Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation

Saif Mohammad
University of Toronto
Toronto, ON M4M2X6 Canada
smm@cs.toronto.edu
http://www.cs.toronto.edu/~smm

Ted Pedersen
University of Minnesota
Duluth, MN 55812 USA
tpederse@d.umn.edu
http://www.d.umn.edu/~tpederse

Abstract

The success of supervised learning approaches to word sense disambiguation is largely dependent on the features used to represent the context in which an ambiguous word occurs. Previous work has reached mixed conclusions; some suggest that combinations of syntactic and lexical features will perform most effectively. However, others have shown that simple lexical features perform well on their own. This paper evaluates the effect of using different lexical and syntactic features both individually and in combination. We show that it is possible for a very simple ensemble that utilizes a single lexical feature and a sequence of part of speech features to result in disambiguation accuracy that is near state of the art.

1 Introduction

Most words in natural language exhibit *polysemy*, that is, they have multiple possible meanings. Each of these meanings is referred to as a *sense*, and *word sense disambiguation* is the process of identifying the intended sense of a target word based on the context in which it is used. The *context* of the target word consists of the sentence in which it occurs, and possibly one or two surrounding sentences. Consider the following sentence:

`Harry cast a bewitching` **spell** (1)

The target word *spell* has many possible senses, such as, *a charm or incantation, to read out letter by letter*, and *a period of time*. The intended sense, *a charm or incantation*, can be identified based on the context, which in this case includes *bewitching* and a reference to a famous young wizard.

Word sense disambiguation is often approached by supervised learning techniques. The training data consists of sentences which have potential target words tagged by a human expert with their intended sense. Numerous learning algorithms, such as, Naive Bayesian classifiers, Decision Trees and Neural Networks have been used to learn models of disambiguation. However, both (Pedersen, 2001a) and (Lee and Ng, 2002) suggest that different learning algorithms result in little change in overall disambiguation results, and that the real determiner of accuracy is the set of features that are employed.

Previous work has shown that using different combinations of features is advantageous for word sense disambiguation (e.g., (McRoy, 1992), (Ng and Lee, 1996), (Stevenson and Wilks, 2001), (Yarowsky and Florian, 2002)). However, less attention is paid to determining what the minimal set of features necessary to attain high accuracy disambiguation are. In this paper we present experiments that measure the redundancy in disambiguation accuracy achieved by classifiers using two different sets of features, and we also determine an upper bound on the accuracy that could be attained via the combination of such classifiers into an ensemble.

We find that simple combinations of lexical and syntactic features can result in very high disambiguation accuracy, via an extensive set of experiments using the SENSEVAL-1, SENSEVAL-2, *line*, *hard*, *serve* and *interest* data. Together, this consists of more than 50,000 sense-tagged instances. This paper also introduces a technique to quantify the optimum gain that is theoretically possible when two feature sets are combined in an ensemble. In the process, we identify some of the most useful part of speech and parse features.

2 Feature Space

We employ lexical and syntactic features in our word sense disambiguation experiments. The lexical features are unigrams, bigrams, and the surface form of the target word, while the syntactic features are part of speech tags and various components from a parse tree.

2.1 Lexical Features

The surface form of a target word may restrict its possible senses. Consider the noun *case* which has the surface forms: *case, cases* and *casing*. These have the following senses: *object of investigation, frame or covering* and *a weird person*. Given an occurrence of the surface form *casing*, we can immediately conclude that it was used in the sense of *a frame or covering* and not the other two. Each possible surface form as observed in the training data is represented as a binary feature, and indicates if that particular surface form occurs (or not).

Unigrams are individual words that appear in the text. Consider the following sentence:

`the judge dismissed the case` (2)

Here *the, judge, dismissed, the* and *case* are unigrams. Both *judge* and *dismissed* suggest that *case* has been used in the *judicial* sense and not the others. Every unigram that occurs above a certain frequency threshold in the training corpus is represented as a binary feature. For example, there is a feature that represents whether or not *judge* occurs in the context of a target word.

