Making Senses: Bootstrapping Sense-tagged Lists of Semantically-Related Words

Nancy Ide

Department of Computer Science, Vassar College, Poughkeepsie, New York USA
ide@cs.vassar.edu

Abstract. The work described in this paper was originally motivated by the need to map verbs associated with FrameNet 1.2 frames to appropriate WordNet 2.0 senses. As the work evolved, it became apparent that the developed method was applicable for a number of other tasks, including assignment of WordNet senses to word lists used in attitude and opinion analysis, and collapsing WordNet senses into coarser-grained groupings. We describe the method for mapping FrameNet lexical units to WordNet senses and demonstrate its applicability to these additional tasks. We conclude with a general discussion of the viability of using this method with automatically sense-tagged data.

1 Introduction

Lists of semantically-related words and phrases are heavily used in many automatic language processing tasks. A common use of such lists in recent work is in attitude or opinion analysis, where words indicative of a given semantic orientation—often, "positive" or negative" polarity—are detected to classify documents such as movie and product reviews as more or less favorable ([1], [2], [3]). Approaches include simple term counting [4] as well as training machine learning algorithms to classify documents. In machine learning approaches, semantically-related words and phrases are often used as a part of the feature set (e.g., [2], [3], [5]. NLP tasks such as event recognition also typically rely on lists of semantically-related verbs coupled with frames or patterns that are used to identify participants, etc. (e.g., [6] [7]).

Largely due to the recent upsurge in work on attitude and opinion analysis, numerous lists of semantically-related words have been made available within the language processing community. The lists are compiled using a variety of means, including extraction from existing resources such as lexicons, thesauri, and pre-compiled content category lists such as the General Inquirer [8]; automated extraction [2] [3]; and manual production; and often include hundreds or even thousands of words.

Whatever the source, available lists of semantically-related words do not identify the sense of the included items, despite the fact that many of the words are highly

polysemous.[1] As a result, work relying on such lists identifies word occurrences that may not represent the phenomenon in question. Sense-tagged lists of words could significantly increase the accuracy of pattern-recognition and learning algorithms, if the data is also sense-tagged. For the moment, we put aside the issue of (accurately) sense-tagging corpora, and return to it in Section 8.

The work described in this paper was originally motivated by the need to map verbs associated with FrameNet 1.2 frames to appropriate WordNet 2.0 senses. As the work evolved, it became apparent that the developed method was applicable for a number of other tasks, including assignment of WordNet senses to word lists used in attitude and opinion analysis and collapsing WordNet senses into coarser-grained groupings. In the sections that follow, we describe our method and demonstrate its applicability to these additional tasks. We conclude with a general discussion of the viability of using our sense-tagged lists with automatically sense-disambiguated data.

2 Background

The work reported here was undertaken in the context of the FDR/Pearl Harbor Project[2], which is enhancing a range of image, sound, video and textual data drawn from the *Franklin D. Roosevelt Library and Digital Archives*. The project is undertaking the encoding, annotation, and multi-modal linkage of a portion of the collection, and development of a web-based interface that enables exploitation of sophisticated data mining techniques. The project focuses on a collection of 1,446 internal administration documents concerned with US-Japanese relations between 1931 and 1941, including memoranda of conversations, letters, diplomatic correspondence, intelligence reports, and economic reports. The corpus has been annotated for a wide range of entities and linguistic phenomena, and all words have been automatically tagged[3] with WordNet2.0 senses. To support retrieval, an ontology including persons, locations, roles, organizations, and events and other entities specific to our data (ships, treaties, etc.) has been created, by extending and refining SUMO and MILO categories such as *government and military organizations* and *people related to organizations*. All annotation and ontology development in the project has been accomplished using the GATE (General Architecture for Text Engineering) system [9] developed at the University of Sheffield.

Historical research on Japanese-American relations in the ten years prior to the bombing of Pearl Harbor focuses on the nature of the relationship between representatives of the two countries. In particular, historians and political scientists are interested in the interplay of the dialogue between the two countries and how it conveys attitudes such as power and control vs. submission, hostility vs. friendliness and openness, cooperation vs. non-cooperation, etc., not only at a given time, but as these attitudes varied during interactions over the ten-year pre-war period. The FDR Project

---

[1] The General Inquirer includes sense tags using a sense inventory developed by the project; however, only words appearing in more than one sense in the same list are tagged.

[2] Supported by U.S. National Science Foundation grant ITR-0218997.

[3] WordNet::SenseRelate (all words) [11] was used to provide WordNet sense annotations.

is therefore concerned with identifying evidence of such attitudes in the wording of documents in the corpus, and attributing this information to the appropriate person or entity. Because a large portion of the documents in the collection consists of so-called "memoranda of conversations", many consist of near-transcriptions of meetings between Japanese and US officials.[4] We have therefore focused on identifying *communication events*, down to the level of the utterance (e.g., "X asked that…") and apply attitude-recognition procedures to each utterance attributed to a given speaker. Historians may thus request a synopsis of, for example, the attitudes conveyed by a Japanese official in conversations with the US Secretary of State over the ten-year period, and consider their development and change.

