Complementarity of Lexical and Simple Syntactic Features: The SyntaLex Approach to SENSEVAL-3

Saif Mohammad
University of Toronto
Toronto, ON M5S1A1 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

This paper describes the SyntaLex entries in the English Lexical Sample Task of SENSEVAL-3. There are four entries in all, where each of the different entries corresponds to use of word bigrams or Part of Speech tags as features. The systems rely on bagged decision trees, and focus on using pairs of lexical and syntactic features individually and in combination. They are descendants of the Duluth systems that participated in SENSEVAL-2.

1 Introduction

The SyntaLex systems are supervised learners that identify the intended sense of a word (target word) given its context. They are derived from the Duluth systems that participated in SENSEVAL-2, and which are more fully described in (Pedersen, 2001b).

The context of a word is a rich source of discrete features which lend themselves nicely to decision tree learning. Prior research (e.g., (McRoy, 1992), (Ng and Lee, 1996), (Stevenson and Wilks, 2001), (Yarowsky and Florian, 2002)) suggests that use of both syntactic and lexical features will improve disambiguation accuracies. There has also been considerable work on word sense disambiguation using various supervised learning algorithms. However, both (Pedersen, 2001a) and (Lee and Ng, 2002) show that different learning algorithms produce similar results and that the use of appropriate features may dramatically improve results. Thus, our focus is not on the learning algorithm but on the features used and their dynamics.

Our systems use bigrams and Part of Speech features individually, in a simple ensemble and as part of single classifier using both kinds of features. We also show that state of the art results (72.1%, coarse grained accuracy) can be achieved using just these simple sets of features.

2 Feature Space

Simple lexical and syntactic features are used to represent the context. The lexical features used are word bigrams. The Part of Speech (PoS) of the target word and its neighbors make up the the syntactic features. Bigrams are readily captured from the text while Part of Speech taggers are widely available for a variety of languages.

2.1 Bigrams

A bigram is a pair of words that occur close to each other in text and in a particular order. Consider:

`the interest rate is lower in state banks` (1)

It has the following bigrams: *the interest, interest rate, rate is, is lower, lower in, in state* and *state banks*. Note that the bigram *interest rate* suggests that *bank* has been used in the *financial institution* sense and not the *river bank* sense.

All features are binary valued. Thus, the bigram feature *interest rate* has value 1 if it occurs in the context of the target word, and 0 if it does not. The learning algorithm considers only those bigrams that occur at least twice in the training data and have a word association ratio greater than a certain predecided threshold. Bigrams that tend to be very common are ignored via a stop list. The Ngram Statistics Package[1] is used to identify statistically significant bigrams in the training corpus, for a particular word.

2.2 Part of Speech Features

The Part of Speech (PoS) of the target word and its surrounding words can be useful indicators of its intended sense. Consider the following sentences where *turn* is used in *changing sides/parties* and *changing course/direction* senses, respectively:

Did/VBD Jack/NNP **turn**/VB **against/IN his/PRP$ team/NN** ?/. (2)

Did/VBD Jack/NNP **turn**/VB **left/NN at/IN the/DT** crossing/NN ?/. (3)

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

Notice that the Part of Speech of words following *turn* in the two sentences are significantly different. We believe that words used in different senses may be surrounded by words with different PoS. Therefore, PoS of words at particular positions relative to the target word are used as features to identify the intended sense. The PoS of the target word is denoted by P_0. The Part of Speech of words following it are represented by P_1, P_2 and so on, while that of words to the left of the target word are P_{-1}, P_{-2}, etc. Like bigrams, the Part of Speech features are binary. For example, the feature (P_1 = JJ) has value 1 if the target word is followed by an adjective (JJ), and 0 otherwise.

3 Data and its Pre-processing

The English lexical sample of SENSEVAL-3 has 7,860 sense-tagged training instances and 3,944 test instances. The training data has six pairs of instances with identical context (different instance ID's). These duplicates are removed so as not to unfairly bias the classifier to such instances. The test data has one pair of with the same context but no instances were removed from the test data in order to facilitate comparison with other systems. The data also has certain instances with multiple occurrences of a word marked as the target word. We remove all such markings except for the first occurrence of the target word in an instance. Thus, our systems identify the intended sense based solely on how the target word is used in the first occurrence.