Bigrams are pairs of words that occur in close proximity to each other, and in a particular order. For example, in the following sentence:

`the interest rate is lower in state banks` (3)

the interest, interest rate, rate is, is lower, lower in, in state and *state banks* are bigrams, where *interest rate* suggests that *bank* has been used in the *financial institution* sense and not the *river bank* sense. Every bigram that reaches a given frequency and measure of association score threshold is represented as a binary feature. For example, the bigram feature *interest rate* has value of 1 if it occurs in the context of the target word, and 0 if it does not.

We use the Ngram Statistics Package[1] to identify frequent unigrams and statistically significant bigrams in the training corpus for a particular word. However, unigrams or bigrams that occur commonly in text are ignored by specifying a stop list composed mainly of prepositions, articles and conjunctions.

2.2 Part of Speech Features

The parts of speech of words around the target word are also useful clues for disambiguation. It is likely that when used in different senses, the target word will have markedly different configuration of parts of speech around it. The following sentences have the word *turn* in *changing sides/parties* sense and *changing course/direction* senses, respectively:

`Did/VBD Jack/NNP` **turn/VB against/IN**

$$\mathbf{his/PRP\$\ team/NN}\ ?/. \quad (4)$$

`Did/VBD Jack/NNP` **turn/VB left/NN**

$$\mathbf{at/IN\ the/DT}\ \texttt{crossing/NN}\ ?/. \quad (5)$$

Observe that the parts of speech following each occurrence of *turn* are significantly different, and that this distinction can be captured both by individual and combinations of part of speech features.

The parts of speech of individual words at particular positions relative to the target word serve as features. The part of speech of the target word is P_0. The POS of words following the target are denoted by P_1, P_2 and so on. The POS of words to the left of the target word are P_{-1}, P_{-2}, etc. There is a binary feature for each part of speech tag observed in the training corpus at the given position or positions of interest.

Suppose we would like to use part of speech features for the target word and one word to the right of the target. If the target word has 3 different parts of speech observed in the training data, and the word to the right (without regard to what that word is) has 32 different part of speech tags, then there will be 35 binary features that represent the occurrence of those tags at those positions.

We also consider combinations of part of speech tags as features. These indicate when a particular sequence of part of speech tags occurs at a given set of positions. These features are boolean, and indicate if a particular sequence of tags has occurred or not. In the scenario above, there would be 96 different binary features represented, each of which indicates if a particular combination of values for the two positions of interest, occurs.

2.3 Parse Features

A sentence is made up of multiple phrases and each phrase, in turn, is made of phrases or words. Each phrase has a *head word* which may have strong syntactic relations with other words in the sentence. Consider the phrases, *her hard work* and *the hard surface*. The head words *work* and *surface* are indicative of the *calling for stamina/endurance* and *not easily penetrable* senses of *hard*.

Thus, the head word of the phrase housing the target word is used as a feature. The head word of its parent phrase is also suggestive of the intended sense of the target word. Consider the sentence fragments *fasten the line* and *cross the line*. The noun phrases (*the line*) have the verbs *fasten* and *cross* as the head of parent phrases. Verb *fasten* is indicative of the *cord* sense of *line* while *cross* suggests the *division* sense.

The phrase housing the target word and the parent phrase are also used as features. For example, phrase

[1] http://ngram.sourceforge.net

housing the target word is a noun phrase, parent phrase is a verb phrase and so on. Similar to the part of speech features, all parse features are boolean.

3 Experimental Data

We conducted experiments using part of speech tagged and parsed versions of the SENSEVAL-2, SENSEVAL-1, *line, hard, serve* and *interest* data. The packages `posSenseval` and `parseSenseval` part of speech tagged and parsed the data, respectively. `posSenseval` uses the Brill Tagger while `parseSenseval` employs the Collins Parser. We used the training and test data divisions that already exist in the SENSEVAL-2 and SENSEVAL-1 data. However, the *line, hard, serve* and *interest* data do not have a standard division, so we randomly split the instances into test (20%) and training (80%) portions.