3 The Word List Problem

Annotation of the FDR document collection as described in Section 2 required automatic identification of semantically-related words signifying events and attitudes, followed by the application of pattern-recognition rules to extract contextual information (including role-fillers in the case of event recognition, polarity influencers [10] for attitude analysis, etc.). To detect lexical items indicative of a given attitude, we need to compile lists of words, in particular for specific affective categories such as "hostility", "cooperation", "power/control", etc. For events, we require lists of verbs associated with a given event type, in particular, different categories of communication verbs (e.g., questioning, persuasion, reporting, etc.).

Rather than starting from scratch, we gathered information from available resources such as the General Inquirer, FrameNet[12], VerbNet [13], WordNet, and Levin's verb list [14], and various additional lists compiled by individual researchers[5], with the goal of merging as much of the information provided in these resources as possible. Existing resources from which word lists can be extracted come in several forms:

1. Flat word lists including no additional information. Some sources provide lists for relatively specific categories, such as "hostility", "military", etc, as, for example, one finds in the General Inquirer; others—especially lists that are becoming increasingly available within the NLP community—provide lists of words deemed to denote a positive or negative attitude. Typically, words in such lists are unlemmatized and may contain several inflectional variants of the same lexical item.
2. Word lists including a measure of relevance/relatedness, such as lists of positive/negative words that provide an associated measure of degree, or lists providing measures of semantic similarity (e.g., [15]).

[4] Note that the memoranda represent a complex communication event, in which, for example, Secretary Welles reports to FDR what the Japanese Ambassador said and how Secretary Hull replied. We make no judgment concerning the degree to which reports of, say, the Japanese Ambassador's wording may have been colored by the reporter; our job is to simply provide the information to the historian and allow him to draw his or her own conclusions.

[5] Our thanks go to Diana Inkpen, David Nadeau, and Maite Taboada for providing their lists of positive and negative words, and to Janyce Wiebe for her lists of subjective elements.

3. Computational lexicons such as WordNet, FrameNet, and VerbNet. Depending on their intended use, computational lexicons contain additional syntactic and/or semantic information, such as definitions and examples and verb argument structure, and therefore, extracting a simple list of related words typically demands some degree of processing. Both WordNet and FrameNet additionally support their lexicons with hierarchical ontologies that provide several levels of increasingly general semantic classes associated with each word.

Merging information from these various resources is not a trivial matter, especially when the semantic categories involved go beyond broad categories such as "positive" and "negative". First, semantic categories in different resources are determined using different criteria, and as a result, a straightforward mapping of categories is often impossible. This is particularly true for FrameNet and WordNet, whose categories and ontologies are, for practical purposes, virtually disjoint; furthermore, FrameNet's ontology is shallow, often including only two or three levels, whereas WordNet's is substantially deeper. VerbNet assigns FrameNet frames and WordNet senses to some of its entries, thus making a merge simpler; and the association of a given lexical item with a WordNet sense enables a mapping of words assigned to a given FrameNet frame to a WordNet sense. However, VerbNet's coverage is relatively scant, especially for FrameNet categories, and it therefore provides very little information to link the three resources.

Extracting information from available resources can be a time-consuming task, and information from the different sources is various and occasionally conficting. Furthermore, the results do not provide comprehensive lexical coverage, and they are in some cases inaccurate. Therefore, rather than attempting to merge existing resources, we developed a method to generate accurate, extensive sense-tagged word lists for lexical units associated with FrameNet frames, and generalized the method to produce sense-tagged lists for additional semantic categories.

3.1 Bootstrapping from FrameNet

FrameNet is creating a database of semantic frames, in which case roles dictated by the semantics of the lexical units (LUs) associated with the frame are specified. FrameNet provides a shallow inheritance hierarchy of frames corresponding to semantic categories; for example, the frame *complaining* inherits from *statement*, which inherits from the general category *communication*.[6] Each frame is associated with a set of "frame-evoking" LUs consisting of word lemmas. Different senses of a lexical unit are defined on the basis of its association with different frames.

---

[6] FrameNet also specifies a "using" relation among frames in cases where a particular frame makes reference in a general way to the structure of a more abstract frame; for example, the *judgment_communication* frame uses the *judgment* and *statement* frames, although it does not directly inherit from them. For the purpose of constructing a general hierarchy of semantic categories, we treat the "using" relation as inheritance. For a fuller explanation of FrameNet's architecture and rationale, see [12].

