

Manual and Automatic Semantic Annotation with WordNet

Christiane Fellbaum
Princeton University
fellbaum@clarity.princeton.edu

Martha Palmer, Hoa Trang Dang, Lauren Delfs, Susanne Wolf
Department of Computer and Information Science
University of Pennsylvania
mpalmer, htd, lcdelfs, ser@unagi.cis.upenn.edu

Abstract

Semantically annotated corpora are needed to train automatic systems for word sense disambiguation. Truly reliable training corpora must be hand-constructed and require meticulous and time-consuming work. But the annotations of human “taggers” also open a window into the lexical organization of speakers and how mental representations can be mapped onto those of the dictionary from which the taggers select the senses. We examine some inter-annotator discrepancies from a large-scale annotation project. These may reveal the nature of speaker-specific lexical organization and sense representations; they also show the inadequacies of dictionaries and suggest possibilities for systematic entry modification. Such entries would allow for underspecified and clustered entries that a tagger could select when the context is indeterminate. Finally, we report on some initial results with an automatic system trained on the tags of one of the human annotators. The ability to test on clustered senses significantly improves the tagger’s results.

1 Introduction

Most Natural Language Processing applications require large-scale, sophisticated lexical resources that are amenable to reliable word sense disambiguation (WSD). Polysemous words with related but subtly distinct meanings present the greatest hurdle. The most polysemous words are not only the most frequently occurring words, but also domain-independent, making the WSD problem a universal one.

One approach to building a reliable automatic system for WSD is to train a system on corpora that have already been annotated by humans. Such corpora are hard to come by, as human annotation is labor-intensive and expensive. Moreover, manual annotation is not as reliable as one might think—people cannot easily identify a token in a text with a particular sense in a dictionary. Dictionaries by necessity represent word meanings in a way that makes them difficult to map onto words in a text because the kind of lexical information derivable from these

two resources differs so greatly.

Dictionaries are created for the purpose of helping their user identify the meaning of an unknown word or usage. The assumption is that the user has the context but needs to understand the word. For polysemous words, dictionaries list for each sense paradigmatically related words such as synonyms and superordinates, as well as definitions. But they usually do not say enough about the range of possible contexts that differentiate the senses.

Miller and Gildea (1987) have demonstrated the limited use of dictionaries as a source of information for the contextual aspects of lexical knowledge. They had children write sentences using novel words that the children had looked up in a dictionary. Their young subjects had clearly understood the dictionary definitions but the sentences demonstrated that this information is not sufficient for learning a word’s syntagmatic properties. The children wrote sentences like “My family erodes a lot” (*erode* was glossed in the dictionary as *eat out*¹) and “She was meticulous about falling off the cliff” (based on the dictionary definition of *meticulous* as *careful*).

On the other hand, texts or corpora tell us a lot about how a word is used, but they are not explicit about the word’s meaning. When we first encounter a new word in a text, we can usually form only a vague idea of its meaning; checking a dictionary will clarify the meaning. But the more contexts we encounter for a word, the harder it is to match them against only one dictionary sense.

2 Constructing an annotated corpus

It is easy enough to automatically extract from a corpus all the occurrences of a given word. But such a concordance does not distinguish between the different senses of the target word. A human annotator needs to inspect all the corpus lines and distinguish the different senses with respect to a dictionary; the annotator then records a link between a given occurrence of a word and the corresponding sense in the dictionary. This process of semantic annotation is also referred to as “tagging.”

¹as in *remove* or *eat away at*

Fellbaum, Grabowski, and Landes (1997) discuss some of the tacit assumptions underlying the annotation task. Tagging relies on what one might call the “dictionary model” of word representation, namely, that word senses are discrete and enumerable. The dictionary model predicts that annotation is easy: Taggers inspect the occurrences of a (polysemous) string in a corpus, interpret and determine its meanings, and match these against dictionary entries. Tagging should be easy, since it would mimic our everyday behavior of processing language input and “looking up entries” in our mental lexicons. Under this model, tagging is also the inverse of corpus-based lexicography, where the lexicographer gathers the occurrences of a (polysemous) string from a corpus, interprets and determines its meaning(s), and creates an appropriate dictionary entry.