The SENSEVAL-2 and SENSEVAL-1 data were created for comparative word sense disambiguation exercises held in the summers of 2001 and 1998, respectively. The SENSEVAL-2 data consists of 4,328 test instances and 8,611 training instances and include a total of 73 nouns, verbs and adjectives. The training data has the target words annotated with senses from WordNet. The target words have a varied number of senses ranging from two for *collaborate, graceful* and *solemn* to 43 for *turn*. The SENSEVAL-1 data has 8,512 test and 13,276 training instances, respectively. The number of possible senses for these words range from 2 to 15, and are tagged with senses from the dictionary *Hector*.

The *line* data (Leacock, 1993) consists of 4,149 instances where the noun *line* is used in one of six possible WordNet senses. This data was extracted from the 1987-1989 Wall Street Journal (WSJ) corpus, and the American Printing House for the Blind (APHB) corpus. The distribution of senses is somewhat skewed with more than 50% of the instances used in the *product* sense while all the other instances more or less equally distributed among the other five senses.

The *hard* data (Leacock, 1998) consists of 4,337 instances taken from the San Jose Mercury News Corpus (SJM) and are annotated with one of three senses of the adjective *hard*, from WordNet. The distribution of instances is skewed with almost 80% of the instances used in the *not easy - difficult* sense.

The *serve* data (Leacock, 1998) consists of 5,131 instances with the verb *serve* as the target word. They are annotated with one of four senses from WordNet. Like *line* it was created from the WSJ and APHB corpora.

The *interest* data (Bruce, 1994) consists of 2,368 instances where the noun *interest* is used in one of six senses taken from the Longman Dictionary of Contemporary English (LDOCE). The instances are extracted from the part of speech tagged subset of the Penn Treebank Wall Street Journal Corpus (ACL/DCI version).

4 Experiments and Discussion

The `SyntaLex` word sense disambiguation package was used to carry out our experiments. It uses the C4.5 algorithm, as implemented by the J48 program in the Waikato Environment for Knowledge Analysis (Witten and Frank, 2000) to learn a decision tree for each word to be disambiguated.

We use the majority classifier as a baseline point of comparison. This is a classifier that assigns all instances to the most frequent sense in the training data. Our system defaults to the majority classifier if it lacks any other recourse, and therefore it disambiguates all instances. We thus, report our results in terms of accuracy. Table 1 shows our overall experimental results, which will be discussed in the sections that follow. Note that the results of the majority classifier appear at the bottom of that table, and that the most accurate result for each set of of data is shown in bold face.

4.1 Lexical Features

We utilized the following lexical features in our experiments: the surface form of the target word, unigrams and bigrams. The entries under *Lexical* in Table 1 show disambiguation accuracy when using those features individually.

It should be noted that the experiments for the SENSEVAL-2 and SENSEVAL-1 data using unigrams and bigrams are re-implementations of (Pedersen, 2001a), and that our results are comparable. However, the experiments on *line, hard, serve* and *interest* have been carried out for the first time.

We observe that in general, surface form does not improve significantly on the baseline results provided by the majority classifier. While in most of the data (SENSEVAL-2, *line, hard* and *serve* data) there is hardly any improvement, we do see noticeable improvements in SENSEVAL-1 and *interest* data. We believe that this is due to the nature of the feature. Certain words have many surface forms and senses. In many such cases, certain senses can be represented by a restricted subset of possible surface forms. Such words are disambiguated better than others using this feature.

4.2 Part of Speech Features

Word sense disambiguation using *individual part of speech* features is done in order to compare the effect of single POS features versus possibly more powerful combination part of speech features. They are not expected to be powerful enough to do very good classification but may still capture certain intuitive notions. For example, it is very likely that if the noun *line* is preceded by a wh

