Corpus-based Techniques for Word Sense Disambiguation

Gina-Anne Levow

January 23, 1997

1 Introduction

1.1 The Problem and Why it Matters

Consider the task of building a speech-to-speech translation system. One significant problem confronting the designer is the absence of a one-to-one mapping from word sounds to text strings to word meanings. The following examples reveal the ubiquity of this problem . In a highly homophonous language like Chinese the single sound sequence 'shi' maps to 56 different characters, each of which in turn has at least one meaning. In English, not only are there many text strings with context-dependent pronunciations and meanings ("record": the verb - "re-córd" and the noun "ré-cord"), but there are also many words like "bank" which have only one pronunciation but take on numerous meanings. For example, "bank" can be used as "the bank of a river", "bank account", and "bank a plane". The most extreme form of this ambiguity appears in pronous like "it", which take meaning only by reference to another element of the discourse. These mismatches multiply across languages , where in English the word "sentence" has two meanings, but in French, these meanings must be realized as two different words *peine*, in the criminal sense, and *phrase* in the grammatical. So, performing dictation,speech recognition, machine translation, or Web-search document retrieval, all require the ability to correctly select word senses.

1.2 Roadmap

In the course of this paper, I will describe a set of three techniques which claim to use corpus-based statistical methods to try to solve the problem of word sense disambiguation. As an introduction I will describe briefly some of the range of approaches which have been applied to this task. I will further indicate the limitations on scalability of these techniques which motivated the shift to automatic machine-learning techniques trained of large corpora for this task. Then I will embark on a careful assessment of the three techniques described in the examination papers addressing the issues below:

- Description of the techniques and some preliminary results

- Operation on an illustrative example
- Important Sources of Disambiguation Information
- Contrasts between wide and narrow windows of context
- Limitations of Surface Statistics
- Different definitions of similarity
- New Senses and the Importance of Generalization
- Lack of a Model

1.3 A Variety of Attempts

Early word sense disambiguation (WSD) approaches emphasized working from large amounts of hand-coded knowledge. Scripts, as developed by Schank & Abelson [31],[7] encoded topic-based world knowledge about word uses in typical instances of common activities, such as going to a restaurant. Others used a mix of syntactic and semantic constraints embodied in parser rules or semantic frames [38], while a number of researchers collected lists of words strongly associated with a sense of a word or synonyms and looked for matches between the lists and the words near a target word. Others picked a variety of these constraints and combined them either in complex pieces of computer code [33], [18] or complex networks of spreading activation [5],[37]. Many of these techniques performed accurate sense selection in tests, but the task of manually encoding all of the information to handle any significant portion of English was far too large. These techniques were also not robust enough handle imperfect input.

To ease this problem, many turned to precoded knowledge sources, such as machine-readable dictionaries (MRDS), thesauri, or semantic networks [19],[30]. However, Lesk's (1986) [23] seminal approach which relied on word overlap between the current use of the word in a sentence and the dictionary definition text to identify the correct sense in the context illustrates the difficulty of using precoded knowledge source. The technique depends crucially on similarity in wording between the two texts. Negotiating coincidence of word choice is something which, as Brennan [1] notes, is actually a key part of negotiating the form of a dialogue for the participants. In order to surmount the problem of lexical choice in MRDS, it became necessary for their users to embark of the huge task learning to understand the dictionary itself in order to extract useful information from it. [40],[14] Dictionaries also suffered from limits in domain specific coverage and in the ability to adapt to the introduction of new words. In rejection of these limitations, a number of researchers decided that instead of trying to convert a precoded knowledge base to their needs, they would simply build their own from corpus data in the world and replaced pre-coded or hand-coded information about word sense and uses with learned information derived from statistics over large corpora.

1.4 Why Learn from Large Corpora?

As noted above, a major challenge and stumbling block for many WSD systems has been the ability to handle a large number of words in a wide variety of contexts. In part the problem is exacerbated by hand-coding, where the designer must produce individually all the necessary information about disambiguation. It is this issue, among others, which has inspired a number of researchers to turn to the collection and exploitation of large corpora (of text or speech) to help extend the coverage of existing models or even to bootstrap or train the design of new ones. An early proponent of the corpus-based approach in linguistics was Z. Harris (1968)[16] who tried to extract groups from text corpora which corresponded to syntactic categories. Techniques which merged machine-learning techniques and large training corpora have proven successful in areas ranging from finding faces in crowded scenes (Sung & Poggio 1995)[34] to speech recognition systems [28] and part-of-speech taggers [2]. In the remainder of this paper, we will explore the issues raised by techniques of this class which will, we hope, shed further light on WSD needs and the ability of corpus-based techniques to meet them.

1.4.1 A Caveat about Statistics

However, let us first interject a word of warning about the ability of these techniques to succeed in the task we have set for them. Scalability is a tremendous challenge for statistical and corpus-based approaches. In speech recognition, there are approximately 625 triphone contexts which can appear for English and which the system must be trained to recognize. This task can be achieved with greater than 95% word accuracy on a 1,000-word speaker-independent recognition task with thousands of sentences of recorded speech from more than 109 talkers. (Rabiner & Juang 1993)[28]. Part-of-speech taggers likewise have 64 part-of-speech tags to assign and work on short sequences of parts of speech, often just pairs, to make their decisions. These systems can be trained to 97% accuracy on corpora of 1.5 million words. (Brill et al. 1991)[2] In contrast, to do word sense tagging there are more than 55,000 words and 74,000 senses [40] in even a learner's dictionary, much less something like the OED with hundreds of thousands of senses. The problem is compounded by the fact that constraints on word sense can easily come from as far away as 40 words, say the previous sentence in the Wall Street Journal. Even storing only the pairwise word-word co-occurrence matrix for this task is beyond the capacity of most contemporary workstations. This problem is far larger than those to which large corpus statistic-based techniques have been applied with such high success rates, even before considering the size of the corpus required to exhibit all those interactions which could be useful for word sense disambiguation. More abstract relations which allow us to make more useful generalizations than word-word co-occurrence relations seem crucial to making the problem of word sense disambiguation tractable.