But a comparison of different dictionaries, including WordNet, shows up significant differences with respect to entries for polysemous words. First, not all senses are represented in each dictionary; lexicographers and editors presumably choose those senses they consider the most important and most frequent. Second, a single sense in one dictionary may be broken up into distinct subsenses in another dictionary. For example, Webster’s and the American Heritage Dictionary distinguish a transitive (causative) and an intransitive sense for many verbs of change and motion, while Collins Dictionary merges them into a single sense. Finally, different dictionaries often cover the same semantic space in the entry for a polysemous word, but they carve it up into different and only partially overlapping senses (Palmer et al., 2000). We must conclude from these facts that there is no unequivocal mental lexical representation that lexicographers, and by extension, all speakers, can consult in a straightforward look-up fashion.

Alternative models of meaning representation, such as prototype theory, are perhaps more realistic and could account better for speakers’ capacity to interpret a large number of conventional and novel usages of polysemous words. But we have no way to represent such a model, which does not assume fixed correspondences between a word form and a meaning, in a dictionary that can be used in semantic annotation. An interesting theory of word meaning is represented by the Generative Lexicon (Pustejovsky, 1995). This model explores the systematic extension of underspecified senses based on the context. For example, sentence (1a) leaves open whether the book’s contents or its physical make-up are of good quality; sentences (1b) and (1c) pick out only these specific meanings, respectively.

1. (a) This book is good.
- (b) This book is interesting.
- (c) This book is torn.

The Generative Lexicon theory suggests ways to represent word meanings more flexibly and allow for broader, underspecified senses, as well as more specific senses. Such lexical representations might lead to higher annotation agreement and accuracy, but awaits large-scale implementation.

WordNet’s entries resemble those of a traditional dictionary, though its organization is not alphabetical but that of a semantic network. WordNet was constructed largely without the benefit of corpus data. But WordNet has been tested against corpora and used as a dictionary for semantic annotation (Landes et al., 1998). Such efforts should not only build on WordNet but also be exploited to improve its contents.

3 SENSEVAL

A recent exercise in automatic Word Sense Disambiguation, SENSEVAL, provided valuable experience in the complementary task of creating annotated corpora which could be used to train the automatic taggers (Kilgarriff and Palmer, 2000). The evaluation format followed DARPA style evaluations in that the participants were provided with hand annotated training data and test data and a pre-defined metric for evaluation.² The evaluation scheme provided three different scorings: one for exact matches to fine-grained senses; one for matches with only the more coarse-grained senses which ignored more fine-grained sense distinctions; and a mixed-match that gave partial credit for coarse-grained matches. ROMANSEVAL, an evaluation for French and Italian, was run in parallel.

The lexical inventory used was the Hector lexicon, developed jointly by DEC and OUP. After selecting the 34 target lexical items, highly trained Oxford University Press (OUP) lexicographers tagged sentences containing those items that had been extracted from the Hector corpus. By allowing for revision of confusing lexical entries during the tagging of the training data, inter-annotator agreement of over 90% was achieved. The training data consisted of at least 100 instances of each word, with the sentence that the word appeared in and its preceding sentence having been extracted from the Hector corpus. The test data included at least 30 instances. This corpus and the sense inventories for the lexical items are now available on the web for research and development purposes. At the workshop there were 54 attendees who presented the results of 24 systems, from universities and companies in Europe, North America and Asia. In general the systems did surprisingly well, with several of the supervised

²A consensus on a standardized evaluation metric - following loosely the Resnik and Yarowsky proposal from SIGLEX-97 as amended by Dan Melamed and implemented by Joseph Rosenzweig - was finalized by e-mail.

systems getting precision and recall numbers in the high 70's and low 80's (Kilgariff and Rosenzweig, 2000). There was also much discussion of the need for establishing meaningful baselines for the task, and clearer criteria for making sense distinctions. These criteria should include relevance to information processing tasks such as information retrieval and extraction and machine translation. The evaluation task should also include running text as well as corpus instances.

The workshop provided convincing evidence that automatic systems can perform well, given clear, consistent sense distinctions and suitable training data. This does not, however, solve the problem of making available to the public a broad-coverage English (or any other language) lexical resource with the requisite clear, consistent sense distinctions, and the corresponding tagged corpora. A new workshop, SENSEVAL2, which will include WSD tasks for 10 languages, is planned for July, 2001 in conjunction with ACL'01 in Toulouse, France. A concerted effort is being made to use existing WordNets as sense inventories because of their wide adoption internationally. Data tagged with a widely-used public domain resource could be especially useful to the broader community, but only if the tagging is of a quality similar to the previous SENSEVAL data, so careful studies of inter-annotator agreement are necessary.