Table 1: Supervised WSD Accuracy by Feature Type

Features	SENSEVAL-2	SENSEVAL-1	line	hard	serve	interest
Lexical						
Surface Form	49.3%	62.9%	54.3%	81.5%	44.2%	64.0%
Unigrams	**55.3**%	66.9%	**74.5**%	83.4%	73.3%	75.7%
Bigrams	**55.1**%	66.9%	72.9%	**89.5**%	72.1%	79.9%
POS						
P_{-2}	47.1%	57.5%	54.9%	81.6%	52.1%	56.0%
P_{-1}	49.6%	59.2%	56.2%	82.1%	54.8%	62.7%
P_0	49.9%	60.3%	54.3%	81.6%	47.4%	64.0%
P_1	53.1%	63.9%	54.2%	81.6%	55.6%	65.3%
P_2	48.9%	59.9%	54.3%	81.7%	48.9%	62.3%
POS Combos						
P_{-1}, P_0	50.8%	62.2%	56.5%	82.3%	60.3%	67.7%
P_0, P_1	54.3%	66.7%	54.1%	81.9%	60.2%	70.5%
P_1, P_2	53.2%	64.0%	55.9%	82.2%	58.0%	68.6%
P_{-1}, P_0, P_1	54.6%	**68.0**%	60.4%	84.8%	73.0%	78.8%
P_{-2}, P_{-1}, P_0, P_1, P_2	54.6%	**67.8**%	62.3%	86.2%	**75.7**%	**80.6**%
Parse						
Head (H)	51.7%	64.3%	54.7%	87.8%	47.4%	69.1%
Head of Parent (HP)	50.0%	60.6%	59.8%	84.5%	57.2%	67.8%
Phrase POS (P)	52.9%	58.5%	54.3%	81.5%	41.4%	54.9%
Parent Phrase POS (PP)	52.7%	57.9%	54.3%	81.7%	41.6%	54.9%
Parse Combos						
H + HP	52.6%	65.1%	60.4%	87.7%	58.1%	73.2%
H + P	51.9%	65.1%	54.7%	87.8%	45.9%	69.1%
H + HP + P	52.9%	65.5%	60.4%	87.7%	57.6%	73.2%
H + P + HP + PP	52.7%	65.6%	60.5%	87.7%	56.7%	73.5%
Majority Classifier	47.7%	56.3%	54.3%	81.5%	42.2%	54.9%

word such as *whose* or *which*, it is used in the *phone line* sense. If the noun *line* is preceded by a preposition, say *in* or *of*, then there is a good chance that *line* has been used in the *formation* sense. The accuracies achieved by part of speech features on SENSEVAL-2, SENSEVAL-1, *line*, *hard*, *serve* and *interest* data are shown in Table 1. The individual part of speech feature results are under *POS*, and the combinations under *POS Combos*.

We observe that the individual part of speech features result in accuracies that are significantly better than the majority classifier for all the data except for the *line* and *hard*. Like the surface form, we believe that the part of speech features are more useful to disambiguate certain words than others. We show averaged results for the SENSEVAL-2 and SENSEVAL-1, and even there the part of speech features fare well. In addition, when looking at a more detailed breakdown of the 73 and 36 words included in these samples respectively, a considerable number of those words experience improved accuracy using part of speech features.

In particular, we observed that while verbs and adjectives are disambiguated best by part of speech of words one or two positions on their right (P_1, P_2), nouns in general are aided by the part of speech of immediately adjacent words on either side (P_{-1}, P_1). In the case of transitive verbs (which are more frequent in this data than intransitive verbs), the words at positions P_1 and P_2 are usually the objects of the verb (for example, *drink water*). Similarly, an adjective is usually immediately followed by the noun which it qualifies (for example, *short discussion*). Thus, in case of both verbs and adjectives, the word immediately following (P_1) is likely to be a noun having strong syntactic relation to it. This explains the higher accuracies for verbs and adjectives using P_1 and would imply high accuracies for nouns using P_{-1}, which too we observe. However, we also observe high accuracies for nouns using P_1. This can be explained by the fact that nouns are often the subjects in a sentence and the words at positions P_1 and P_2 may be the syntactically related verbs, which aid in disambiguation.

To summarize, verbs are aided by P_1 and P_2, adjectives by P_1 and nouns by P_{-1} and P_1. Thus, P_1 is the the most

potent individual part of speech feature to disambiguate a set of noun, verb and adjective target words.