2 Three Ways to Pick Senses

2.1 Schutze: Context Vector Representations

In his paper "Word Space", Schutze describes a technique which builds a vector space representation of word meanings. Specifically, he begins by bootstrapping his representations by building a co-occurrence matrix for 5,000 frequent yet informative letter fourgrams. These are simply four character sequences which occur in a large corpus of New York Times news stories. The most frequent 300 are excluded as too frequent to be informative and include sequences such as ' the', ' and', common affixes, and function words. Essentially, these sequences provide a controlled vocabulary, since the actual word-word full co-occurrence matrix for this corpus would be far too large for modern computational techniques. The co-occurrence matrix value of A_{ij} is incremented each time w_i occurs within 200 fourgrams to the left of w_j. A singular value decomposition is then performed, allowing each letter fourgram to be represented as a vector of 97 real values. Word context vectors are, in turn, built by summing and normalizing the vectors of all fourgrams within a 1001 character window of the encoded word. A sum of the context vectors of all observed instances of a word in the corpus form that word's confusion.

To apply this representation to WSD, an automatic clustering algorithm operates on the context vectors of all observed instances of the target word. The distance metric is vector distance within the 97 dimensional space. Each cluster is then (hand-)labelled with a sense tag as appropriate. For each new word occurrence to be disambiguated, a context vector is constructed as before and is assigned the sense tag of the closest cluster. On a task disambiguating instances of 10 well-known ambiguous words, mostly in 2-way sense distinctions, the system achieved an average accuracy of greater than 92%.

2.2 Resnik: Interpreting Clusters with a Semantic Network

The next paper, by Resnik, basically provides an extension to Schutze's "word space", or any other distributional clustering algorithm for word sets, to eliminate the need for hand-labelling of sense clusters. It starts from the observation that when presented with a cluster of words, people naturally and automatically interpret them as a coherent group and assign a sense to polysemous words that suits the meaning of the group. For instance, in a cluster such as "attorney, counsel, court, trial, judge", cited in his paper and extracted by (Brown *et al.* 1992)[3], readers naturally assign the legal senses of 'counsel', 'trial', and 'judge', under the influence of the surrounding words.

To perform this assignment automatically, Resnik uses the IS-A hierarchy of WordNet [25], specifically the noun component of this carefully hand-crafted semantic network, to assess the similarity between word senses. For each node in the network, a measure of informativeness is computed as follows: $I(C) = -\log(\frac{\sum_{n \in words(C)} count(n)}{N})$, where N is the size of the corpus. This measure corresponds to the log of the inverse frequency of the concept and all of its child words in the corpus. Infrequent concepts are presumed to be more informative in a representative corpus. This approach to similarity in WordNet tries to avoid the pitfalls of path length distance metrics where concepts high in the hierarchy may be very close in terms of path length, but may be such abstract concepts as to be very weak indicators of relatedness.

To perform disambiguation within a word cluster, for each pair of words in the cluster, do the following:

1. Get the most informative common ancestor of any senses of the two words and its I measure.

2. Add this I value to all senses of the two words subsumed by this concept.

Then assign the highest scoring sense to each instance, after accumulating values over all pairs of words. The author presented a variety of example labellings as a qualitative evaluation, and also conducted a formal evaluation in comparison with human labellers, using Roget's Thesaurus categories as clusters and labelling 23 senses of 'line' in those contexts, in which the system approached the human level of performance, with man achieving 67% accuracy and the machine 60% on this difficult task.

2.3 Yarowsky: Making Senses with Decision Lists

The third and final approach, described by Yarowsky in "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods", proceeds under two main assumptions: that there is one sense per discourse and one sense per collocation. The first means that in a given text document an ambiguous word will probably only appear in one of its senses. In addition to statistical evidence which Yarowsky cites, this observation makes intuitive sense. A given topic generally selects one sense of a word, and also co-operative speakers and writers do not try to confuse their partners by intentionally mixing senses of a word. The second observation means that if a word w appears in sense $s1$ in some collocation or "word configuration" much more frequently than in $s2$, when w appears elsewhere in the same collocation, it will probably have the same sense $s1$. Again there is intuitive support for this claim in that when we are asked to explain the meaning of a word, we often use a short characteristic phrase which includes the word, as in "river bank" vs. "bank account".