We use the set of LUs associated with each frame as defined by FrameNet as a starting point to develop more comprehensive word lists representing semantic categories. To be maximally useful for our application in the FDR project, this demanded several enhancements:

1. Extension of the lists of lexical units to provide more comprehensive coverage of words representative of a given category. As FrameNet is in the process of development, the number of lexical units associated with a given frame varies considerably, and coverage is incomplete.
2. Sense-tagging lexical units in order to eliminate "false hits" in our analysis. Many of the lexical items associated with the various FrameNet frames are highly polysemous, and identifying un-sense-tagged occurrences leads to considerable noise in our analysis. Because our corpus has been annotated with WordNet senses, it is desirable to associate each lexical unit in a given frame with a WordNet sense or senses.
3. Refinement of the FrameNet categories. FrameNet associates lexical units with a given frame on the basis of frame semantics, which often leads to word lists containing, for example, words with both positive and negative connotations that correspond to the same general semantic category. For example, the lexical units for the frame *judgment_communication* include not only "acclaim", "commend", and "praise" but also "condemn" and "denounce". In addition, the possibility for more subtle semantic distinctions is apparent in many of the lists; in the same *judgment_communication* frame, we can isolate distinguishable semantic groupings of lexical items, such as "deride", "mock", and "ridicule"; "belittle", "disparage", "denigrate"; "blast" and "slam"; etc.

We developed a procedure to address issues (1) and (2) and applied a clustering algorithm to the results in order to accomplish (3), as described in the following sections.

3.2 Mapping FrameNet Lexical Units to WordNet Senses

Attempts to automatically or semi-automatically derive lists of semantically-related words and phrases has a long history in NLP, starting with a series of projects during the 1990's (e.g., [15] [16] [17]) using similarity of distributional patterns in large corpora and clustering techniques. However, the use of distributional patterns in corpora has one major drawback: words may follow similar patterns not only because of semantic similarities, but also syntactic or pragmatic ones. As a result, many of the lists compiled using this strategy contain words that are not necessarily related semantically; for example, "achieve", "frighten", "invite", and "penalize" are among the top-rated words in Lin's publicly-available similarity list for the word "encourage". For our purposes, where *semantic* similarity is the focus, corpus evidence is therefore not an ideal source.

WordNet::Similarity. WordNet::Similarity(WNS) [18] is a freely available package that includes six measures of similarity and three measures of relatedness that use information in WordNet, including links and path lengths for the various WordNet relations (synonymy, hyperonymy, etc.) and overlap among glosses and examples, shortest WordNet path length, information content, depth in the WordNet is-a hierarchy, and semantic density, to determine the degree to which two words are alike. The various measures are described and compared in [18]. Given a pair of words, WordNet::Similarity returns their most similar WordNet senses together with a numeric value reflecting their degree of similarity. The measure of similarity can also be determined for a pair of WordNet sense-tagged words, one sense-tagged word and one untagged word, etc., or for all senses of the input pair.

We use WNS to determine the "most similar" WordNet senses for each of the LUs associated with a particular FrameNet frame. To do this, we create a set of pairs $P_F = LU_F \times LU_F$, where LU_F is the set of lexical units associated with FrameNet frame**,** and feed P_F to WNS. The result set R_F includes the most similar senses for each pair of words and a measure of their similarity. The hypothesis is that since the words in the pair sets have been associated with a specific FrameNet frame, they should be mutually disambiguating, and the most appropriate WordNet sense for the frame can be determined.

Preliminary experimentation with all nine of the similarity and relatedness measures provided in WNS confirmed that the *lesk* measure provided the most accurate results, which corresponds to the determination based on similar experiments reported in [REF]. The *lesk* measure [19] assigns a relatedness score by scoring overlaps between glosses of two senses and senses of other words that are directly linked to them in WordNet, according to a user-chosen set of relation "pairs" that specify which WordNet relations determine the score (for example, overlaps between synsets of the two words, overlaps in the gloss of the first word and example text of the second, etc.), any of which may be optionally weighted. WNS provides a default relation set for use with the *lesk* measure that determines the relatedness score based on overlaps among all possible relation pairs, a total of 88 in all.

We devised a reduced relation set that includes the following relation pairs: example - example, gloss - gloss, hypernym - hypernym, hypernym - hyponym, hyponym - hypernym, hyponym - hyponym, synset - example, synset – gloss, and synset - synset. Greatest weight (0.7) was given to synset overlaps, and additional weight (0.5) was given to overlaps in example texts, glosses, and synset overlaps with examples and glosses. The rationale for this choice was to focus on synonymy (same concept) and *is-a* relations (more/less general expression of the same concept). We also determined that gloss and example overlaps, as well as synset overlaps with glosses and overlaps, are highly reliable indicators of relatedness, often capturing commonalities that are not otherwise direct or explicit (e.g., the synset for urge#v#3 includes "inspire", which appears in the gloss for encourage#v#2, "inspire with confidence").

Computing Sense Lists. We determine sense-tagged lists for LUs associated with FrameNet categories using WNS's *lesk* procedure. A *suresense* ss_{w_i} is identified for each word $w_i \in LU_F$ when any one of the following holds:

1. w_i has more than one sense and $freq(s_{w_i}) = 1$
2. w_i has only one sense and $simscore(s_{w_i}) > .2$
3. $freq(s_{w_i}) > T_{freq}$, and $tsim(s_{w_i}) > T_{sim}$

where T_{freq} and T_{sim} are user-defined threshold values in range 0-1 for the frequency and total similarity values, respectively.