4 Tagging the Penn TreeBank

Currently, we are engaged in adding semantic annotations to the syntactically annotated Penn TreeBank (Marcus et al., 1993) with the goal of preparing a corpus that can be used to train automatic systems (Palmer et al., 2000). The underlying assumption is that the discriminators that people use when distinguishing senses can be learned by a system and successfully used. People presumably use these discriminators unconsciously, as they do when processing language in natural contexts, or, for that matter, when tagging.

The Penn TreeBank semantic annotation project differs in several respects from previous tagging efforts such as SemCor (Miller et al., 1994; Landes et al., 1998). First, the Penn annotators are linguistically trained, whereas most of the Princeton taggers were linguistically naïve. In evaluating the annotations of the Princeton taggers, Fellbaum, Grabowski, and Landes (1997; 1998) found that, although the taggers agreed with each other to a significant extent, there was a statistically significant difference between the tags of the two supervising linguists and the naïve tagger group.

Second, the Princeton taggers tagged running text. This required them to (a) familiarize themselves with many different lexical entries in each tagging session, and (b) refamiliarize themselves with

the entry for a frequently occurring word each time it came up in the text, instead of considering multiple occurrences (with different senses) and weighing these against each other.

These considerations suggest, in hindsight, that serial tagging puts an unnecessary burden on the annotators. Targeted tagging, as in SENSEVAL, where all occurrences of one polysemous word are tagged at the same time, allows the annotators to familiarize themselves with the lexical entry for a given word, examine all occurrences of this word in the corpus, and analyze the entire dictionary entry in light of the data. When all occurrences of one word are being tagged in one session, potential errors may be eliminated that arise merely from the fact that the taggers have to examine the entire verb entry each time they hit upon a given verb in serial tagging. In the case of targeted tagging, the annotators can “learn” one dictionary entry at a time and have it at their fingertips. The Penn TreeBank is being tagged in a targeted fashion, for which, incidentally, the taggers expressed a strong preference.

5 Annotating verbs

Fellbaum, Grabowski, and Landes (1997; 1998) showed that the Princeton taggers disagreed among each other and with the “expert” taggers far more often on the annotation for verbs than for words from other lexical categories. This was partly due to the fact that verbs are far more polysemous than nouns, both in WordNet and in other dictionaries (Fellbaum, 1998b). But, more interestingly, verb meanings appear to be more flexible and less fixed than the meanings of nouns.

As part of the preparation for the SENSEVAL2, instances of thirty highly polysemous English verbs from the Penn TreeBank were hand-tagged. In the first phase of the tagging project, two linguistically trained annotators each tagged the same tokens independently of each other. They tagged between 75 and 300 tokens of each verb, depending on the number of senses. The verbs included some of the most polysemous ones (such as *call* and *draw*).

The annotations were compared and the discrepancies were examined. Our goal was to discern patterns of disagreement in the way the WordNet senses were interpreted against the tokens in the corpus. Specifically, we hoped to learn which senses the taggers interpreted as being semantically close or overlapping. Such senses should either be merged or grouped into clusters. Senses that are members of a cluster each represent a specific reading that arises from particular semantic or syntactic contexts. The cluster as a whole represents a broader, underspecified sense.

WordNet currently contains several thousand clustered verb senses. Clustering was done follow-

ing both syntactic and semantic criteria.

6 Syntactic clustering of related verbs

WordNet has a distinct entry for each syntactic use of a verb. Entries related by syntactic alternations such as indefinite object drop, cognative object realization, and causative/inchoative are grouped as follows:

1. (a) We ate fish and chips.
(b) We ate at noon.
2. (a) They danced a wild dance.
(b) They danced.
3. (a) He chilled the soup.
(b) The soup chilled.

Note that WordNet’s design forces sense distinctions for a given verb based on argument alternations (Levin, 1993) to appear in different hierarchies (Fellbaum, 1998a). For example, *wash* in one of its transitive uses (“He washed the dirty shirts”) has the superordinate *clean*; as a middle (“Do these shirts wash by hand?”), where the verb denotes a certain property of the subject, the verb’s hyponym is *be*; when the washed entity is projected as an oblique argument (“He washed the spots from the shirts”), the superordinate is *remove*. These different senses are clearly related, but this relation is not revealed by a search for their hypernyms; independent semantic criteria must be applied. But by itself, syntactic clustering is uncontroversial.

