differences in semantic classes of arguments (abstract versus concrete, animal versus human, animacy versus inanimacy, different instrument types...), differences in the number and type of arguments (often reflected in the subcategorization frame as discussed above), differences in entailments (whether an argument refers to a created entity or a resultant state), differences in the type of event (abstract, concrete, mental, emotional...), whether there is a specialized subject domain, etc.

Senseval-2 verb inter-annotator disagreements were reduced by more than a third when evaluated against the groups, from 29% to 18%, and by over half in a separate study, from 28% to 12%. A similar number of random groups provided almost no benefit to the inter-annotator agreement figures (74% instead of 71%), confirming the greater coherence of the manual groupings.

3 Mapping of Sense Groups to Framesets

Groupings of senses for Senseval-2, as discussed above, use both syntactic and semantic criteria. Propbank, on the other hand, uses mostly syntactic cues to divide verb senses into framesets. As a result, framesets are more general than sense-groups and usually incorporate several sense groups. We have been investigating whether or not the groups developed for SENSEVAL-2 can provide an intermediate level of hierarchy in between the PropBank Framesets and the WN 1.7 senses, and our initial results are promising. Based on our existing WN 1.7 tags and frameset tags of the Senseval2 verbs in the Penn TreeBank, 95% of the verb instances map directly from sense groups to framesets, with each frameset typically corresponding to two or more sense groups, as illustrated by the tables 1-4 for the verbs 'serve', 'leave', 'pull', and 'see'[1] below.

As the tables 1-4 illustrate, the criteria used to split the Framesets into groups are as follows:

1) Syntactic Frames. Most verb senses which allow syntactic alternations (such as transitive/inchoative, unspecified object deletion, etc) are analyzed as one sense group. However, in some cases, as illustrated by the verb *leave*, intransitive and transitive uses are distinguished as different sense groups:

Group 1: DEPART (*Ship leaves at midnight*)
Group 2: LEAVE BEHIND (*She left a mess*.)

The DEPART sense of the verb can be used transitively if the object specifies the place of departure. The LEAVE BEHIND sense is more general and allows syntactic variation as well as different semantic types of NPs. In PropBank, these groups are unified as one frameset (Frameset 1 MOVE AWAY FROM).

[1] All these verbs have one or more additional framesets, which correspond to one group or sense, and therefore are not included here

Frameset	Senseval-2 Groupings	Examples from WordNet
serve 01: *Act, work* Roles: Arg0:worker Arg1:job, project Arg2:employer	GROUP 1: WN1 (function) WN3(contribute to) WN12 (answer)	 His freedom served him well The scandal served to increase his popularity Nothing else will serve
	GROUP 2: WN2 (do duty) WN13 (do military service)	 She served in Congress She served in Vietnam
	GROUP 5: WN7 (devote one's efforts) WN10 (attend to)	 She served the art of music May I serve you?
	GROUP 3: WN4 (be used by) WN8 (serve well) WN14 (service)	 The garage served to shelter horses Art serves commerce Male animals serve the females for breeding purposes

Table 1. Frameset serve 01.

Frameset	Senseval-2 Groupings	Examples from WordNet
leave 02: *Move away from* Roles: Arg0:entity leaving Arg1:thing left Arg2 :attribute / secondary predication	GROUP 2: WN2 (leave behind) WN12 (be survived by) WN14 (forget)	 She left a mess He left six children I left my keys
	GROUP 1: WN1 (go away) WN5 (exit, go out) WN8 (depart)	 The ship leaves at midnight Leave the room The teenager left home
	GROUP 3: WN3 (to act) WN7 (result in)	 The inflation left them penniless Her blood left a stain on the napkin
	SINGLETON WN4 (leave behind)	 Leave it as is
	SINGLETON WN6 (allow for, provide)	 Leave lots of time for the trip

Table 2. Frameset leave 02.

2. Optional Arguments. In PropBank verbs of manner of motion and verbs of directed motion are usually grouped into one frameset. For example, one of the framesets of the verb *pull* (TRY) TO CAUSE MOTION unifies the following two group senses:
Group 1: MOVE ALONG (*pull a sled*)
Group 2: MOVE INTO A CERTAIN DIRECTION (*The van pulled up*)

Although the frame for the frameset 1 of the verb *pull* has a 'direction' argument, this argument does not have to be present (or implied), and verbs with this frame can also be understood as verbs of manner of motion in PropBank.