The frequency score $freq(s_{w_i})$ is defined as

$$freq(s_{w_i}) = \frac{\sum_{s_{w_i} \in R_F} s_{w_i}}{pc_F} . \quad (1)$$

where pc_F is the number of pairs (w_i,w_j) in P_F for some w_i—i.e., $size(LU_F)$ - 1. When $freq(s_{w_i}) = 1$, s_{w_i} has been returned as the most similar sense for every pair $(w_i,w_j) \in P_F$.

$Tsim(s_{w_i})$ is the sum of *lesk score* (*ls*) values returned by WNS reflecting the degree of relationship between sense s_{w_i} and all other senses s_{w_j}:

$$Tsim(s_{w_i}) = \sum_{\substack{j=1 \\ j \neq i}}^{pc_F} ls(s_{w_i}, s_{w_j}) . \quad (2)$$

$simscore(s_{w_i})$ is defined as:

$$simscore(s_{w_i}) = \frac{e^{simscore(s_{w_i})} - e^{-simscore(s_{w_i})}}{e^{simscore(s_{w_i})} + e^{-simscore(s_{w_i})}} . \quad (3)$$

This scales *simscore* to a value between 0 and 1, and eliminates the impact of the size of the LU sets.

Condition 2 above handles the rare instance when the appropriate *sense* of an LU does not appear in WordNet; e.g., "grill", which is an LU in the category *questioning*, appears in a single sense in WordNet ("cook over a grill"). However, because *simscore*(grill#v#1) is only .04, it does not exceed the threshold, and therefore this sense is not added to the set.

We use the following algorithm to create a list of suresenses for LUs associated with a FrameNet category:

ALGORITHM A:

1. Compute SS_{LU_F}, the set of suresenses ss_{w_i} for lexical units in LU_F[7], using the method described above
2. Generate a new set of pairs P' from $SS_{LU_F} \times LU_F$
3. Compute $SS_{P'}$

[7] In the course of computing the similarity measures, LUs that do not appear in WordNet are eliminated.

Note that some LUs in LU_F may not be assigned a suresense. At the same time, more than one sense for a given word may qualify as a suresense. Step 1 identifies highly-related senses from the original un-tagged list of LUs; since some words are not assigned a sense at this point, in Step 2 relatedness is computed using the set of *sense-tagged* words identified in Step 1 coupled with every un-tagged word in the original set. This strategy both provides better information for computing relatedness for the as-yet unassigned words, and may identify additional senses for words that were tagged in Step 1.

Manual evaluation determined that suresense sets compiled using this method are highly accurate, but that in a few cases, "noise" was introduced into the set in the form of in appropriate sense assignments. This occurred in situations where, for example, two or more words in an LU share a second meaning, which was then introduced into the suresense set. For example, the LUs for the *reasoning* frame include "demonstrate" and "show", which share not only the appropriate sense of proving or establishing something, but also the sense of exhibiting to an audience. Therefore, to maximize the accuracy of our results, we modified the algorithm to include additional information derived from other sources, including WordNet::SenseRelate::WordToSet[8] (SR) and a "master" sense list extracted from VerbNet (VN) and FnWnVerbMap 1.0[9] (VM) [20].

SR determines the relatedness of a given word w to a set of words. The results list all WordNet senses of w together with a relatedness score, sorted from highest to lowest score. We fed SR lists of word-set pairs for each LU_F consisting of (1) each $w_i \in LU_F$, coupled with (2) the set of all $w_j \in LU_F$, $i \neq j$. SR uses WNS to compute the relatedness scores and provides the same choice of similarity measures; we have used the similarity measure and relation set as described above for WNS. We derive two additional "suresense" sets from SR's output:

1. SR_{top}, the sense of each LU determined to be most similar to the remaining words in a given LU_F; and
2. SR_{cutoff}, the senses of each LU with a relatedness score above a pre-determined cutoff value.

Note that because SR computes a single, overall score for each sense based on its relatedness to all other LUs in a given frame, the results from WNS described above and results from SR provide somewhat different results; the correlation of results computed using WNS above and each of the two sets computed from SR is .8. We can characterize results in SR_{top} as highly precise but with low recall; whereas SR_{cutoff} and the *SS* sets computed using WNS have slightly lower precision but better recall.

To address this problem, we created another suresense set by combining the WordNet senses assigned to words in a given LU_F that is also tagged in VN and/or VM into a single set *V*. VN includes slightly over 4000 words, each of which is manually assigned a WordNet sense or senses; FrameNet frames are assigned to only a fraction of entries. VM provides a semi-automatically-assigned WordNet sense or senses for every verb lexical unit in FrameNet 1.2. Originally, we hoped to use VM as a gold standard against which to evaluate our results, but we discovered that the assigned

[8] http://www.d.umn.edu/~tpederse/senserelate.html

[9] Available at http://lit.csci.unt.edu/~rada/downloads/FnWnVerbMap/FnWnVerbMap1.0.tar.gz

signed senses in VM are often incomplete; that is, many senses that are viable alternatives are not included. Also, the identified senses are occasionally incorrect. More importantly, comparison of WordNet sense assignments for words that are tagged in both VN and VM show an agreement level of only 27%, which is no doubt a results of the well known problem with WordNet sense assignments, wherein distinctions are generally regarded as too fine-grained for most NLP applications and problematic for humans to distinguish. Collapsing WordNet senses to produce a sense list more appropriate for NLP applications has been proposed ([21]; see also [22]); in fact, because our method identifies multiple senses for each LU, it potentially identifies at least some senses of a given LU that can be collapsed.

Using information extracted from the various resources, the final set of senses for a FrameNet frame *F* is determined as follows:

ALGORITHM B:

1. Compute SS_{LU_F}, the set of suresenses ss_{w_i} for lexical units in LU_F
2. Set $SS'_{LU_F} = SS_{LU_F} \cap SR_{top_F} \cap SR_{cutoff_F} \cap V_F$
3. Generate a new set of pairs P' from $SS'_{LU_F} \times LU_F$
4. Compute $SS'_{P'}$

3.3 Evaluation

To test the accuracy of our method, we computed sense-tagged lists of LUs for 28 FrameNet categories that are classified as sub-types of *communication*. The number of LUs in the frames ranges from two to 58; the average number of LUs per frame is 12. In this experiment, T_{freq} was set to .3 and T_{sim} was set to .9. The algorithm identified at least one sense for 95% of the LUs and assigned an average of 1.35 senses to each.

Manual evaluation of sense-tagging is a notoriously problematic task, and even among human annotators there is typically no more than 80% agreement on the WordNet sense to be assigned to a given word in context. Our task here is somewhat simplified, for several reasons:

1. Sense assignments are not evaluated for words in context, but rather in terms of the word's association with a FrameNet category and in relation to the set of LUs associated with that category.
2. Multiple senses can be assigned to a given LU; there is no attempt to identify a unique sense assignment.
3. The task consists of validating the assignments produced by the algorithm, rather than assigning a sense or senses to LUs and comparing the results to the automatically-produced set.

Two undergraduate Cognitive Science majors with a background in linguistics performed manual validation of the sense assignments produced by our algorithm. Both verified that 100% of the senses assigned by the algorithm were appropriate for the FrameNet category with which they are associated. We note that given our method of determining the sense assignments, it is possible that some appropriate senses are not included, especially additional senses for words for which at least one sense has been

identified. We address this issue in section 7. However, for our purposes it is preferable to maximize precision at the expense of recall, since the resulting suresense sets are used to augment the lists, as described in the following section.

4 Augmenting the Lists

The highly accurate suresense sets produced using the algorithm described in the previous section provide the base from which to generate additional sense-tagged words in order to augment the FrameNet LU sets. To do this, we apply the following algorithm:

ALGORITHM C:

1. Add synsets for all $s_{w_i} \in SS_{P'_F}$ to $SS_{P'_F}$
2. Generate $HYPE_F$, the set of hypernyms for all $s_{w_i} \in SS_{P'_F}$
3. Generate a new set of pairs P_{HYPE} from $SS_{P'_F} \times HYPE_F$
4. Compute SS_{HYPE} from P_{HYPE}
5. Generate $HYPO_F$, the set of hyponyms for all $s_{w_i} \in SS_{P'_F}$
6. Generate a new set of pairs P_{HYPO} from $SS_{P'_F} \times HYPO_F$
7. Compute SS_{HYPO} from P_{HYPO}
8. Generate a new set $U_F = SS_{P'_F} \cup SS_{HYPE} \cup SS_{HYPO}$
9. Add synsets for all $s_{w_i} \in U_F$ to U_F

Hyponym and hypernym sets occasionally include words that are less related to the category than desirable. For example, the set of hyponyms for sense 3 of "permit", which is included in the category *grant_permission*, includes sense 4 of "pay" ("bear a cost or penalty in recompense for some action"). Verifying the hypernym set against the previously-generated set of suresenses for the category eliminates this and other less related words, including "take_lying_down" and "stand_for". Hypernym sets often include general concepts, such as sense 2 of move ("cause to move, both in a concrete and in an abstract sense"), which is in the hypernym set for the category *attempt_suasion*; verification against the suresense set also eliminates very general senses, as they are typically related only weakly to those suresenses for which they are not the hypernym.

A modified version of algorithm A is used to verify hypernym and hyponym sets, in which frequency scores--which tend to be near or equal to 1 in every case--are ignored; in these cases, relatedness is determined solely on the basis of *simscore*.