7 Semantic distinctions in WordNet

But there are no equally clear criteria for semantic similarity that could guide meaning-based clustering. The verb sense clusters currently in WordNet were created largely on the basis of lexicographic intuitions. An examination of the taggers’ data should provide a firmer basis for capturing meaning similarity as the basis for clusters. The analysis of inter-annotator disagreements below highlights specific types of semantic relationships between closely related senses. Bearing these findings in mind, we are currently experimenting with techniques for semantic grouping of senses. An example of our recent groups for *call*, which appears in WordNet 1.7, is given in Section 9. We will re-examine the existing groupings of all of our SENSEVAL2 verbs to make explicit the criteria for each sense’s inclusion in a group.

8 Inter-annotator disagreements and consequences

A comparison of the annotations of the two taggers showed that, unlike in the case of SemCor, the rate

of disagreement was not proportional to the number of WordNet senses. We found fairly high disagreement rates between the two taggers for words with both large and small numbers of WordNet senses. The annotators’ disagreements could have been due to the impossibility of identifying an unambiguous match for a specific occurrence in the sense inventory of WordNet, i.e., to the lack of a sense matching the occurrence, and the taggers’ choices of what each saw as the best possible match. Alternatively, the disagreements could have arisen from the taggers’ different interpretations of the target words.

The disagreements were considered and, whenever possible, resolved by a third person. After genuine errors were discounted, the remaining discrepancies showed some systematic patterns that could be classified and characterized. The semantic disagreements showed clearly that one tagger turned out to be a “lumper,” who consistently selected fewer senses, while the other was a “splitter,” who chose several senses to the lumper’s single sense. The lumper’s choices often corresponded to a broader, more general sense that arguably included the narrower senses selected by the splitter.

We present a brief typology of semantic similarities among senses of polysemous verbs. These kinds of similarity exemplify sources of the annotators’ disagreements. At the same time, they serve as a guide to the kind of clustering of semantically related senses that corresponds to human intuition and can perhaps be learned by an automatic system. In addition to clustering senses, we allow for double tagging, i.e., for annotations where the context does not allow the distinction between two unrelated senses.

8.1 Sense subsumption and blending

Three senses of the verb *live* were involved in inter-tagger disagreements:

1. *be, live* (have life, be alive; “Grandfather lived til the end of war”)
2. *survive, last, live, live on, go, endure, hold up, hold out* (continue to live; endure or last; “The legend of Elvis lives on”)
3. *exist, survive, live, subsist* (support oneself; Can you live on \$2000 a month in New York City?)

Sense 1 is the broadest sense and subsumes senses 2 and 3, which have an additional meaning component each: living beyond an implicit expectation or norm in sense 2; and the specific “economic survival” meaning in sense 3. Some tokens in the corpus allowed a clear and unequivocal match with one sense; in other cases, the verb’s meaning was a blend of the senses and the annotations differed. There is no reason to assume that an automatic system could discriminate the senses where the taggers were not able

to do so. To tag occurrences with meaning components from different senses, we allow for annotation using a cluster of senses.

8.2 Selectional restrictions and aspectual distinctions

The verb *use* was tagged 116 times by both annotators, producing 30 disagreements. The taggers could choose from the 6 senses of this verb in WordNet; all 6 senses were involved in the discrepancies. The “lumper” chose the following same sense in all but 3 of the discrepant cases:

1. use, utilize, utilise, apply, employ – (put into service; make work or employ for a particular purpose or for its inherent or natural purpose: “use your head!”; “we use Spanish at home”; “use plastic bags to store food”; “use a computer”)

For the same 27 cases, the “splitter” selected these 4 distinct senses:

1. practice, apply, use – (avail oneself to; “use care when going down the stairs”; “use your common sense”)
2. use – (seek or achieve an end by using to one’s advantage; “use one’s influential friends to get jobs”; “use one’s good connections”)
3. use – (take or consume (regularly); “She uses drugs rarely”)
4. use, expend – (use up, consume fully)

Each of the senses selected by the splitter is in fact a more specific subsense of the one sense chosen by the lumper. But the sense distinctions involve two independent parameters. Senses 1 and 2 impose specific selectional restrictions on their direct objects (behavioral or mental attributes, or persons or abstract entities that can serve as the means to an end or goal, respectively). Senses 3 and 4 have specific aspectual properties (habitual and completive, respectively). Both types of meaning component can co-occur in a single usage; the aspectual property of the verb is independent of its selectional restriction. An entity can be used for its inherent purpose (most general sense), and be fully used up (sense 4) or used regularly (sense 3). Many contexts leave the aspectual properties of the verb unclear and do not specify whether something is used up or used regularly. To account for occurrences where otherwise distinct meanings may overlap, annotation must be allowed using not only any one of these senses alone, but also a combination of them.