3) Syntactic variation of arguments. Syntactic variation in objects can also be used to distinguish sense groups, but are not taken into consideration for distinguishing framesets. Here both noun phrases and sen-

Frameset	Senseval-2 Groupings	Examples from WordNet
pull.01: try to cause motion Roles: Arg0:puller Arg1:thing pulled Arg2: direction or predication Arg3:extent, distance moved	GROUP 1: WN1 (draw) WN4(apply force) WN9 (cause to move) WN10 (operate) WN13 (hit)	 Pull a sled Pull the rope A declining dollar pulled down the export figures Pull the oars Pull the ball
	GROUP 2: WN2 (attract) WN12 (rip)	 The ad pulled in many potential customers Pull the cooked chicken into strips
	GROUP 3: WN3 (move) WN7 (steer)	 The car pulls to the right Pull the car over
	GROUP 4: WN6 (pull out) WN15 (extract) WN17(take away)	 The mugger pulled a knife on his victim Pull weeds Pull the old soup cans from the shelf

Table 3. Frameset pull 01.

Frameset	Senseval-2 Groupings	Examples from WordNet
see.01: *view* Roles: Arg0:viewer Arg1:thing viewed Arg2:secondary attribute	GROUP 1: WN1 (perceive by sight) WN7 (watch) WN19 (observe as if with an eye) WN20 (examine)	 Can you the bird? See a movie The camera saw the burglary I must see your passport
	GROUP 3: WN3 (witness) WN6 (learn)	 I want to see the results I see that you have been promoted
	GROUP 4: WN5 (consider) WN24 (interpret)	 I don't see the situation quite as negatively What message do you see in this letter?
	GROUP 5: WN8 (determine) WN10 (check) WN14 (attend)	 See whether it works See that the curtains are closed Could you see about lunch?
	GROUP 6: WN11 (see a professional) WN15 (receive as a guest)	 You should see a lawyer The doctor will see you now

Table 4. Frameset see 01.

tential complements are contained in the same frameset. These could also be distinguished by the type of event, a physical perception vs. an abstract or mental perception, but these would also not distinguished by PropBank.

Group 1: PERCEIVE BY SIGHT (*Can you see the bird?*)
Group 5: DETERMINE, CHECK (*See whether it works*)

4) Semantic classes of arguments. Differences in semantic classes of arguments, such as ANIMACY versus INANIMACY, are also not considered for distinguishing framesets. The verb *serve*, for example, has the following group senses, the second of which requires an ANIMATE agent, which are unified as one frameset in PropBank:

Group 1: FUNCTION (*His freedom served him well*)
Group 2: WORK (*He served in Congress*)

Most of the criteria which are used to split Framesets into groupings, as the tables above illustrate, are se-

mantic. These distinctions, although more fine-grained than Framesets, are still more easily distinguished than WordNet senses.

Mismatches between Framesets and groupings usually occur for the following two reasons. First, some senses can be missing in the PropBank, if they do not occur in the corpus. Second, given that PropBank is an annotation of the Wall Street Journal, it often distinguishes obscure financial senses of the verb as separate senses.

4 Experiments with Automatic WSD

We have also been investigating the suitability of these distinctions for training automatic Word Sense Disambiguation systems. The system that we used to tag verbs with their frameset is the same maximum entropy system as that of Dang and Palmer (2002), including both topical and local features. Topical features looked for the presence of keywords occurring *anywhere* in the sentence and any surrounding sentences provided as context (usually one or two sentences). The set of keywords is specific to each lemma to be disambiguated, and is determined automatically from training data so as to minimize the entropy of the probability of the senses conditioned on the keyword.

The local features for a verb *w* in a particular sentence tend to look only within the smallest clause containing *w*. They include *collocational* features requiring no linguistic prepro essing beyond part-of-speech tagging (1), *syntactic* features that capture relations between the verb and its complements (2-4), and *semantic* features that incorporate information about noun classes for objects (5-6):

1) the word *w*, the part of speech of *w*, and words at positions -2, -1, +1, +2, relative to *w*

2) whether or not the sentence is passive

3) whether there is a subject, direct object, indirect object, or clausal complement (a complement whose node label is S in the parse tree)

4) the words (if any) in the positions of subject, direct object, indirect object, particle, prepositional complement (and its object)

5) a Named Entity tag (PERSON, ORGANIZATION, LOCATION) for proper nouns appearing in (4).

6) all possible WordNet synsets and hypernyms for the nouns appearing in (4).