The sense-tagged training and test data are Part of Speech tagged using the `posSenseval`[2] package. `posSenseval` PoS tags any data in SENSEVAL-2 data format (same as SENSEVAL-3 format) using the Brill Tagger. It represents the PoS tags in appropriate xml tags and outputs data back in SENSEVAL-2 data format. A simple sentence boundary identifier is used to place one sentence per line, which is a requirement of the Brill Tagger. The mechanism of Guaranteed Pre-tagging (Mohammad and Pedersen, 2003) is used to further enhance the quality of tagging around the target words. The experiments performed on this pre-processed data are described next.

4 Experiments and Discussion

The `SyntaLex` systems are used to perform a series of word sense disambiguation experiments using lexical and syntactic features both individually and in combination. The C4.5 algorithm, as implemented by the J48 program in the Waikato Environment for Knowledge Analysis (Witten and Frank, 2000) is used to learn bagged decision trees for each word to be disambiguated.

Ten decision trees are learned for each task based on ten different samples of training instances. Each sample is created by drawing N instances, with replacement, from a training set consisting of N total instances. Given a test instance, weighted scores for each sense provided by each of the ten decision trees are summed. The sense with the highest score is chosen as the intended sense.

A majority classifier which always chooses the most frequent sense of a word in the training data, achieves an accuracy of 56.5%. This result acts as a baseline to which our results may be compared. The decision trees learned by our system fall back on the most frequent sense in case the identified features are unable to disambiguate the target word. Thus, the classification of all test instances is attempted and we therefore report our results (Table 1) in terms of accuracies. The break down of the coarse and fine grained accuracies for nouns, verbs and adjectives is also depicted.

4.1 `SyntaLex-1`: Part of Speech Features (Narrow Context)

`SyntaLex-1` uses bagged decision trees to classify a target word based on its Part of Speech and that of its immediate neighbors. The nodes in the decision trees are features of form: P_{-1} = <Tag>, P_0 = <Tag> or P_1 = <Tag>, where <Tag> represents any Part of Speech. Consider a sentence where the target word *line* is used in the plural form, has a personal pronoun preceding it and is not followed by a preposition. A decision tree based on such Part of Speech features as described above is likely to capture the intuitive notion that in such cases *line* is used in the *line of text* sense, as in, *the actor forgot his lines* or *they read their lines slowly*. Similarly, if the word following *line* is a preposition, the tree is likely to predict the *product* sense, as in, *the line of clothes*.

The system achieves a fine grained accuracy of 62.4% and a coarse grained accuracy of 69.1%.

4.2 `SyntaLex-2`: Part of Speech Features (Broad Context)

`SyntaLex-2`, like `SyntaLex-1`, uses bagged decision trees based on part of speech features for word sense disambiguation. However, it relies on the Part of Speech of words within a broader window around the target word. The Part of Speech of words in a sentence have local influence. The Part of Speech of words further away from the target word are not expected to be as strong indicators of intended sense as the immediate neighbors. However,

[2] http://www.d.umn.edu/ tpederse/pos.html

inclusion of such features has been shown to improve accuracies (Mohammad and Pedersen, 2004). The nodes in the decision trees are features of the form: P_{-2} = <Tag>, P_{-1} = <Tag>, P_0 = <Tag>, P_1 = <Tag> or P_2 = <Tag>.

The system achieves a fine grained and coarse grained accuracy of 61.8% and 68.4%, respectively.

4.3 `SyntaLex-3`: Ensemble of Lexical and Simple Syntactic Features

Prior research has shown that both lexical and syntactic features can individually achieve a reasonable quality of disambiguation. Further, some of the work (e.g., (McRoy, 1992), (Ng and Lee, 1996)) suggests that using both kinds of features may result in significantly higher accuracies as compared to individual results.

`SyntaLex-3` utilizes Part of Speech features and bigrams. Individual classifiers based on both kinds of features are learned. Given a test instance, both classifiers assign probabilities to every possible sense. The probabilities assigned to a particular sense are summed and the sense with the highest score is chosen as the desired sense. A narrow context of Part of Speech features is used for the syntactic decision tree that has features of the form: P_{-1} = <Tag>, P_0 = <Tag> or P_1 = <Tag>.

`SyntaLex-3` achieves a fine grained accuracy of 64.6% and a coarse grained accuracy of 72.0%.

4.4 `SyntaLex-4`: Combination of Lexical and Simple Syntactic Features

`SyntaLex-4` also relies on a combination of PoS and bigram features but uses unified decision trees that can have either kind of feature at a particular node. In an ensemble, for a sense to be chosen as the intended one, both classifiers must assign reasonably high probabilities to it. A low score for a particular sense by any of the classifiers will likely entail its rejection. However, in certain instances, the context may be rich in useful disambiguating features of one kind but not of the other.

A unified decision tree based on both kinds of features has the flexibility of choosing the intended sense based on one or both kinds of features and hence likely to be more successful. It must be noted though that throwing in a large number of features intensifies the data fragmentation problem of decision trees.

`SyntaLex-4` achieves a fine grained and coarse grained accuracies of 63.3% and 71.1%, respectively.

5 Discussion

Observe that even though `SyntaLex-2` uses a larger context than `SyntaLex-1` it does not do much better than the latter, in fact, its accuracies are slightly lower. We believe this is due to the low training data per task ratio, which usually means that the weak indicators (P_{-2} and P_2) are likely to be overwhelmed by idiosyncrasies of the data. (Mohammad and Pedersen, 2004) show results to the same conclusions for SENSEVAL-1 and SENSEVAL-2 data that have similar low training data per task, while, the *line, hard, serve* and *interest* data which have much larger training data per task are shown to benefit from a larger context.

`Duluth-ELSS` (a sister system of `SyntaLex`) achieves an accuracy of 61.7%. It creates an ensemble of three bagged decision trees, where one tree is based on unigrams, another on bigrams, and a third on co–occurrences with the target word. Observe that its accuracy is comparable to `SyntaLex-2` (62.4%) which use only Part of Speech features. However, these results alone do not tell us if both kinds of features disambiguate the same set of instances correctly, that is, they are mutually redundant, or they classify differing sets of instances correctly, that is, they are mutually complementary. Significant complementarity implies that a marked increase in accuracies may be achieved by suitably combining the bigram and Part of Speech features. We have shown earlier (Mohammad and Pedersen, 2004) that there is indeed large complementarity between lexical and syntactic features by experiments on *line, hard, serve, interest*, SENSEVAL-1 and SENSEVAL-2 data. We use the measures *Optimal Ensemble* and *Baseline Ensemble*, introduced there, to quantify the complementarity and redundancy between bigrams and Part of Speech features in the SENSEVAL-3 data.

The *Baseline Ensemble* of bigram and PoS features is the accuracy of a hypothetical ensemble that correctly disambiguates an instance only when the individual classifiers based on both kinds of features correctly identify the intended sense. The *Optimal Ensemble* of bigrams and PoS features is the accuracy of a hypothetical ensemble that accurately disambiguates an instance when any of the two individual classifiers correctly disambiguates the intended sense. We find the Baseline Ensemble of bigrams and PoS features on SENSEVAL-3 data to be 52.9% and the Optimal Ensemble to be 72.1%. Thus, given 100 instances, almost 53 of them would be correctly tagged by both kinds of classifiers and up to 72 may be correctly disambiguated using a powerful ensemble technique.

Table 1: Disambiguation Accuracies

System	Granularity	Overall	Nouns	Verbs	Adjectives
Majority Classifier		56.5%	55.0%	58.0%	54.1%
`SyntaLex-1`	Fine	62.4%	58.7%	67.0%	48.0%
	Coarse	69.1%	65.1%	73.3%	61.7%
`SyntaLex-2`	Fine	61.8%	57.7%	66.5%	50.0%
	Coarse	68.4%	64.1%	73.1%	60.1%
`SyntaLex-3`	Fine	64.6%	62.5%	67.6%	51.6%
	Coarse	72.0%	69.6%	74.9%	64.2%
`SyntaLex-4`	Fine	63.3%	62.2%	65.3%	49.1%
	Coarse	71.1%	69.5%	73.4%	62.0%

In order to capitalize on the significant complementarity of bigrams and Part of Speech features, `SyntaLex-3` uses a simple ensemble technique, while `SyntaLex-4` learns a unified decision tree based on both bigrams and Part of Speech features. Observe that both `SyntaLex-3` and `4` achieve accuracies higher than `SyntaLex-1` and `2`. Further, `SyntaLex-3` performs slightly better than `SyntaLex-4`. We believe that `SyntaLex-4` may be affected by data fragmentation caused by learning decision trees from a large number of features and limited training data. We also note that the Optimal Ensemble is markedly higher than the accuracies of `SyntaLex-3` and `4`, suggesting that the use of a more powerful combining methodology is justified.

References

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.

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

S. Mohammad and T. Pedersen. 2003. Guaranteed Pre-Tagging for the Brill Tagger. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2003*.

S. Mohammad and T. Pedersen. 2004. Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation. (To appear) in *Proceedings of the Eighth Conference on Natural Language Learning at HLT-NAACL*.

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).