Algorithm C could be repeated one or several times to further augment the lists, although we have not tested this option: iterative addition of hypernyms and hyponyms could introduce increasing noise, and accuracy of the sets may degrade.

Lists of un-sense-tagged words from other sources can also be run against the suresense sets to augment the suresense sets. For example, we have run the list of verbs appearing in the FDR corpus (with the exception of "do", "be", "have", and modal verbs) against the suresense sets for the FrameNet communication categories, in order to ensure full coverage of our lexicon. Here, because the vast majority of the words

in the list are unrelated to the suresense set, we increased the threshold for eliminating words with one sense given in Algorithm A, step 2, to .5.

Similarity lists for each set of LUs associated with a FrameNet communication categories were also extracted from Lin's data and run against the suresense sets in order to extract additional word senses appropriate for the categories. The results were judged to be about 90% accurate overall, somewhat less than the accuracy rate for the suresense sets, presumably because the words in Lin's lists had already been selected for similarity to the target word by using contextual information. The failures typically involve words that have no sense that is relatively synonymous to the target word or with opposite polarity (e.g., (e.g., "engage" and "frighten" in relation to "encourage"). We are currently experimenting with Lin's lists in order to improve accuracy, before adding the results to the FrameNet suresense sets.

Our sense-tagged lists of words for each of the FrameNet communication categories is available at http://www.cs.vassar.edu/~ide/FnWnSenselists. Both the original suresense lists, including only the FrameNet LUs, and the augmented lists including synsets, hypernyms, and hyponyms, are available on the website.

5 Refining Categories

The LUs associated with FrameNet frames often fall into semantic sub-categories that are not isolated in FrameNet. The similarity measures produced by WNS can be exploited to produce a similarity matrix, which in turn can be used to identify semantic sub-groups of LUs for a given frame via clustering.

We applied a clustering algorithm using *weighted arithmetic average*, which assigns a weight to the distance between samples S1 (in A) and S2 (in B) of $(1/2)^G$, where G is the sum of the nesting levels (number of enclosing clusters) of S1 and S2, which reduces the influence of groups of similar samples on the clustering process.

Table 1 shows the clustering results for the *judgment_communication* suresenses, obtained by "pruning" the cluster tree at edges with a weighted distance > .85 according to the algorithm. Each column contains word senses (WordNet2.0 sense number appended) included in one of the pruned sub-clusters.[10] The results identify intuitively sensible semantic groupings, and correctly isolate the positive and negative senses. Further pruning within a sub-cluster could yield even finer semantic distinctions; for example, the "acclaim" sub-cluster includes two sub-clusters: *acclaim1*, *extol1*, *laud1*, and *commend4*; and *commend1*, *praise1*, and *cite2*.

[10] Senses *damn1*, *harangue1*, and *criticize1*, each of which appears in a cluster by itself, are not included in the table.

Table 1. Clustering results for *judgment_communication*

ACCLAIM	DENIGRATE	BELITTLE	CONDEMN	CHARGE	ACCUSE	RIDICULE
acclaim1	denigrate1	belittle2	condemn1	accuse2	accuse1	deride1
extol1	deprecate2	disparage1	decry1	charge2	denigrate2	ridicule1
laud1	execrate2	reprehend1	excoriate1	recriminate1		gibe2
commend4		censure1	deprecate1			scoff1
commend1		denounce1				mock1
praise1		remonstrate3				scoff2
cite2		blame2				remonstrate2
		castigate1				

6 Generating Word Lists for Attitude Analysis

The procedure outlined in sections 3 and 4 can be applied to generate sense-tagged word lists for use in tasks such as attitude analysis. Here, the user provides an initial list of "seed" words to replace the FrameNet lists of LUs. An obvious source of seed words is the categorized word lists in the General Inquirer; however, the GI lists are extensive, some including over 1000 words, and often the semantic range of items in a given category is quite broad. In addition, the lists contain words in various parts of speech as well inflectional variants, the latter of which are not usable to retrieve information from WordNet.

To test the viability of creating sense-tagged lists of words for attitude analysis, we created lists of seed words by intersecting lemmas from a 150,000 word sub-corpus of the FDR data with the GI word lists for the categories "hostile", "power/cooperation", "submit", "weak", and "strong". A seed suresense list is created using Algorithm B, replacing *LU* with the list of seed words, and using only *SS*, SR_{top}, and SR_{cutoff} in step 2. In step 3, the seed suresense list is run against the remaining words in the original list. Note that in processing the FrameNet categories, only verb senses were considered for inclusion, whereas here, senses of a given word as a noun, verb, adjective, or adverb are considered if they exist. Following the application of Algorithm B, the resulting suresense sets were split into subsets according to part of speech, and each subset was individually augmented by applying Algorithm C.

The resulting lists, averaging about 80 senses in length, were judged to be 98% accurate by the student validators. Our sense-tagged lists for GI categories are available at http://www.cs.vassar.edu/~ide/GIsenselists/.

Lists of "positive" and "negative" words are commonly used in opinion analysis, and several extensive lists are in circulation within the research community. The General Inquirer also provides lists of words with positive and negative connotations. Such lists include words from a broad range of semantic categories, since the only criteria for inclusion is their mutual polarity. For this reason it was not clear that the suresense procedure would be as effective in identifying relevant word senses. However, experimentation has so far shown that the results are better than anticipated. Inappropriate suresenses typically involve words whose inclusion in the list is questionable—for example, words like "colony" and "desire" in the GI's list of negatives—although the procedure often fails to identify a suresense in such cases. We continue to experiment with producing suresense sets from polarity word lists; in

particular, we are experimenting with threshold values as well as breaking the sure-sense sets into sub-sets based on clustering. Results will be posted on the website as they become available.

7 WordNet Sense Ambiguation

For the purposes of NLP, many WordNet senses can be viewed as identical—indeed, many of the problems with manual sense-tagging and word sense disambiguation that use WordNet senses arise from the at-times imperceptible shades of meaning that WordNet distinguishes. In an attempt to determine WordNet senses for the same word that can be, for practical purposes, collapsed, we applied WNS to determine the similarity among all *senses* of a given word. In this experiment, the full set of relations provided in WNS was used, rather than the reduced set applied in the experiments reported above.

Similarity scores between senses of the same word computed by WNS proved to be extremely low, which is not surprising given that the criterion for distinguishing senses in WordNet is membership in different synsets, one of the main criteria by which similarity is measured by WNS. Clustering based on the similarity matrix for the scores, however, indicated that the method holds some promise for collapsing WordNet senses: for example, the topology of the cluster tree for the verb "press" is given in Figure 1. The tree topology reflects sense grouping that are intuitively obvious, especially the close association of senses 10 and 7, 4 and 5, and 3 and 8; even more striking is the division between senses concerned with physical pressure and senses in which the use of "press" is abstract or metaphorical (excepting sense 11).[11] The clusters can be separated on the basis of varying distance cutoffs to create sense grouping at different levels of granularity.

We are currently experimenting with clustering WNS similarity measures to "ambiguate" WordNet senses, together with the incorporation of multiple suresenses for the same word identified by the algorithm. Based on the results so far, the method shows considerable promise for creating sense lists that are more usable for NLP.

[11] Note that the "press" example was randomly chosen, and is typical of the results we have seen so far in our experiments.

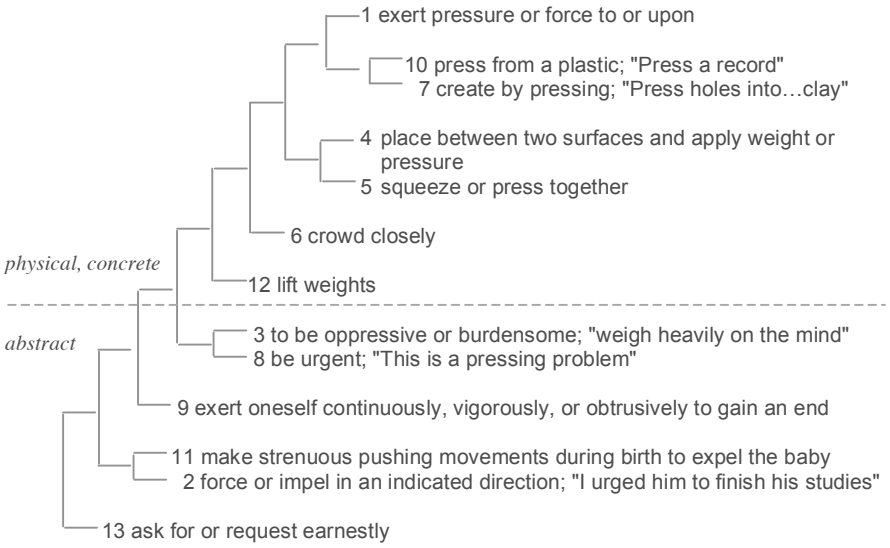

Fig. 1. Cluster tree topology for WordNet senses of "press"

8 Summary

The methods outlined in this paper demonstrate that similarity measures and clustering are effective methods for creating sense-tagged word lists for semantic analysis. Sense-tagged word lists are a valuable resource in their own right, but to be used in applications such as attitude an opinion analysis and event recognition, the corpus under analysis must be sense-tagged as well. This would seem to be a drawback to using sense-tagged word lists to accomplish these and other corpus-analytic tasks, since automatic disambiguation algorithms currently achieve, at best, about 80% accuracy, and the cost of hand-tagging or hand-validating sense-tags in even a modestly-sized corpus is prohibitive. However, automatic sense-tagging may soon cease to be the insurmountable problem it has traditionally been thought to be, if a "common" sense inventory is agreed upon within the community, and if the accuracy of entirely automatic sense disambiguation software is improved. We believe that both of these obstacles can be addressed, at least to a workable degree, with a single solution: adopt a set of senses derived from WordNet, in which senses that can be regarded as identical (at least, for the purposes of NLP) are collapsed.