Yarowsky describes an algorithm that, given a small set of sense-tagged "seed" instances, can build a decision list procedure to label a full corpus and disambiguate new instances of a word. The seed instances are examples of the word in each of its senses in sentence context. Each instance is examined by the algorithm to find collocations of different forms, such as "word-to-the-left", "word-to-the-right", "word in ±k words", etc. Such a collocation is deemed informative if for all currently labelled instances $abs(\log \frac{Pr(sense_1|collocation)}{Pr(sense_2|collocation)})$ is large; that is, if one sense appears in the specific collocation much more often than the other. For instance, for "plant", in the biological and factory senses, "manufacturing plant" is highly informative, but "the plant" would not be. These rules are ordered from most to least informative based on this maximum likelihood estimator and are placed in a decision list. The algorithm then loops over any remaining untagged instances labelling as many as possible with the new decision list, and then using the contexts of the newly labelled instances as sources for new collocations which can be inserted into the decision list. The one sense per discourse constraint can also be applied either at each iteration or when no new instances in the residual can be labelled. This constraint can (re-)label all instances of a word in an article where a majority sense had emerged. Once the decision list has been trained, disambiguation proceeds by presenting the target instance in its context to the decision list for labelling by the highest ranked informant. Like Schutze's, this algorithm was evaluated on a set of infamous pairwise ambiguous words and achieved an accuracy of 95%.

Let us quickly try to put these results into perspective. Miller et al. [26] ran three simple experiments to establish baseline performance measures for statistical techniques. First they observe that while 82% of the words in WordNet have only one sense, in a typical corpus only 27% of the words have a single sense. Thus, the need to disambiguate senses clearly arises very frequently. These researchers applied two simple statistical heuristics for sense selection: sense frequency from a labelled corpus and co-occurrence within a sentence. Both simple metrics achieved an accuracy of 70% on labelling all senses in the corpus, reflecting a 60% accuracy on words with more than one sense. Although applying our three techniques to a full corpus labelling task would be instructive, even these basic results illustrate that the algorithms are faring well above the baseline.

2.4 Example: "Plant" Disambiguation

The two text segments below were taken from Web pages and will illustrate the operation of each of the three disambiguation techniques described above. We will assume that the training stage has already completed and we will disambiguate the uses of plant in each passage.

There are more kinds of plants and animals in the rainforests than anywhere else on Earth. Over half of the millions of known species of plants and animals live in the rainforest. Many are found nowhere else. There are even plants and animals in the rainforest that we have not yet discovered.
Many of the plants from the rainforest are used for medicines by both people in the forest and hospitals throughout the world. One-fourth of the drugs that you can buy at the drugstore have products that come from the rainforest. Medicines that fight heart disease and treat cancer patients are made from rainforest plants. Aspirin originally came from the rainforest. A flower called the rosy periwinkle helps treat children with Leukemia (a kind of cancer).

Text 1

The Paulus company was founded in 1931. Since those days the product range has been the subject of constant expansions and is brought up continously to correspond with the state of the art. We're engineering, manufacturing and commissioning worldwide ready-to-run plants packed with our comprehensive know-how. Our Product Range includes pneumatic conveying systems for carbon, carbide, sand, lime and many others. We use reagent injection in molten metal for the purpose of desulphurising and recarburising. We also build dust extraction and filter plants in dry and wet implementations, sand reclamation plants for the foundry industry, and movable sandrecovery machines. Our industrial automation for the iron/steel, foundry and chemical industries, including Switchgears, PLC/DCS and MMI systems are the best. We will provide special industrial designs to meet your requirements upon request.

Text 2

Let us start with Schutze's approach. Running some quick statistics on the two texts we note first that each is smaller than the 1001 character window over which the algorithm builds a context vector. Also, we observe that there are only 7 instances of content words which appear in both, three of which are the target word itself. Clearly the context vectors formed by the fourgrams in these contexts will be far apart in "word space", and will be assigned to different clusters as appropriate. A scan of the words indicates typical contexts associated with the biological and manufacturing senses of "plant", leading to correct labeling.

Next let us apply Resnik's approach. Since we do not have a cluster for plant, let us construct one for each passage from the nouns which appear there. For text 1, that gives us "plants, animals, rainforests, species, medicines, people, forest, ..."; for text 2, we have "product,range, systems, carbon, ..., metal, purpose, ...,filter,extraction, industry,machines, plants, automation, ...". Even without a corpus to generate informativeness scores for the subsumers, it is clear that the best subsumer for plants in cluster from text 1 will be the biological sense, and in text 2 the manufacturing sense.

Finally, let us try the decision list provided in [11] to assign the senses to plant. For each text let us consider the first occurrence of the word plants. Going down the (initial) decision list, we match on "animal (within $\pm$2-10 words) $\rightarrow$ sense A" for text 1, and "manufacturing (within $\pm$2-10 words)" $\rightarrow$ sense B. It is interesting to note that both of these rules are placed much lower in the final decision list, and none of the top decision rules in that list match in our texts. Also, curiously, although "manufacturing" does occur close to "plants" in text 2, this meaning is surely not the one anticipated in the training data.

All three approaches easily accomplish this simple sense disambiguation task.

3 Information They Try to Use

All three of the current papers use different sets of co-occurrence statistics, sometimes augmented with other linguistic or world knowledge, to try to capture some of the types of disambiguating information described by Hirst (1987) as listed below. These are:

- Knowledge of Context: this refers to global topic, such as the information captured by Schank and others in scripts.
- Association with nearby words
- Syntactic disambiguation cues, which include subcategorization
- Selectional restrictions between ambiguous words
- Inference and World Knowledge[1]

Psycholinguists also recognize that frequency of sense can play a role in sense selection.

3.1 Finding the Topic in a Window

3.1.1 A Needle in a Haystack?

Schutze captures information about letter fourgrams within a 1001 character window of the target word as a representation of the context in which the word appears. These vectors are treated as an unordered bag of words. Thus any information which is encoded in ordering or adjacency relations between the context words and the target word is lost. The technique is unable to make use of most cues based on syntactic structure or selectional restrictions and close word associations. Thus what the vector space model is most effectively equipped to capture is some notion of a general knowledge of the topic. The use of this type of information is reflected in the word groupings which he cites as examples, and the relative success of the system at disambiguating noun rather than verb senses. Statistical studies such as those by Yarowsky in "One Sense Per Collocation" [42] indicate that while noun sense disambiguation can be aided by word co-occurrences up to hundreds of words away, useful information for selecting verb and adjective senses falls off rapidly with distance from the target word. As

[1] [18], p. 80

frame theory identifies, and as used in Preference Semantics [39], the verb in a sentence interacts strongly with its objects, and adjectives likewise are tightly associated with the nouns they modify. These associations provide the best evidence for their senses. Thus, by focusing on wide-window co-occurrences Schutze ignores the best sources of disambiguating information for verbs, resulting in the weakest reported result of 69% for the verb sense of 'train', which could easily be bettered by a simple part-of-speech tagger.

Resnik's approach again relies heavily on a knowledge of topic. The distributional clusters he uses are, as with Schutze, unordered bags of words which are presumed to be related under some topic, which when inferred will provide evidence for the most appropriate sense. WordNet, in turn, provides both a fairly simple knowledge base on which to perform some inference and an additional source of information about context. The secondary use of the corpus provides a more motivated way of measuring similarity of objects within the knowledge base. However, since WordNet separates parts of speech strictly into different hierarchies, and the similarity metric operates only within the one dimension of the IS-A hierarchy, the disambiguation technique is restricted to operating on nouns alone and can derive no information from either syntax or other possible relations between verbs or adjectives and the current target noun.

Yarowsky's technique allows for the potential incorporation of a wide range of different information sources through the decision list mechanism. The current implementation described here, though, makes use of a smaller subset of the available disambiguators. In particular, most of the discriminants are short-range collocates of the form - x target-word, target-word x, or x ± 10 words from the target-word. These features explicitly capture information on nearby words, and implicitly try to access some bits of syntactic information, selectional restriction, and topic. The "one sense per discourse" constraint more explicitly captures topical information on a very broad, article-level scale.

Clearly, no single source of information will be able to disambiguate every utterance, so those techniques which rely heavily on, for instance, "global" topical information will be unsuccessful in cases where the context of the target word is narrow or underspecified , as is often the case in information retrieval queries. Disambiguation will also fail for words which appear in a wide variety of topical contexts, such as common verbs, which conversely may be easily identified by local collocational, syntactic or selectional cues, or even by frequency information.

A comparison of three different "wide-window" statistically-based techniques was conducted by Leacock *et al* [22]. Specifically, they compared the performance of a Bayesian classifier, a context vector, and a neural network, trained on the same corpus with the same context window of the current and preceding sentence. On a two-way sense disambiguation task, all achieved greater than 90% accuracy. For three and six- way sense selection tasks, performance for all systems quickly dropped to around 70%. These results are consistent with other attempts to use "wide-window" context schemes with other machine learning techniques such as simulated annealing [6], indicating that most of these machine learning techniques, while differing in implementation, are similar in power and in disambiguation ability when given the same information on which to operate. It also suggests the limitations of the pair-wise disambiguation task as a metric for evaluating the techniques; clearly, even a small increase in the number of senses dramatically changes the difficulty of the task.

Particularly revealing is an additional study by Leacock *et al* [22] in which human subjects were given the same disambiguation tasks to perform with three different types of information available: first the two original context sentences, then the context sentences with the words all randomly ordered, and finally only the randomly ordered content words. While the subjects performed almost perfectly on the first task, once ordering information and original syntactic structure were removed, the performance of human subjects became comparable to that of the computer systems, falling to an error rate of almost 32%. These results indicate two significant points: first that the systems are doing as well as possible with the limited input they are given, and second, that crucial information is lost when the context is treated simply as an unordered "bag of words."

4 Surface Co-occurrence, not Meaningful Disambiguators

In the preceding discussion we described how these corpus-based algorithms selected different categories of disambiguation information from the environment in which the target word occurred. Specifically, we noted that use of wide-window techniques which concentrate on the use of bags of content words can capture some constraints associated with topic while narrow-window approaches can capture information of the type encoded in selectional restrictions and word associations. However, as tempting as it may be for these authors to claim that they are using "topic" or "global context" or "syntactic cues", it is important to remember that these techniques are really capturing statistical regularities about the sentences in and near which these words occur. While topic, selectional restrictions, syntax, etc. interact in the mind of the writer/speaker to cause the sentences to take the form they do, statistics of word co-occurrence capture *only* the surface regularities. There is no distinction between significant regularities - in this case, those co-occurrences which are directly the result of the interaction of the word sense and, say, one feature of its environment - and unimportant regularities. As an example of the latter, DeMarcken(1995)[12] in a corpus-based lexical learning task notes that "scratching her nose" appears in his corpus much more frequently than expected, i.e. it is a statistically significant regularity; however, it isn't a meaningful regularity in the same sense as the fact that "kicking the bucket" can be expected to occur much more often in a corpus than expected since it is an idiom.

In this case, it is important to differentiate between regularities that have impact on the sense of the word and those which do not; it would also be profitable to differentiate among sources of regularities, since , just as all co-occurrences are not equally relevant , not all sources of information are equally relevant. We will find examples in these techniques of the inability to treat meaningful and coincidental regularities differently, and also the lack of weighting between information sources.

Schutze's "word space" provides a number of insights into what is really being learned by corpus-based techniques. These issues strongly impact Resnik's work as well, since he acts as a consumer of these distributional clusters. Let us consider some illustrative examples of randomly selected nearest neighbors in "word space."

4.1 People Interpret Clusters, Algorithms Don't

For "burglar" the 10 related items are: "burglars thief rob mugging stray robbing lookout chase crate thieves". The majority of these neighbors look "reasonable" - burglar, mugging, rob, robbing, even lookout - logically are related to each other in the context of criminal activities. First we should note that we have interpreted this set in such a way as to make it coherent. Secondly, consider that one of the top 10 scorers for 'burglar' is 'crate.' The inclusion of this word, to us, is clearly anomalous, and we may even be willing to accept this cluster as "good" since "it only got one wrong." However, to the system, and any others which make use of this output, this entry is as valid as all the others on the list. If it were differentiable, it would not have been included in the first place.

4.2 Learning the Corpus, not the Sense

Now consider a less successful cluster, the one for "Ste." (Sainte) which is as follows: "dry oyster whisky hot filling rolls lean float bottle ice". First, observe that none of these words has anything to do with the lexical meaning of "Ste." as a beatified woman. In fact, the system here has simply learned that in its huge corpus of words "Ste." appears in the context of these other words; Schutze notes that these contexts are in reference to the river Ste. Marguerite. A post-hoc labelling algorithm such as Resnik's would likely treat the cluster as a bunch of food-related terms, which do not relate to "Ste." in any way. Since his algorithm assumes that distributional clusters are semantically meaningful, they are interpreted as coherent even when they are not.

4.3 Learning Nothing, by Asking the Wrong question

Finally, consider the case of "keeping" with neighbors "hoping bring wiping could some would other here rest have". Even for a person knowing that this is a cluster for "keeping" , it is difficult, if not impossible, to find any relation either among the words as a group, or even between any of the words and "keeping" . Schutze remarks that it is difficult for his technique to handle words which appear in a wide variety of contexts. This example highlights the need for multiple knowledge sources and the need to apply different information to different tasks. Even a hand-labelled description of topic for each instance in which this word appears would not handle the two senses so clearly differentiated by "keeping up" vs. "in keeping with," for example. This problem with words appearing in a wide variety of contexts is of particular concern since, distributionally, a relatively small proportion of the words in a language is used very frequently, and these words in fact are the most polysemous and therefore in need of disambiguation.

4.4 All collocations are not created equal

Turning to Yarowsky's approach, we find a technique that can possibly incorporate any number of different knowledge sources, but here again we find little distinction between knowledge sources, and no differentiation between relevant and irrelevant regularities. There is a broad-brush distinction between sense determination based on discourse and that of more localist collocational information. However, there is no difference for the algorithm between localist co-occurrences that arise due to topical constraint as in "astronomer" and "star", those that occur in common colloquialisms, "the North Star", and those from selectional restrictions such as "married a star", which forces the "famous person" interpretation. Clearly, these constraints can interact, and we need a method which allows us to model these interactions at least sufficiently to choose between senses based on competing constraints that are weighted by more than relative frequency of co-occurrence in collocation.

5 Measures of Similarity

Let us next consider the definitions of similarity defined and used by these approaches. What criteria do they use to determine whether two instances of a single text or phoneme string or two different words are similar? This question greatly influences the tasks for which the techniques can be used and also how easily the approach can generalize to new words, additional senses, and new domains. Each method builds up its own notion of similarity from training data in cooperation with any precoded knowledge.

5.1 Vector Distances in Word Space

Schutze has a straightforward definition of similarity which is a natural outgrowth of his choice of representation. Since he builds a high dimensional vector space, he used vector distance within this space to assess similarity. Once the representation is computed by training, it is simple to compute similarities. Since word instances are represented by vectors derived from "wide-window" co-occurrence information, we can say that things appearing in similar context are similar.

5.2 WordNet IS-A Hierarchy: Similarity in 2-D

Resnik combines two components to establish his notion of similarity. The first is simply co-occurrence within a cluster. The second is derived with the WordNet IS-A hierarchy as sharing an informative subsumer, so two senses are, intuitively, more similar if they can be found to share an ancestor deep in the WordNet tree, preferably one which also occurs infrequently. This metric is actually quite restrictive. Specifically it depends on both the exact structure of the IS-A hierarchy and the idiosyncracies of the training corpus. To illustrate this problem, consider again the example Resnik himself used to introduce his stance - "attorney, counsel, trial, court, judge". Curiously, the algorithm can not assign the correct sense to trial, even though it has very strong semantic associations with the other members of the cluster. This is because the other list elements all fall within the "person" hierarchy in WordNet , while none of the senses of 'trial' does. Thus the most informative subsumer is the empty root node.

The system will, conversely, label "lookout" in the "burglar" cluster correctly but for the "wrong" reason - here this is the only person sense available and the cluster provides many supporters for the "person" interpretation. This raises the dangerous possibility of unanticipated interactions between coincidental similarities in clusters and the structure of WordNet. More generally, the problem here is that meaning can be viewed as multi-dimensional, reminiscent of the feature vectors of Katz & Fodor -style semantics[20], and structurally similar to Schutze's word space. The WordNet hierarchy, however, only forms IS-A links along a certain dimension. When the cluster is related along the same dimension as the WordNet hierarchy, correct disambiguation is promoted; when the cluster is related along a different, perhaps orthogonal, dimension, coincidental support is given to senses.

5.3 Who Needs Similarity When We have Difference?

Finally, we have Yarowsky's approach. Here, we find no general notion of similarity at all. Two words may have collocations in common, but the decision lists are built independently for each set of senses to be partitioned. Since such a wide variety of surface collocations is used, it would be difficult to say that two words have a similar sense and thus should share a decision list. The algorithm can, of course, be instructed to find any collocational features which can discriminate between members of pairs. Looking assiduously for differences, as this algorithm does, will not lead one to identify similarities. The programmer may be able to identify similar classes, as in the case of accent restoration for pairs of Spanish verb tenses, and apply the same decision list to all members of the class, but this process requires hand-coding to prevent inadvertently including word-specific cues in the class-level decision list.

6 Key to Generalization: Recognizing Similarity

It is important to be able identify and interpret new words and senses. One also needs to be able to easily extend the system to handle new ambiguities and tasks. Let us consider how each approach would respond, having encountered the two noun sense pairs of "river bank"/ "financial bank" and "manufacturing plant"/ "living plant", to the verb sense of each ambiguous word.

Schutze would, as usual, compute a context vector for the new instance and compare it under the vector distance metric to other established senses and other words. In fact, even unseen words have a defined place in the word space, according to their "wide-window" context. No new representations need to be created and, while the verb sense of "plant" would likely be viewed, reasonably, as fairly close to the biological sense of the word, the verb form of "bank" would probably be easily identifiable as a new sense. Having a general similarity metric allows a system to flexibly adapt to new words or senses of a known word. This particular metric is weakened by the problems of the "wide-window" bag of words definition of context discussed above, but could still prove useful.

Resnik, by using WordNet as a filter for all actions, gains a lot of information for free, avoiding the need to build a representation as both Schutze and Yarowsky must. However, the converse problem is that WordNet forms a closed semantic representation and compresses the many dimensions of word meaning into its hierarchy. A contextually appropriate sense which is novel to WordNet, as in the verb sense of "bank" (since Resnik uses only the noun hierarchy and hierarchies by definition have zero cross-similarity) in the context of, say, "plane","flight",etc.. will be identified with any noun sense of "bank" which coincidentally overlaps with one or more senses of other words in the cluster. The algorithm will never know the difference. Thus, any new word or sense must be explicitly hand-coded into WordNet, a highly complex task, before it can participate in sense labelling. One could conceivably label any unknown word in a cluster with some sense tag which is "dominant" across the cluster, but that would not distinguish between unknown words, missing senses and words that didn't belong in the cluster in the first place. Further, choosing such a "dominant" sense simply returns us to the sense disambiguation problem again.

Lastly, for Yarowsky the issue of adding new senses of words to existing pairs, identifying such senses, and adding new words is both simple and complex at once. Since there is no notion of similarity but only of discriminants, which have been selected to identify a particular pairwise contrast, it is quite possible for the verb sense of "plant" to masquerade well enough to be labelled as the noun. Severe problems could arise from the appearance of unanticipated senses in the training corpus, since they would eventually be tagged one way or another by the system if they shared any contexts with known senses.

The algorithm certainly can be straightforwardly extended to handle multiple senses, and one can train a new decision list for any new sense pair represented in the corpus. However, it would be desirable to not have to start from stratch to learn decision rules for each new ambiguous pair. Instead, we would like to be able to share or duplicate appropriate parts of the decision lists we have learned for other "similar" words, with weights appropriately adjusted for the collocations. However, simply by inspection of the rule, we can not tell what information source gave rise to this collocation. Thus we can not tell which rules are transferrable or what relations must hold between two words in order to share rules. For instance, many word association rules would hold, say, for the both the verb and noun senses of "plant" since they both relate to agriculture , but none of the rules that related to adjacent content words because these would be most heavily influenced by syntax, in which these senses differ. Likewise, "brook" and "river" are very similar with respect to most information sources, but a list which learned "babbling" as a good collocate for "brook" should not transfer that to "river".

Clearly, a well-supported, general, extensible notion of similarity provides major advantages for word sense disambiguation systems it terms of identifying and incorporating new words and senses. This discussion has also highlighted the utility of an "open" method, rather than one which rigidly encapsulates all that it learns. Finally, it again points up the need to identify the underlying sources of disambiguating surface structures, since words we wish to handle with our algorithms may be similar in their reaction to some environments, while differing in others. As in the "plant" example above, the broad topics in which both the noun and verb form occur are quite similar, but they differ dramatically in position in predicate-argument structure.

7 The Big Picture

Stepping back from the detailed examination of definitions of similarity, use and capture of underlying knowledge sources from surface phenomena, and questions of extensibility and generalization, we can now evaluate the overall contributions of these techniques to identifying what is needed for an effective, trainable, and extendable technique for word sense disambiguation. We can also identify some key points of failure.

7.1 Where's the model?

Curiously, none of these approaches undertakes to define what really constitutes a sense. Schutze generates distributional clusters and tags them, while Resnik uses WordNet as a source of senses. Yarowsky likewise uses the defined seeds as "senses". All three make reference to different levels of sense distinctions, without defining the criteria for fine-grained vs. coarse-grained senses, while saying that the former are less important than the latter. One would think that such a definition of the task you are trying to solve would be a key component of the experimental structure. Further, none of the approaches tries to describe, say, how senses are learned by people or model the development of selection of sense. As a result, one finds the lack of generalizabilty and lack of coherent representation that leads to the problems we have detailed.

7.2 Some Pieces of a Model

Just as Yarowsky rightly criticizes Schutze and other users of "wide-window" co-occurrences for using a "bag of words", he in turn is guilty of using a "bag of rules" in his decision lists and a "bag of classifiers" to hold all of his different pair-specific decision procedures. By treating the words around a target word as an unordered list, one loses the opportunity to exploit or model the influence of syntax, word association, and other factors which depend on order and position. Even people experience a severe degradation in their ability to perform sense assignmentstasks when normal sentence environments are replaced with unordered groups of words. How can we expect computers to fare better? Crucial information is missing.

Likewise, it is wrong to treat all decision rules the same, distinguishing only on the basis of surface statistics. Yarowsky's ability to incorporate any sort of rule into the decision list paradigm is very powerful, but can not fully solve the problem. As the example of "The astronomer married the star" illustrates, some constraints are simply stronger than others, even though there may be no additional support from surface statistics. Further, different types of constraints generalize differently. A collocation based on sound similarity like "babbling brook" is very unlikely to be informative for other words, but selectional restrictions like the requirement that the object be a person can be used for a variety of words. However, one can only make such a generalization if one knows that the relationship which led to the meaning of "married the star" is different from the one which led to the "babbling brook." If you can not make the distinction and generalize from it, you are forced to relearn decision rules for each new word.

Schutze's description of a general metric for similarity of words is a significant contribution. This type of comparison allows one to identify old versus new senses of a word and allows easy extension to new words, by placing them in relation to already known words. In addition, this metric is scalar, permitting degrees of similarity in contrast to Yarowsky's binary distinction between $sense_1$ and $sense_2$. The problem is that the metric is limited. Only "wide-window" information is available, so one can not recognize, much less use or relate, words based on other factors - such as similarity in selectional restriction or part of speech. Resnik too recognizes the importance of defining a general, scalable similarity metric. Unfortunately, he is again hampered by an even more restrictive notion of similarity, by using only WordNet's IS-A hierarchy which reduces the number of dimensions along which we can assess similarity. Yarowsky fails to address the issue of similarity between words, in some ways his approach is a logical descendant of Word-Expert parsing (Small & Reiger) in that each word learns all about how to disambiguate itself, and each new word requires encoding all of the possible cues for it.

It is necessary to recognize that words vary along many dimensions and identifying similarity along any dimension allows one to generalize in that domain. For instance, if one knows only word-word cooccurrences, one, like Yarowsky, can not generalize without external intervention. However, if one can capture both the surface relations between words and the underlying constraints which lead to them, one can generalize appropriately for words which are similar with respect to that relation. Dagan *et al*[10] reach toward this notion when they define similarity as being between words with high mutual information and which occur in similar predicate argument structures. This allows them to substantially increase the applicability of their target word selection technique. In order to build successful, extensible disambiguation techniques, we must model not just the surface co-occurrences which arise from deeper constraints, but also the constraints themselves and their interactions, and then tie this to a robust notion of word similarity. Otherwise, we will be forced to constantly relearn constraints for each new task.

8 Conclusion: The Last Resort

Corpus-based techniques, like those we have discussed here, succeed or fail based on their ability to capture regularities in observed surface word co-occurrences. However, there are disambiguation tasks where surface co-occurrence phenomena provide no cues. Consider an example from Hebrew, where the word *hagira* is ambiguous between immigration and emigration, as follows: "According to the new *hagira* bill every Soviet citizen will have the automatic right to receive a passport valid for five years." [2] One must reason that a bill about passports for Soviet citizens must be a soviet bill and thus passport issuing should be related to leaving rather than entering the country. Also, consider an example from Chinese, "Gou chi ji." which could be translated variously as "Dog/Dogs eat/ate/eats/have eaten chicken/chickens." Chinese has no surface inflection related to singular/plural or tense distinctions, and all of these combinations are valid. Only general inference from knowledge about the event can resolve this multi-way ambiguity. Here we see that although the techniques we reviewed have made use of many sources of disambiguation information based on surface co-occurrence statistics, one source of information is closed to them, forever: inference, which was identified by Hirst, as the source of "last resort". [18] Sometimes there is no substitute for knowing what the sentence means.

[2] dagan1991

References

[1] S. Brennan and J. Ohaeri. Effects of mesage style on users' attributions toward agents. In *CHI '94 Human Factors in Computing Systems Conference Companion*. ACM, 1994.

[2] Eric Brill, David Magerman, Mitchell Marcus, and Beatrice Santorini. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the 29th Annual meetin gof the Association for Computational Linguistics*, 1991.

[3] P. F. Brown, V. J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 20(4), 1992.

[4] Peter F. Brown, Stephen A. DellaPietra, Vincent J. DellaPietra, and Robert L. Mercer. Word sense disambiguation using statistical methods. In *Proceedings 29th Annual Meeting of the Association for Computational Linguistics*, pages 265–270, Berkeley, CA, June 1991.

[5] G. W. Cottrell and S. L. Small. Viewing parsing as word sense discrimination: A connectionist approach. In B. G. Bara and G. Guida, editors, *Computational Models of Natural Language Processing*, pages 91–119. North-Holland, Amsterdam, 1984.

[6] Jim Cowie, Joe Guthrie, and Louise Guthrie. Lexical disambiguation using simulated annealing. In *Proceedings of COLING-92*, pages 359–365, Nantes, 1992.

[7] R. Cullingford. *Script Application: Computer Understanding of newspaper stories*. PhD thesis, Yale University, 1978.

[8] Ido Dagan and Alon Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20:563–596, 1994.

[9] Ido Dagan, Alon Itai, and Ulrike Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 130–137, 1991.

[10] Ido Dagan, Shaul Marcus, and Shaul Markovitch. Contextual word similarity and estimation form sparse data. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 164–171, 1993.

[11] david Yarowsky. Unsupervised word sense disambiguiation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for COmputational Linguistics*, pages 189–196, 1995.

[12] C. G. DeMarcken. Unsupervised acquisition of a lexicon from continuous speech. AI Lab Memo 1558, M.I.T, 1995.

[13] William Gale, Kenneth Church, and David Yarowsky. One sense per discourse. Proceedings of the 4th DARPA Speech and Natural Language Workshop, February 1992.

[14] J. Guthrie, L. Guthrie, Y. Wilks, and H. Aidinejad. Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991.

[15] Z. Harris. *Structural Linguistics*. University of Chicago Pres, Chicago, 1951.

[16] Z. Harris. *Mathematical Structure s of Language*. Wiley, New York, 1968.

[17] M. Hearst. Noun homgraph disambiguation using local context in large corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, 1991.

[18] Graeme Hirst. *Semantic Interpretation and the resolution of Ambiguity*. Cambridge University Press, Cambridge, 1987.

[19] N. M. Ide and J. Veronis. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International COnference on Computational Linguistics*, 1990.

[20] J. J. Katz and J. A. Fodor. The structure of a semantic theory. In J. A. Fodor and J. J. Katz, editors, *The Structure of Language*, chapter 19, pages 479–518. Prentice Hall, 1964.

[21] Robert Krovetz and W. Bruce Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141, April 1992.

[22] Claudia Leacock, Geoffrey Towell, and Ellen M. Voorhees. Toward building contextual representations of word senses using statistical models. In *Proceedings of the 1993 ACL SIGLEX Workshop - Acquisition of Lexical Knowledge from Text*, 1993.

[23] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, 1986.

[24] K. McKeown and V. Hatzivassiloglou. Augmenting lexicons automatically: clustering semantically related adjectives. In M. Bates, editor, *ARPA Workshop of Human Language Processing*, 1993.

[25] G. Miller. Wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.

[26] George Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert Thomas. Using a semantic concordance for sense identification. In *ARPA Workshop on human Language Technology*, Plainsboro, NJ, March 1994.

[27] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of ACL-93*, 1993.

[28] Lawrence Rabiner and Bing-Hwang Juang. *Fundamentals of SPeech Recognition.* Prentice Hall, New Jersey, 1993.

[29] Philip Resnik. Disambiguating noun groupings with respect to WordNet senses. In *Third Workshop on Very Large Corpora.* Association for Computational Linguistics, 1995. (cmp-lg/9511006).

[30] R. Richardson, A.F. Smeaton, and J. Murphy. Using wordnet as a knowledge base for measuring conceptual similarity. In *Proceedings of the 7th Annual Conference on Artificial Intelligence and Cognitive Science*, 1994.

[31] R. Schank and R. Abelson. *Scripts, Plans, goals, and understanding: An enquiry into human knowledge structures.* Lawrence Erlbaum Associates, 1977.

[32] Hinrich Schütze. Word space. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, San Mateo CA, 2993. Morgan Kaufmann.

[33] S.L. Small and C.J. Reiger. *Parsing and Comprehending with Word Experts (a Theory and its Realization).* Lawrence Erlbuam Associates, 1982.

[34] Kah-Kay Sung and Tomaso Poggio. Example based learning for view-based human face detection. AI Lab Memo 1521, M.I.T., 1995.

[35] Michael Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93)*, Arlington, Virginia, 1993.

[36] Ellen M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In Robert Korfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 171–180, 1993.

[37] D.L. Waltz and J.B. Pollack. Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, 9(1):51–74, 1985.

[38] Yorick Wilks. An intelligent analyzer and understander of English. In B. Grosz, K. Sparck Jones, and B. Webber, editors, *Readings in Natural Language Processing*, pages 193–204. Morgan Kaufmann, 1986. Originally appeared in *CACM* 18(5), pp. 264–274, 1975.

[39] Yorick Wilks and Dan Fass. The preference semantics family. *Computers & Mathematics with Applications*, 23(2–5):205–221, 1992.

[40] Yorick Wilks, Dan Fass, Cheng-Ming Guo, James E. Mcdonald, Tony Plate, and Brian M. Slator. Providing machine tractable dictionaries tools. *Machine Translation*, 5(2):99–154, 1990.

[41] David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes, France, July 1992.

[42] David Yarowsky. One sense per collocation. DARPA Workshop on Human Language Technology, March 1993. Princeton.

[43] David Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 1994.

[44] Uri Zernik, editor. *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.* Erlbaum, 1991.