4.2.1 Combining Part of Speech features

A combination of parts of speech of words surrounding (and possibly including) the target word may better capture the overall context than single part of speech features. Following is an example of how a combination of part of speech features may help identify the intended sense of the noun *line*. If the target word *line* is used in the plural form, is preceded by a personal pronoun and the word following it is not a preposition, then it is likely that the intended sense is *line of text* as in *the actor forgot his lines* or *they read their lines slowly*. However, if the word preceding *line* is a personal pronoun and the word following it is a preposition, then it is probably used in the *product* sense, as in, *their line of clothes*. *POS Combos* in Table 1 shows the accuracies achieved using such combinations with the SENSEVAL-2, SENSEVAL-1, *line, hard, serve* and *interest* data. Again due to space constraints we do not give a break down of the accuracies for the SENSEVAL-2 and SENSEVAL-1 data for noun, verb and adjective target words.

We note that decision trees based on binary features representing the possible values of a given sequence of part of speech tags outperforms one based on individual features. The combinations which include P_1 obtain higher accuracies. In the the case of the verbs and adjectives in SENSEVAL-2 and SENSEVAL-1 data, the best results are obtained using the parts of speech of words following the target word. The nouns are helped by parts of speech of words on both sides. This is in accordance with the hypothesis that verbs and adjectives have strong syntactic relations to words immediately following while nouns may have strong syntactic relations on either side. However, the *hard* and *serve* data are found to be helped by features from both sides. We believe this is because of the much larger number of instances per task in case of *hard* and *serve* data as compared to the adjectives and verbs in SENSEVAL-1 and SENSEVAL-2 data. Due to the smaller amount of training data available for SENSEVAL-2 and SENSEVAL-1 words, only the most potent features help. The power of combining features is highlighted by the significant improvement of accuracies above the baseline for the *line* and *hard* data, which was not the case using individual features (Table 1).

4.3 Parse Features

We employed the following parse features in these experiments: the head word of the phrase housing the target word, the type of phrase housing the target word (Noun phrase, Verb Phrase, etc), the head of the parent phrase, and the type of parent phrase. These results are shown under *Parse* in Table 1.

The head word feature yielded the best results in all the data except *line*, where the head of parent phrase is most potent. Further, the nouns and adjectives benefit most by the head word feature. We believe this the case because the head word is usually a content word and thus likely to be related to other nouns in the vicinity. Nouns are usually found in noun phrases or prepositional phrases. When part of a noun phrase, the noun is likely to be the head and thus does not benefit much from the head word feature. In such cases, the head of the parent phrase may prove to be more useful as is the case in the *line* data. In case of adjectives, the relation of the head word to the target word is expected to be even stronger as it is likely to be the noun modified by the adjective (target word). The verb is most often found in a verb phrase and is usually the head word. Hence, verb target words are not expected to be benefited by the head word feature, which is what we find here. The phrase housing the target word and the parent phrase were not found to be beneficial when used individually.

4.3.1 Combining Parse Features

Certain parse features, such as, the phrase of the target word, take very few distinct values. For example, the target word *shirt* may occur in at most just two distinct kinds of phrases: noun phrase and prepositional phrase. Such features are not expected to perform much better than the majority classifier. However, when used in combination with other features, they may be useful. Thus, like part of speech features, experiments were conducted using a combination of parse features in an effort to better capture the context and to identify sets of features which work well together. Consider the parse features *head word* and *parent word*. Head words such as *magazine, situation* and *story* are indicative of the *quality of causing attention to be given* sense of *interest* while parent words such as *accrue* and *equity* are indicative of the *interest rate* sense. A classifier based on both features can confidently classify both kinds of instances. Table 1 has the results under *Parse Combos*. The Head and Head of Parent combinations have in general yielded significantly higher accuracies than simply the head word or any other parse feature used individually. The improvement is especially noteworthy in case of *line, serve* and *interest* data. The inclusion of other features along with these two does not help much more. We therefore find the Head and Head of Parent combination to be the most potent parse feature combination. It may be noted that a break down of accuracies (not shown here for sake of brevity) for noun, verb and adjective target words, of the SENSEVAL-1 and SENSEVAL-2 data revealed that the adjectives were disambiguated best using the Head word and Phrase combination. This is observed in the *hard* data results as well, albeit marginally.