The system performed well on the English verbs in Senseval-2, achieving an accuracy of 60.2% when tagging verbs with their fine-grained WordNet senses, and 70.2% when tagging with the more coarse-grained sense groups.

Verb	Framesets	Instances	Accuracy
call	11	522	0.835
carry	4	195	0.933
develop	2	240	0.938
draw	3	94	0.926
dress	3	15	0.800
drive	2	99	0.808
keep	5	136	0.919
leave	3	147	0.762
live	4	125	0.888
play	5	98	0.806
pull	6	88	0.784
see	2	187	0.995
serve	2	150	0.967
strike	10	59	0.610
train	2	17	0.941
treat	2	51	0.863
turn	14	141	0.638
use	2	820	0.988
wash	2	8	0.875
work	7	398	0.955

Table 5. Frameset tagging results

For frameset tagging, we collected a total of 3590 instances of 20 verbs in the PropBank corpus that had been annotated with their framesets. The verbs all had more than one possible frameset and were a subset of the ones used for the English lexical sample task of Senseval-2. Local features for frameset taging were extracted using the gold-standard part-of-speech tags and bracketing of the Penn Treebank. Table 5 shows the number of framesets, the number of instances, and the system accuracy for each verb using 10-fold cross-validation. The overall accuracy of our automatic frameset tagging was 90.0%, compared to a baseline accuracy of 73.5% if verbs are tagged with their most frequent frameset. While the data is only a subset of that used in Senseval-2, it is clear that framesets can be much more reliably tagged than fine-grained WordNet senses and even sense groups.

Conclusion

This paper described an hierarchical approach to WordNet sense distinctions that provided different types of automatic Word Sense Disambiguation (WSD) systems, which perform at varying levels of accuracy. We have described three different levels of sense granularity, with PropBank Framesets being the most syntactic, the most coarse-grained, and most easily reproduced. A set of manual groupings devised for Senseval2 provides a middle level of granularity that mediates between Framesets and WordNet. For tasks where fine-grained sense distinctions may not be essential such as an AskJeeves information retrieval task, an accurate coarse-grained WSD system such as our Frameset tagger may be sufficient. There is evidence, however, that machine translation of languages as diverse as Chinese and English might require all of the fine-grained sense distinctions of WordNet, and even more (Ng, et al 2003, Palmer, et. al., to appear).

References

Apresjan, J. D. .(1974) Regular polysemy, *Linguistics*, 142:5—32.

Atkins, S. (1993) Tools for computer-aided corpus lexicography: The Hector Project. *Actu Linguistica Hunguricu*, 41:5-72.

Buitelaar, P.P (2000). Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification. In *Poceedings of the ANLP Workshop on Syntactic and Semantic Complexity in NLP Systems.* Seattle, WA.

Cruse, D. A., (1986), *Lexical Semantics*, Cambridge University Press, Cambridge, UK, 1986.

Dang, H. T. and Palmer, M., (2002). Combining Contextual Features for Word Sense Disambiguation. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, Pa.

Edmonds, P. and Cotton, S. (2001). SENSEVAL-2: Overview. In Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems, ACL-SIGLEX, Toulouse, France.

Hanks, P., (2000), Do word meanings exist? Computers and the Humanities, Special Issue on SENSEVAL, 34(1-2).

Geeraerts, D., (1993), Vagueness's puzzles, polysemy's vagaries, *Cognitive Linguistics*, 4.

Kilgarriff, A., (1997), I don't believe in word senses, *Computers and the Humanities*, 31(2).

Kilgarriff, A. and Palmer, M., (2000), Introduction to the special issue on Senseval, *Computers and the Humanities*, 34(1-2):1-13.

Kingsbury, P., and Palmer, M, (2002), From TreeBank to PropBank, *Third International Conference on Language Resources and Evaluation, LREC-02*, Las Palmas, Canary Islands, Spain, May 28- June 3.

Marcus, M, (1994), The Penn TreeBank: A revised corpus design for extracting predicate argument structure, In *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, NJ.

Ng, H. T., & Wang, B., & Chan, Y. S. (2003). Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In the *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03).* Sapporo, Japan, July.

Palmer, M., Dang, H. T., and Fellbaum, C., (to appear, 2004), Making fine-grained and coarse-grained sense distinctions, both manually and automatically, under revision for *Natural Language Engineering*.

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

Pustejovsky, J. (1995) *The Generative Lexicon*, Cambridge, MIT Press, Mass.