8.3 World knowledge

Sometimes, the meaning of the verb depends not only on its syntactic use or the semantics of its arguments, but on world knowledge. For example, two

senses of the verb *face* differ on whether the person or entity faced usually presents an adversarial situation or not. *Face your partner* may be part of an instruction for a dance; *face your boss/the media/reporters/judge* conveys the sense of an embarrassing confrontation. The verb’s objects here carry subtle implications that affect the taggers’ choice between a neutral sense of the verb and a “confrontation” sense.

Similarly, a sentence like “He faced Smith in the courtroom” differs crucially from “He faced the teacher in the classroom/at the beginning of this scene.” The context in the form of the adjunct provides a human with the intended reading, but it would be very hard to make an appropriate distinction between *courtroom* and *classroom* in a dictionary or to train an automatic system to make this distinction. A multiple tag must be allowed.

8.4 Vague contexts

Even if we imagine a “perfect” tagger, it would not be the case that each occurrence of a word in a context can be uniquely identified with a dictionary sense, even by a human tagger. Consider the following context for the verb *see*:

- As for dancing – holy mackerel, he ought to see the gypsies in Jerez; they danced on the sand till your blood got hot and danced with them.

It is unclear (even from the larger context), whether *see* here has the sense glossed in WordNet as “to perceive by sight” or the one referring to watching a show. The interpretation hinges on whether the gypsies dance informally or as part of a performance. The human taggers each chose a different sense.

The senses here are distinct and cannot be clustered. But when the context does not allow for an unequivocal annotation, double tagging is permitted.

9 A complete grouping of *call*

The analysis of the inter-tagger annotation differences provides some criteria that can guide us towards the clustering of semantically related senses in WordNet. Keeping these types of relations in mind we did a careful study of *call* which resulted in the following WordNet 1.7 grouping. We have given a general gloss for each group to indicate what is shared by all the members, and also specified what each sense in the group uniquely contributes over and above the general gloss. For instance, in Group 1 the senses are distinguished by the semantic categorization of the kind of thing being ascribed. Is it an attribute or property, or a name? In Group 2, Senses 2 and 3 are distinguished by the type of

instrument used to make the call. Sense 2, communicate by phone, is more specific than Sense 3, which could generalize to radios or other devices. In Group 6 the senses are distinguished by whether the sound is produced naturally or artificially. In general, semantically related verb senses may be abstract rather than concrete, may have different entailments (is anything created, are there secondary predications, are there additional intentions or goals), or may involve different types of participants, (animal versus human, animacy versus inanimacy, different instrument types). The more explicit we can be about the criteria we use for both grouping senses and distinguishing them, the more consistent we can be in tagging and in the categorizations of new usages.

- Group 1 (1,3,12,19,22) - ascribe an attribute (19) or a label (1,3,22) to another. (12, *call the roll*, does not fit as neatly, but it is very specialized, perhaps idiomatic, and difficult to fit anywhere).
- Group 2 (2,3) - communicate by just phone (2), or phone or radio, etc. (3).
- Group 3 (4,7,9) - request an individual (4) or group (7) meeting, or a formal appearance, i.e., jury duty (9).
- Group 4 (5,16) - utter a loud cry (5), or loudly cry something (16).
- Group 5 (6,23) - a visit by a person (6) or a ship (23).
- Group 6 (15,26) - characteristic call of a bird or animal (15) or its imitation (26)
- Group 7 (18,27) - challenge the veracity of, usually *call on*.
- Group 8 (20,25) - initiate a financial change of state with respect to loans (20) or bonds (25).
- Individual senses: 17 - predict an outcome and 28 - wake up (could possibly be grouped with 2 or 4?)
- Specialized domain individual senses: 8 - call a strike, 11 - call a game (as in stop the game), 21 - call a dance, 24 - call a trump in a card game.

10 Automatic word sense disambiguation

In addition to examining annotation discrepancies between human taggers, we looked at the output of an automatic sense tagger on some of the same data, evaluating performance using fine-grained senses as well as the more coarse-grained groups, similarly to the SENSEVAL three-way evaluation discussed above. This provides another point of comparison

for inter-annotator agreement, where one of the annotators is a computer and the other is human³, shedding light on verb meaning from a computational perspective.

10.1 System description

We developed an automatic word sense disambiguation system that uses a maximum entropy framework to combine linguistic contextual features from corpus instances of each verb to be tagged. Under the maximum entropy framework (Berger et al., 1996), evidence from different features can be combined with no assumptions of feature independence. The automatic tagger estimates the conditional probability that a word has sense x given that it occurs in context y , where y is a conjunction of features. The estimated probability is derived from feature weights which are determined automatically from training data so as to produce a probability distribution that has maximum entropy, or equivalently, is maximally noncommittal, under the constraint that it is consistent with observed evidence.

Following work by Chodorow, Leacock and Miller (2000), we divided the possible model features into topical and local contextual features. The local features for a verb w in a particular sentence tend to look only within the smallest clause containing w . They include:

- the word w , the part of speech of w , and whether or not the sentence is passive
- whether there is a sentential complement, subject, direct object, or indirect object
- the words (if any) in the positions of subject, direct object, indirect object, particle, prepositional complement (and its object)
- WordNet synsets and hypernyms for the nouns⁴ appearing in the positions above
- a Named Entity tag (PERSON, ORGANIZATION, LOCATION) for proper nouns appearing in the positions above
- words at positions -2, -1, +1, +2, relative to w

In addition, topical features looked for the presence of keywords (determined automatically from training data) occurring *anywhere* in the sentence. Extra-sentential context was not examined, since human taggers said this was almost never necessary to determine the correct sense of a verb.

³In this experiment, the human annotator was the one who was the “splitter.”

⁴Our use of WordNet was very naive since these nouns were not disambiguated in any way. This clearly limits the benefits of our particular usage of word classes. A more sophisticated use of noun classes might try to do some sense disambiguation of the nouns first.

10.2 System performance

The maximum entropy WSD system's performance on the verbs from the evaluation data for the first SENSEVAL exercise rivaled that of the best-performing systems, with an average accuracy of over 72%. The sense inventory and tagging for SENSEVAL had been done by a team of highly trained lexicographers, resulting in inter-annotator agreement figures of 88% to 100%, with almost all agreement figures above 95%. Given that our WSD system performed so well on this data, we then tested it on 19 of the 30 verbs that had been tagged in preparation for SENSEVAL2, to see the effect of using different sense inventories and corpora.

We allowed the option of using just local features with or without the WordNet noun class information, or combining both local and topical features. In addition, we experimented to see whether tagging the corpus instances with the sense group number instead of the fine-grained senses would improve performance. Table 1 shows the variant of the system that produced the best performance on each verb, as well as its accuracy on unseen test data.

We found that grouping the senses for "call" significantly improved performance over evaluation with respect to fine-grained senses; the system achieved 75% accuracy when trained and tested on grouped senses and 72.5% accuracy when trained and tested on fine-grained senses, but evaluated according to whether the proposed and correct sense tags were in the same sense group; this is in contrast to the 47.5% accuracy achieved when the system was trained, tested, and evaluated on fine-grained senses. When trained and evaluated on fine-grained senses, the system got 22 out of 41 instances wrong, but 10 of the "incorrect" instances were tagged with senses that were actually in the same group as the correct sense. The most common confusion (7 instances) were between the following senses (all in the same sense group):

- name, call – (assign a specified proper name to; "They named their son David"; "The new school was named after the famous Civil Rights leader")
- call – (ascribe a quality to or give a name that reflects a quality; "He called me a bastard"; "She called her children lazy and ungrateful")
- call – (consider or regard as being; "I would not call her beautiful")
- address call – (greet, as with a prescribed form, title, or name; "He always addresses me with 'Sir'"; "Call me Mister"; "She calls him by his first name")

For each of these senses, the maximum entropy model assigned very high weights to the feature "has

sentential complement" (in the top 65 of over 800 features in the model).

11 Conclusion

The traditional dictionary model of meaning representation, with its discrete senses, is clearly not adequate for semantic annotation by human taggers, and there is little reason to assume that automatic systems can map dictionary senses of polysemous words onto tokens in a corpus in a one-to-one fashion. Results from an annotation task performed by two trained humans show high rates of disagreements. But these annotation results can inform the development of dictionaries that are intended for use in automatic word sense identification tasks. Many natural occurrences of polysemous words are embedded in underspecified contexts and could correspond to several of the more specific senses. Annotators and automatic systems need the option to select a cluster of specific senses or a single, broader sense, where specific meaning nuances are contained but hidden. Sense clustering, already present in much of WordNet's verb component, can be much enhanced and guided by the analysis of inter-annotator disagreements and the development of explicit sense distinction criteria that such an analysis provides. Comparing the annotations of an automatic system with that of a human tagger provides further insights into the nature of verb meanings from a computational perspective.

12 Acknowledgments

This work has been supported by DARPA grant N66001-00-1-8915 and NSF AWARD 9800658 at the University of Pennsylvania. We would also like to thank Scott Cotton and Joseph Rosenzweig for system infrastructure support.

References

- Adam Berger, Steven A. della Pietra, and Vincent J. della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.
- Martin Chodorow, Claudia Leacock, and George A. Miller. 2000. A topical/local classifier for word sense identification. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.
- Christiane Fellbaum, Joachim Grabowski, and Shari Landes. 1997. Analysis of a hand-tagging task. In *Proceedings of the ACL/Siglex workshop*, Somerset, NJ.
- Christiane Fellbaum, Joachim Grabowski, and Shari Landes. 1998. Performance and confidence in a semantic annotation task. In Christiane Fellbaum, editor, *WordNet*. MIT Press.
- Christiane Fellbaum. 1998a. The organization of verbs and verb concepts in a semantic net. In

verb	corpus size	senses	% correct	best variant
call	198	22	75.0	-wn+topic+group
carry	198	33	50.0	-wn-topic+group
collaborate	12	2	100.0	-wn-topic-group
develop	202	17	73.2	+wn-topic-group +groupeval
draw	123	39	16.0	-wn+topic-group
dress	55	9	63.6	-wn-topic-group +groupeval
drift	27	8	50.0	-wn-topic-group
drive	126	14	73.1	+wn-topic+group
face	279	7	83.9	-wn-topic-group +groupeval
keep	200	33	57.5	-wn-topic-group
match	82	7	47.1	-wn+topic-group +groupeval
pull	182	34	51.3	-wn-topic-group +groupeval
strike	74	23	60.0	+wn+topic-group +groupeval
train	28	6	83.3	-wn-topic-group +groupeval
treat	82	5	82.4	+wn-topic+group
use	224	8	75.6	-wn+topic+group
wander	21	4	100.0	+wn+topic-group +groupeval
wash	37	14	62.5	-wn-topic-group +groupeval
work	180	24	63.9	-wn-topic+group

Table 1: Precision of maximum entropy model on some Penn TreeBank verbs. Corpus size is the total number of instances in both training and test data (in a ratio of 4:1); number of senses is counted by the number of senses actually appearing in the tagged corpus, rather than in the sense inventory; +wn indicates that the best-performing variant of the system included WordNet noun classes instead of just lexical features (-wn); +topic indicates use of topical features instead of just local features (-topic); +group indicates training and testing on verbs tagged with sense group numbers instead of fine-grained sense numbers (-group). For variants not trained on sense groups, +groupeval indicates that performance improved if measured by the number of words tagged with senses that were in the same sense group.

- Patrick Saint-Dizier, editor, *Predicative Forms in Natural Language and in Lexical Knowledge Bases*. Dordrecht.
- Christiane Fellbaum. 1998b. *WordNet*. MIT Press, Cambridge, MA.
- Adam Kilgarriff and Martha Palmer. 2000. Introduction to the special issue on senseval. *Computers and the Humanities*, 34(1-2):1–13, April. Special Issue on SENSEVAL.
- Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.
- Shari Landes, Claudia Leacock, and Randee Teng. 1998. Building a semantic concordance of english. In Christiane Fellbaum, editor, *WordNet*. MIT Press.
- Beth Levin. 1993. *English Verb Classes and Alternation: A Preliminary Investigation*. The University of Chicago Press.
- Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english. *Computational Linguistics*, 19(2):313–330.
- George A. Miller and Patricia M. Gildea. 1987. How children learn words. *Scientific American*, pages 94–99, September.
- George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the Human Language Technology Workshop*, Princeton, NJ.
- Martha Palmer, Hoa Dang, and Joseph Rosenzweig. 2000. Sense tagging the penn treebank. In *Proceedings of the Second Language Resources and Evaluation Conference*, Athens, Greece.
- James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.