Table 2: The Best Combinations of Syntactic and Lexical Features

Data	Feature-Set Pair				Baseline Ens.	Maj. Class.	Simple Ens.	Optimal Ens.	Best
	Set 1	Acc.	Set2	Acc.					
SVAL-2	Unigram	55.3%	P_{-1}, P_0, P_1	54.6%	43.6%	47.7%	57.0%	67.9%	66.7%
SVAL-1	Unigram	66.9%	P_{-1}, P_0, P_1	68.0%	57.6%	56.3%	71.1%	78.0%	81.1%
line	Unigram	74.5%	P_{-1}, P_0, P_1	60.4%	55.1%	54.3%	74.2%	82.0%	88.0%
hard	Bigram	89.5%	Head, Parent	87.7%	86.1%	81.5%	88.9%	91.3%	83.0%
serve	Unigram	73.3%	P_{-1}, P_0, P_1	73.0%	58.4%	42.2%	81.6%	89.9%	83.0%
interest	Bigram	79.9%	P_{-1}, P_0, P_1	78.8%	67.6%	54.9%	83.2%	90.1%	89.0%

5 Complementary/Redundant Features

As can be observed in the previous results, many different kinds of features can lead to roughly comparable word sense disambiguation results.

Different types of features are expected to be *redundant* to a certain extent. In other words, the features will individually classify an identical subset of the instances correctly. Likewise, the features are expected to be *complementary* to some degree, that is, while one set of features correctly disambiguates a certain subset of instances, use of another set of features results in the correct disambiguation of an entirely distinct subset of the instances.

The extent to which the feature sets are complementary and redundant justify or obviate the combining of the feature sets. In order to accurately capture the amount of redundancy and complementarity among two feature sets, we introduce two measures: the *Baseline Ensemble* and the *Optimal Ensemble*. Consider the scenario where the outputs of two classifiers based on different feature sets are to be combined using a simple voting or ensemble technique for word sense disambiguation.

The *Baseline Ensemble* is the accuracy attained by a hypothetical ensemble technique which correctly disambiguates an instance only when both the classifiers identify the intended sense correctly. In effect, the Baseline Ensemble quantifies the redundancy among the two feature sets. The *Optimal Ensemble* is the accuracy of a hypothetical ensemble technique which accurately disambiguates an instance when any of the two classifiers correctly disambiguates the intended sense. We say that these are hypothetical in that they can not be implemented, but rather serve as a post disambiguation analysis technique.

Thus, the Optimal Ensemble is the upper bound to the accuracy achievable by combining the two feature sets using an ensemble technique. If the accuracies of individual classifiers is X and Y, the Optimal Ensemble can be defined as follows:

$$OptimalEnsemble = (X - BaselineEnsemble) + (Y - BaselineEnsemble) + BaselineEnsemble$$

We use a simple ensemble technique to combine some of the best lexical and syntactic features identified in the previous sections. The probability of a sense to be the intended sense as identified by lexical and syntactic features is summed. The sense which attains the highest score is chosen as the intended sense. Table 2 shows the best results achieved using this technique along with the baseline and optimal ensembles for the SENSEVAL-2, SENSEVAL-1, *line*, *hard*, *serve* and *interest* data. The table also presents the feature sets that achieved these results. In addition, the last column of this table shows representative values for some of the best results attained in the published literature for these data sets. Note that these are only approximate points of comparison, in that there are differences in how individual experiments are conducted for all of the non–SENSEVAL data.

From the Baseline Ensemble we observe that there is a large amount of redundancy across the feature sets. That said, there is still a significant amount of complementarity as may be noted by the difference between the Optimal Ensemble and the greater of the individual accuracies. For example, in the SENSEVAL-2 data, unigrams alone achieve 55.3% accuracy and part of speech features attain an accuracy of 54.6%. The Baseline Ensemble attains accuracy of 43.6%, which means that this percentage of the test instances are correctly tagged, independently, by both unigrams and part of speech features. The unigrams get an additional 11.7% of the instances correct which the part of speech features tag incorrectly.