We contend that a substantial portion of the "errors" generated using current disambiguation systems would be eliminated if WordNet senses were grouped into coarser-grained semantic categories. This is not a novel idea; the word sense disambiguation (WSD) community has long been aware that WordNet senses pose significant problems for the field because of their granularity, both for evaluating WSD systems and achieving agreement among human annotators. At the same time, the

community is aware that homograph recognition—which can be automatically achieved with high rates of accuracy—is not enough for NLP. What has been recognized less frequently is that for most NLP applications, something not far from homograph-level distinction *is* adequate, and that we have some good clues concerning what those distinctions are and how to identify them from a variety of sources, including cross-lingual information, psycholinguistic experiments, and, possibly, clustering "similar" WordNet senses as described in section 7, above (see [22] for a fuller discussion of this point). If the community can turn its attention to creating a usable sense inventory for NLP, then there is a future for automatic WSD.

In summary, we have within our means ways to significantly improve accuracy rates for WSD systems in the not-too-distant future. If this is done, it will in turn open the door to the use of systems such as WordNet::SenseRelate to perform accurate automatic WSD, and to the exploitation of these results in tasks such as attitude analysis and event recognition.

References

1. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (2002) 417-424
2. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of Empirical Methods for Natural Language Processing (2002) 79-86
3. Pang, B., Lee, L., Vaithyanathan, S.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (2004) 271-278
4. Turney, P., Littman, M.: Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus, National Research Council, Institute for Information Technology, Technical Report ERB-1094 (NRC #44929) (2002)
5. Wiebe, J., Riloff, E.: Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Gelbukh, A. (ed.): Computational Linguistics and Intelligent Text Processing, 6th International Conference Proceedings. Lecture Notes in Computer Science, Vol. 3406. Springer, Berlin Heidelberg New York (2005) 486-497
6. Vargas-Vera, M., Celjuska, D.: Event Recognition on News Stories and Semi-Automatic Population of an Ontology. In Shadbolt, N., O'Hara, K. (eds.): Selected Advanced Knowledge Technology Papers (2004) 159-163
7. Alphonse, E., Aubin, S., Bessières, P., Bisson, G., Hamon, T., Lagarrigue, S., Nazarenko, A., Manine, A.-P., Nédellec, C., Vetah, M., Poibeau, T., Weissenbacher, D.: Event-Based Information Extraction for the Biomedical Domain: The Caderige Project. In Collier, N., Ruch, P., Nazarenko, A. (eds.): Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. (2004)
8. General Inquirer: http://www.wjh.harvard.edu/~inquirer/ (2000)
9. Cunningham, H.: GATE, a General Architecture for Text Engineering. Computers and the Humanities, Vol. 36 (2002) 223-254
10. Polanya, L., Zaenen, A.: Contextual valence shifters. Proceedings of the AAAI Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI technical report SS-04-07) (2004) 106-111
11. http://www.d.umn.edu/~tpederse/senserelate.html

12. Ruppenhofer, J., Ellsworth, M., Petruck, M.,, Johnson, C.: FrameNet: Theory and Practice. On-line publication at http://framenet.icsi.berkeley.edu/ (2005)
13. Kipper, K., Dang, H.T., Palmer, M.: Class-based construction of a verb lexicon. Proceedings of Seventeenth National Conference on Artificial Intelligence (2000) 691-696
14. Levin, B.: English Verb Classes and Alternation: A Preliminary Investigation. University of Chicago Press (1993)
15. Lin, D.: Automatic Retrieval and Clustering of Similar Words. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (1998) 768-773
16. Pereira, F., Tishby, N., Lee, L.: Distributional Clustering of English Words. Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (1993) 183-190
17. Lee, L.: Measures of distributional similarity. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (1999) 25-332
18. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity - Measuring the Relatedness of Concepts. Proceedings of the Nineteenth National Conference on Artificial Intelligence (2004) 1024-1025
19. Banerjee, S., Pedersen, T.: Extended gloss overlaps as a measure of semantic relatedness. Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (2003) 805–810
20. Shi, L., Mihalcea, R.: Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. Computational Linguistics and Intelligent Text Processing, 6th International Conference Proceedings. Lecture Notes in Computer Science, Vol. 3406. Springer, Berlin Heidelberg New York (2005) 100-111
21. Palmer, M., Ng, H.T., Dang, H. Evaluation. In Agirre, E., Edmonds, P. (eds.): Word Sense Disambiguation: Algorithms and Applications. Springer, Berlin Heidelberg New York (forthcoming)
22. Ide, N., Wilks, Y.: Making Sense About Sense. In Agirre, E., Edmonds, P. (eds.): Word Sense Disambiguation: Algorithms and Applications. Springer, Berlin Heidelberg New York (forthcoming)