Similarly, the part of speech features are able to correctly tag an additional 11% of the instances which are tagged erroneously when using only bigrams. The above values suggest a high amount of redundancy among the unigrams and part of speech features but not high enough to suggest that there is no significant benefit in combining the two kinds of features. The difference between the Optimal Ensemble and the accuracy attained by unigrams

is 12.6% (67.9% - 55.3%). This is a significant improvement in accuracy which may be achieved by a suitable ensemble technique. The difference is a quantification of the complementarity between unigram and part of speech features based on the data. Further, we may conclude that given these unigram and part of speech features, the best ensemble techniques will not achieve accuracies higher than 67.9%.

It may be noted that a single unified classifier based on multiple features may achieve accuracies higher than the Optimal Ensemble. However, we show that an accurate ensemble method (Optimal Ensemble), based on simple lexical and syntactic features, achieves accuracies comparable or better than some of the best previous results. The point here is that using information from two distinct feature sets (lexical features and part of speech) could lead to state of the art results. However, it is as yet unclear how to most effectively combine such simple classifiers to achieve these optimal results.

Observation of the pairs of lexical and syntactic features which provide highest accuracies for the various data suggest that the part of speech combination feature - P_{-1}, P_0, P_1, is likely to be most complementary with the lexical features (bigrams or unigrams).

The *hard* data did particularly well with combinations of parse features, the Head and Parent words. The Optimal Ensemble attains accuracy of over 91%, while the best previous results were approximately 83%. This indicates that not only are the Head and Parent word features very useful in disambiguating adjectives but are also a source of complementary information to lexical features.

6 Related Work

(McRoy, 1992) was one of the first to use multiple kinds of features for word sense disambiguation in the semantic interpretation system, TRUMP. The system aims at disambiguating all words in the text and relies extensively on dictionaries and is not corpus based. Scores are assigned based on morphology, part of speech, collocations and syntactic cues. The sense with the highest score is chosen as the intended sense. TRUMP was used to tag a subset of the Wall Street Journal (around 2500 words) but was not evaluated due to lack of gold standard.

The LEXAS system of (Ng and Lee, 1996) uses part of speech, morphology, co-occurrences, collocations and verb object relation in nearest neighbor implementation. The system was evaluated using the *interest* data on which it achieved an accuracy of 87.3%. They studied the utility of individual features and found collocations to be most useful, followed by part of speech and morphological form.

(Lin, 1997) takes a supervised approach that is unique as it did not create a classifier for every target word. The system compares the context of the target word with that of training instances which are similar to it. The sense of the target word most similar to these contexts is chosen as the intended sense. Similar to McRoy, the system attempts to disambiguate all words in the text. Lin relies on syntactic relations, such as, subject-verb agreement and verb object relations to capture the context. The system achieved accuracies between 59% and 67% on the SemCor corpus.

(Pedersen, 2001b) compares decision trees, decision stumps and a Naive Bayesian classifier to show that bigrams are very useful in identifying the intended sense of a word. The accuracies of 19 out of the total 36 tasks in SENSEVAL-1 data were greater than the best reported results in that event. Bigrams are easily captured from raw text and the encouraging results mean that they can act as a powerful baseline to build more complex systems by incorporating other sources of information. Pedersen points out that decision trees can effectively depict the relations among the various features used. With the use of multiple sources of information this quality of decision trees gains further significance.

(Lee and Ng, 2002) compare the performances of Support Vector Machines, Naive Bayes, AdaBoost and Decision Trees using unigrams, parts of speech, collocations and syntactic relations. The experiments were conducted on SENSEVAL-2 and SENSEVAL-1 data. They found the combination of features achieved highest accuracy (around 73%) in SENSEVAL-1 data, irrespective of the learning algorithm. Collocations(57.2%), part of speech tags(55.3%) and syntactic relations(54.2%) performed better than decision trees using all features in the SENSEVAL-2 data.

(Yarowsky and Florian, 2002) performed experiments with different learning algorithms and multiple features. Three kinds of Bayes Classifier, Decision lists and Transformation Based Learning Model (TBL) were used with collocations, bag of words and syntactic relations as features. Experiments on SENSEVAL-2 data revealed that the exclusion of any of the three kinds of features resulted in a significant drop in accuracy. Lee and Ng as well as Yarowsky and Florian conclude that the combination of features is beneficial.

(Pedersen, 2002) does a pairwise study of the systems that participated in SENSEVAL-2 English and Spanish disambiguation exercises. The study approaches the systems as black boxes, looking only at the assigned tags whatever the classifier and sources of information may be. He introduces measures to determine the similarity of the classifications and optimum results obtainable by combining the systems. He points out that pairs of systems having low similarity and high optimal accuracies are of interest as they are markedly complementary and the combination of such systems is beneficial.

There still remain questions regarding the use of mul-

tiple sources of information, in particular which features should be combined and what is the upper bound on the accuracies achievable by such combinations. (Pedersen, 2002) describes how to determine the upper bound when combining two systems. This paper extends that idea to provide measures which determine the upper bound when combining two sets of features in a single disambiguation system. We provide a measure to determine the redundancy in classification done using two different feature sets. We identify particular part of speech and parse features which were found to be very useful and the combinations of lexical and syntactic features which worked best on SENSEVAL-2, SENSEVAL-1, *line, hard, serve* and *interest* data.

7 Conclusions

We conducted an extensive array of word sense disambiguation experiments using a rich set of lexical and syntactic features. We use the SENSEVAL-2, SENSEVAL-1, *line, hard, serve* and *interest* data which together have more than 50,000 sense tagged instances. We show that both lexical and syntactic features achieve reasonably good accuracies when used individually, and that the part of speech of the word immediately following the target word is particularly useful in disambiguation as compared to other individual part of speech features. A combination of part of speech features attains even better accuracies and we identify (P_0, P_1) and (P_{-1}, P_0, P_1) as the most potent combinations. We show that the head word of a phrase is particularly useful in disambiguating adjectives and nouns. We identify the head and parent as the most potent parse feature combination.

We introduce the measures *Baseline Ensemble* and *Optimal Ensemble* which quantify the redundancy among two feature sets and the maximum accuracy attainable by an ensemble technique using the two feature sets. We show that even though lexical and syntactic features are redundant to a certain extent, there is a significant amount of complementarity. In particular, we showed that simple lexical features (unigrams and bigrams) used in conjunction with part of speech features have the potential to achieve state of the art results.

8 Acknowledgments

This work has been partially supported by a National Science Foundation Faculty Early CAREER Development award (#0092784).

References

R. Bruce and L. Wiebe. 1994 Word-Sense Disambiguation using Decomposable Models In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.

C. Leacock and M. Chodorow and G. Miller. 1998 Using Corpus Statistics and WordNet Relations for Sense Identification *Computational Linguistics*, 24(1):147–165.

C. Leacock and E. Voorhees. 1993 Corpus-Based Statistical Sense Resolution In *Proceedings of the ARPA Workshop on Human Language Technology*.

K.L. Lee and H.T. Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 41–48.

D. Lin. 1997. Using syntactic dependency as a local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Madrid, July.

S. McRoy. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47.

T. Pedersen. 2001a. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 79–86, Pittsburgh, July.

T. Pedersen. 2001b. Machine learning with lexical features: The duluth approach to senseval-2. In *Proceedings of the Senseval-2 Workshop*, pages 139–142, Toulouse, July.

T. Pedersen. 2002. Assessing system agreement and instance difficulty in the lexical samples tasks of senseval-2. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 40–46, Philadelphia.

M. Stevenson and Y. Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349, September.

I. Witten and E. Frank. 2000. *Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan–Kaufmann, San Francisco, CA.

D. Yarowsky and R. Florian. 2002. Evaluating sense disambiguation performance across diverse parameter spaces. *Journal of Natural Language Engineering*, 8(2).

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA.