Chapter 5
N-Gram Features for Unsupervised WSD with an Underlying Naïve Bayes Model

Abstract The feature selection method we are presenting in this chapter relies on web scale N-gram counts. It uses counts collected from the web in order to rank candidates. Features are thus created from unlabeled data, a strategy which is part of a growing trend in natural language processing. Disambiguation results obtained by web N-gram feature selection will be compared to those of previous approaches that equally rely on an underlying Naïve Bayes model but on completely different feature sets. Test results corresponding to the main parts of speech (nouns, adjectives, verbs) will show that web N-gram feature selection for the Naïve Bayes model is a reliable alternative to other existing approaches, provided that a "quality list" of features, adapted to the part of speech, is used.

Keywords Bayesian classification · Word sense disambiguation · Unsupervised disambiguation · Web-scale N-grams

5.1 Introduction

The present chapter focuses on an entirely different way of performing feature selection for the Naïve Bayes model, that relies on using web scale N-gram counts. The presented feature selection method was introduced in (Preoţiuc and Hristea 2012). To our knowledge, it represents a first attempt of using web N-gram features in unsupervised WSD in general, and in conjunction with the Naïve Bayes model as clustering technique for unsupervised WSD in particular. While creating features from unlabeled data, we are "helping" a simple, basic knowledge-lean disambiguation algorithm, hereby represented by the Naïve Bayes model, to significantly increase its accuracy as a result of receiving easily obtainable knowledge.

The proposed feature selection method (Preoţiuc and Hristea 2012) is based on the intuition that the most frequently occurring words near the target can give us a better indication of the sense which is activated than words being semantically

F. T. Hristea, *The Naïve Bayes Model for Unsupervised Word Sense Disambiguation*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-33693-5_5, © The Author(s) 2013

similar that may not appear so often in the same context with the target word. The corresponding disambiguation method is unsupervised and knowledge-lean in the sense that it just requires the existence or the possibility to estimate N-gram counts for the target language corresponding to which the disambiguation process takes place. No information regarding the actual word senses will be used at any stage of the process. When using such features, the Naïve Bayes model will not require any sense definitions or sense inventories.

5.2 The Web as a Corpus

With respect to feature selection it is necessary to use those words that are the most relevant and distinctive for the target word. So, it is intuitive to think that these words are the ones that co-occur most often with the target. These words can be found by searching and performing an estimate over large corpora and the largest corpora available is the whole Web itself.

While the web provides an imense linguistic resource, collecting and processing data at web-scale is very timeconsuming. Previous research has relied on search engines to collect online information, but an alternative to this that has been developed more recently is to use the data provided in an N-gram corpus. An N-gram corpus is an efficient compression of large amounts of text as it states how often each sequence of words (up to length N) occurs.

The feature selection method that we are presenting here makes use of the Google Web 1T 5-gram Corpus Version 1.1, introduced in (Brants and Franz 2006), that contains English word N-grams (with N up to 5) and their observed frequency counts, calculated over 1 trillion words from the web and collected by Google in January 2006. The text was tokenized following the Penn Treebank tokenization, except that hyphenated words, dates, email addresses and URLs are kept as single tokens. The sentence boundaries are marked with two special tokens <S> and </S>. Words that occurred fewer than 200 times were replaced with the special token <UNK>. The data set has a N-gram frequency cutoff, that is N-grams that have a count that is less than 40 are discarded.

This corpus has been used in a variety of NLP tasks with good results. Yuret (2007) describes a WSD system that uses a statistical language model based on the Web 1T 5-gram dataset. The model is used to evaluate the likelihood of various substitutes for a word in a given context. These likelihoods are then used to determine the best sense for the word in novel contexts. (Bergsma et al. 2009) presents a unified view of using web-scale N-gram models for lexical disambiguation and uses the counts of 2–5 grams in a supervised method on the task of preposition selection, spelling correction or non-referential pronoun detection. In (Bergsma et al. 2010) web-scale N-gram data is used for supervised classification on a variety of NLP tasks such as: verb part-of-speech disambiguation, prenominal adjective ordering or noun compound bracketing. Islam and Inkpen (2009) have used the N-gram data for spelling correction, while Chang and Clark (2010) have made use of this data to check the acceptability of paraphrases in context.

Web-scale N-gram counts are used for the first time in unsupervised word sense disambiguation, as a mean of feature selection for the Naïve Bayes model, in (Preoţiuc and Hristea 2012).

In order to find the most frequent words that co-occur with the target word within a distance of N−1 words, one must take into consideration the N-grams in which the target word occurs. Thus, we can build different feature sets depending on the size of N and on the number of words to include in the feature set. These sets will be referred using the following convention: **n-w-t** represents the set containing the top **t** words occurring in **n**-grams together with the word **w**.

For example, *5-line-100* is the set constituted by the most frequent 100 (stemmed) words that co-occur in the Web with the word *line* within a distance of, at most, 4 words.

In order to build the feature set corresponding to the top t words occurring in N-grams of size n with the target word w, (n-w-t), Preoţiuc and Hristea (2012) have used the following processing directions:

- they have lowercased every occurrence in the N-gram corpus and have combined the counts for identical matches;
- for every number $k(k < n)$, they have built a list of words and counts, each representing word counts occurring at a distance of exactly k on each side of the target word;
- they have merged the counts from all $n - 1$ lists to get a complete list of words and counts that co-occur in a context window of size $n - 1$ with the target word w;
- they have removed the numbers, the punctuation marks, the special tokens (eg. <s>, <unk>), the words starting with special characters or symbols and the stopwords from the list;
- they have performed stemming using the Porter Stemmer on each feature set, merging counts for similar words whenever the case;
- they have sorted the word and counts pairs in descending order of their counts and have extracted the top t words.

Let us note the fact that, while in the context window only content words exist, within the N-grams stopwords may also occur. So it is not guaranteed that the N-grams show the counts of words appearing in a context window of N−1. Preoţiuc and Hristea (2012) have chosen to eliminate stopwords because they appear much too often in the corpora and, by using them as features, the model tends to put too much weight on these, as opposed to the content words that are the ones indicative of the word sense.

Despite the fact that the target words and the dataset we refer to in the experiments are in English, the feature selection method we are discussing here is language independent and can be applied with no extra costs to other languages for which we know or can estimate N-gram counts from large data. Recently, Google has released Web 1T 5-gram, 10 European Languages Version 1 (Brants and Franz 2009) consisting of word N-grams and their observed frequency counts for other ten European languages: Czech, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Spanish and Swedish. The N-grams were extracted from publicly accessible web pages from October 2008 to December 2008 using the same conventions as for the English data set, with only the data being approximately 10 times smaller. Thus, the presented method can be used with no changes whatsoever to extract features for performing sense disambiguation corresponding to these languages as well.

Using a Web scale N-gram corpus implies performing counts that take into account all the possible senses of the target word. Automatically, when computing these counts, high frequency senses will have more words indicative of those senses than low frequency senses have. If the disambiguation setting is restricted to a specific domain (eg. medicine), the discussed method of feature extraction could be used with a N-gram corpus derived from large corpora of texts in that domain.

5.3 Experimental Results

Preoţiuc and Hristea (2012) have tested their proposed feature sets for the three main parts of speech: nouns, adjectives and verbs. They have drawn conclusions, that we shall be presenting here, with regard to each of these parts of speech.

5.3.1 Corpora

In order to compare their results with those of other previous studies (Pedersen and Bruce 1998; Hristea et al. 2008; Hristea 2009; Hristea and Popescu 2009) that have presented the same Naïve Bayes model, trained with the EM algorithm, but using other methods of feature selection, Preoţiuc and Hristea (2012) try to disambiguate the same target words using the same corpora.

In the case of nouns they have used as test data the *line* corpus (Leacock et al. 1993). This corpus contains around 4,000 examples of the word *line* (noun) sense-tagged with one of the 6 possible WordNet 1.5 senses. Examples are drawn from the WSJ corpus, the American Printing House for the Blind, and the San Jose Mercury. The description of the senses and their frequency distribution[1] are shown in Table 5.1.

In (Pedersen and Bruce 1998; Hristea et al. 2008) tests are also performed for only 3 senses of *line*. Preoţiuc and Hristea (2012) do not perform this comparison as their method is not relying on sense inventories. Therefore it is not possible to distinguish and take out the words that co-occur with the specific senses represented in the test set.

In the case of adjectives and verbs the mentioned authors have used as test data the corpus introduced in (Bruce et al. 1996) that contains twelve words taken from the ACL/DCI Wall Street Journal corpus and tagged with senses from the Longman Dictionary of Contemporary English.

[1] Which are the same as those considered in Chap. 3.

Table 5.1 Distribution of senses of *line*

Sense	Count	Pct. (%)
Product	2,218	53,47
Written or spoken text	405	9,76
Telephone connection	429	10,34
Formation of people or things; queue	349	8,41
An artificial division; boundary	376	9,06
A thin, flexible object; cord	371	8,94
Total count	4,148	100

Table 5.2 Distribution of senses of *common*

Sense	Count	Pct. (%)
As in the phrase "common stock"	892	84
Belonging to or shared by 2 or more	88	8
Happening often; usual	80	8
Total count	1,060	100

Table 5.3 Distribution of senses of *public*

Sense	Count	Pct. (%)
Concerning people in general	440	68
Concerning the government and people	129	19
Not secret or private	90	13
Total count	659	100

Tests have been conducted for two adjectives, *common* and *public*, the latter being the one corresponding to which Pedersen and Bruce (1998) obtain the worst disambiguation results.

The senses of *common* and *public* that have been taken into consideration and their frequency distribution[2] are shown in Table 5.2 and in Table 5.3, respectively. In order to compare their results to those of (Pedersen and Bruce 1998; Hristea et al. 2008; Hristea and Popescu 2009), Preoţiuc and Hristea (2012) have also taken into account only the 3 most frequent senses of each adjective, as was the case in those studies.

For verbs, the part of speech which is known as being the most difficult to disambiguate, Preoţiuc and Hristea (2012) have performed tests corresponding to the verb *help* while considering the most frequent two senses of this word. The definition of the senses and the frequency distribution[3] are presented in Table 5.4.

In order for the experiments to be conducted, the data set was preprocessed (Preoţiuc and Hristea 2012) in the usual way: the stopwords, words with special characters and numbers were eliminated and stemming was applied to all remaining words, using the same Porter Stemmer as in the case of stemming the lists of feature words.

[2] Which are the same as those considered in Chaps. 3 and 4.

[3] Which are the same as those considered in Chap. 3.

Table 5.4 Distribution of senses of *help*

Sense	Count	Pct. (%)
To enhance-inanimate object	990	78
To assist-human object	279	22
Total count	1,269	100

5.3.2 *Tests*

As was the case in the mentioned previous studies that examine unsupervised WSD with an underlying Naïve Bayes model, studies to the results of which they are comparing their own disambiguation results, Preoţiuc and Hristea (2012) also evaluate performance in terms of accuracy. As it is well known, in the case of unsupervised disambiguation defining accuracy is not as straightforward as in the supervised case. The objective is to divide the I given instances of the ambiguous word into a specified number K of sense groups, which are in no way connected to the sense tags existing in the corpus. In the experiments, sense tags are used only in the evaluation of the sense groups found by the unsupervised learning method. These sense groups must be mapped to sense tags in order to evaluate system performance. As in the previously mentioned studies, in order to enable comparison, Preoţiuc and Hristea (2012) have used the mapping that results in the highest classification accuracy.

In the case when none of the words belonging to the feature set are found in the context window of the target, as in (Hristea et al. 2008; Hristea 2009; Hristea and Popescu 2009), the disambiguation method presented by Preoţiuc and Hristea (2012) assigns the instance to the cluster that has the greatest number of assignments. If the target word has a dominant sense, which is the case with all the considered test target words, lower coverage will determine an increase in the performance of the method when results are below the most frequent sense baseline (a very high one in the case of unsupervised WSD using the same underlying mathematical model). With respect to this, Preoţiuc and Hristea (2012) also define coverage as the percentage of instances in which at least one feature word occurs in the context window and, so, the assignment is performed by the Naïve Bayes classifier as opposed to a most frequent sense one.

Preoţiuc and Hristea (2012) show results that couple accuracy with coverage. They use a context window with varying size around the target word, the coverage for a feature set increasing accordingly with the enlargement of the window size.

As in (Hristea et al. 2008) each presented result represents the average accuracy obtained by the disambiguation method over 20 random trials while using a fixed threshold ε having the value 10^{-9}.

In what follows, we show the most significant test results that were obtained (Preoţiuc and Hristea 2012) in the case of all main parts of speech.

Within the graphs, the (Preoţiuc and Hristea 2012) results are designated by solid lines with different markers indicating the various parameters (n or t) that were used. The context window sizes vary and are listed in the corresponding text for

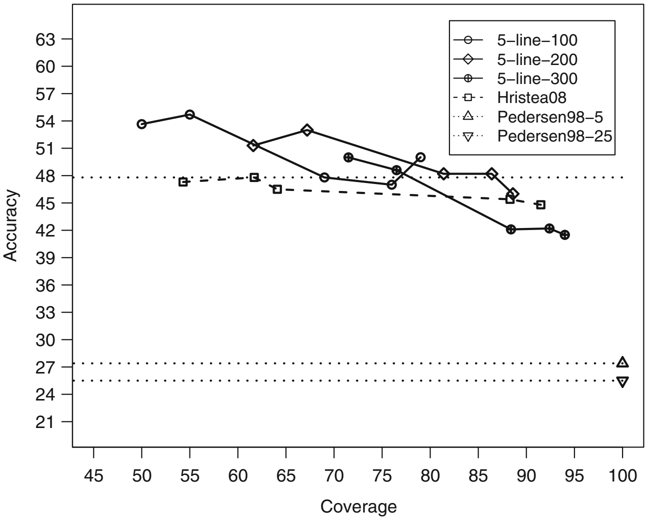

Fig. 5.1 Results for feature sets *5-line*

each part of speech. The Hristea et al. (2008) method is presented with a dashed line and always uses a context window of size 25. The variation in coverage is due to the different type of WordNet relations that were used, resulting in a different number of feature words. The results of the Pedersen and Bruce (1998) method are presented as well. We notice that here we always have just one value, corresponding to a 100 % coverage and to a size of 5 or 25 of the context window. This is due to the fact that the method of feature selection takes into consideration all the words in the vocabulary. Therefore, in this case, there are no contexts with no features. In each graph, corresponding to each of the other two previous methods, and in order to allow an easier visual comparison, Preoţiuc and Hristea (2012) have drawn a dotted black line to illustrate the highest accuracy obtained for that word by the respective method.

5.3.2.1 Test Results Concerning Nouns

In the case of the noun *line* results are presented in Fig. 5.1.[4]

The best results were obtained by using the most frequent words appearing in 5-gram with *line*, although results with a lower *n* were only slightly worse, as reported in (Preoţiuc and Hristea 2012).

Test results are presented (Preoţiuc and Hristea 2012) for context windows of size 4, 5, 10, 15 and 25 corresponding to each feature set. We observe the largest difference in favour of the Preoţiuc and Hristea feature selection method as resulting

[4] Reprinted here from (Preoţiuc and Hristea 2012).

in an accuracy of 54.7 % (for context window 5 and feature set *5-line-100*) as compared to 47.8 % for a similar coverage in (Hristea et al. 2008). For the feature sets *5-line-100* and *5-line-200*, the tests concerning web N-gram feature selection show better performances than any of the results of Hristea et al. (2008) and better, by a wide margin, than those of Pedersen and Bruce (1998). For some experiments, the method outperforms the most frequent sense baseline which, in this case, is situated at 53.47 %.

The graph also shows that by increasing too much the number of features (*5-line-300*), the performance of the system decreases. This performance decreases even more when considering even larger feature sets ($t = 500$ or 1000—not shown on the graph for clarity).

We observe that when web N-gram feature selection is performed in the case of noun disambiguation, increasing the size of the context window (thus bringing more features into the process) does not bring improvements to the disambiguation results (taking into consideration the coverage-accuracy trade-off), as stated in other studies. As reported in (Preoţiuc and Hristea 2012), another interesting aspect is that, by every step in extending the context window, the coverage increases significantly. This remark is not valid, as we shall see, in the case of adjectives and verbs.

The obtained results (Preoţiuc and Hristea 2012) confirm the intuition that, in order to disambiguate a noun, the information in a wide context is useful and can contribute to the disambiguation process. Features taken from wider contexts are also good indicators for disambiguation.

5.3.2.2 Test Results Concerning Adjectives

With respect to adjectives, Preoţiuc and Hristea (2012) have considered the disambiguation of the polysemous words *common* and *public*. Test results are shown in Figs. 5.2[5] and 5.3,[6] respectively.

The best results were achieved by using the most frequent words appearing in bigrams with *common* and in 3-grams with *public* (although results with bigrams for *public* were close in terms of accuracy).

In the case of adjective *common* the results are presented for context windows of size 1, 2, 3, 4, 5 and 10. We observe the largest difference in favour of the Preoţiuc and Hristea (2012) feature selection method as resulting in an accuracy of 87.0 %, as compared to 77.5 %, the best result obtained in (Hristea et al. 2008). Again, almost all scores (16 out of 18 shown) are higher than the ones of the Hristea et al. (2008) method, with almost half of them exceeding the most frequent sense baseline (set at 84.0 % in this case).

Corresponding to the adjective *public* test results are presented for context windows of size 2, 3, 4, 5 and 10. The Preoţiuc and Hristea (2012) best result is 58.7 %

[5] Reprinted here from (Preoţiuc and Hristea 2012).

[6] Reprinted here from (Preoţiuc and Hristea 2012).

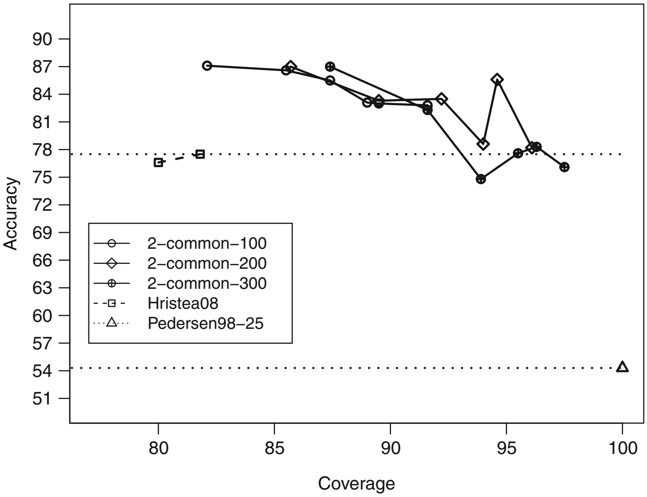

Fig. 5.2 Results for feature sets *2-common*

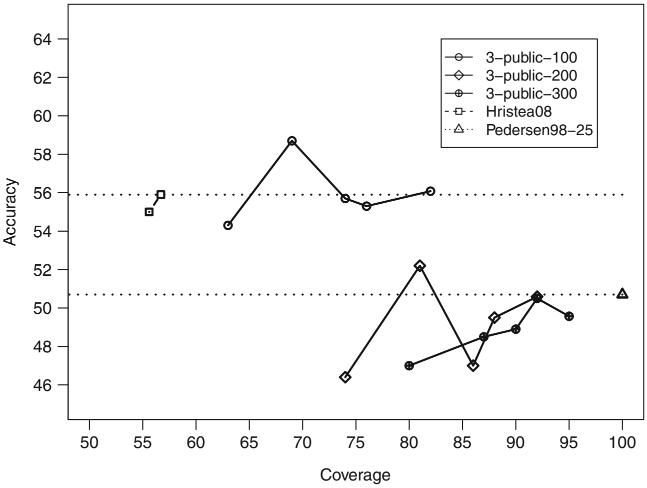

Fig. 5.3 Results for feature sets *3-public*

accuracy as compared to 55.9 % obtained with much smaller coverage in (Hristea et al. 2008).

We must keep in mind that, as we move to the right of the graph (increasing coverage), the results are more significant, because the bias of choosing the most frequent sense baseline for contexts with no features is reduced, due to the fact that the baseline has a very high value (84 and 68 % respectively).

For both adjectives, we observe that just by taking the most frequent 100 words in bigrams or trigrams and a very narrow context window (starting with size 1) we already

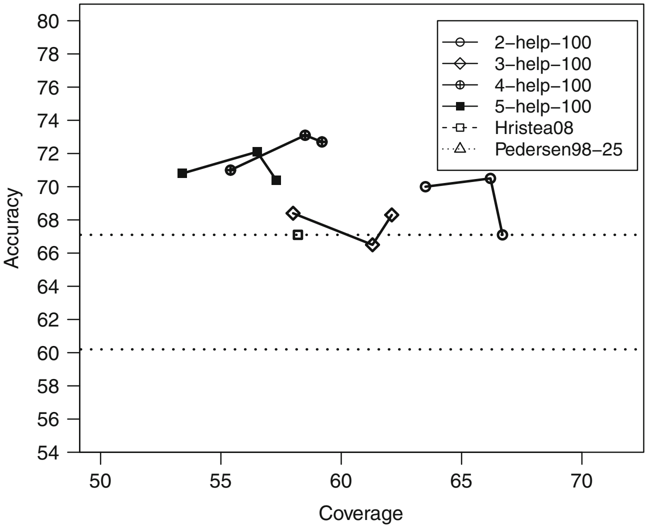

Fig. 5.4 Results for feature sets *help-100*

obtain a very high coverage, that increases at a low rate together with the enlargement of the context window. This corresponds to the linguistic argument that an adjective will appear together with the word it modifies, the latter representing the most frequent and important attribute when disambiguating the respective adjective. Results with wider N-grams were inferior by a distinctive margin (Preoţiuc and Hristea 2012).

5.3.2.3 Test Results Concerning Verbs

Corresponding to the verb *help* test results are shown in Fig. 5.4.[7]

As commented in (Preoţiuc and Hristea 2012), interestingly enough, the best results were achieved by using the top 100 words regardless of the order of the N-grams. The (Preoţiuc and Hristea 2012) top result was 73.1 % when using words from 4-grams and a context window of size 15, as compared to a maximum of 67.1 % in (Hristea et al. 2008), obtained with similar coverage. Out of 12 results, 11 were better than those in (Hristea et al. 2008), confirming the reliability of disambiguating using web N-gram feature sets.

Test results are presented for context windows of size 10, 15 and 25 respectively, as coverage is too low corresponding to smaller context windows. One can notice that coverage for this verb is very low compared to the case of the studied nouns and adjectives and that it increases by a very low margin with the enlargement of the context window.

[7] Reprinted here from (Preoţiuc and Hristea 2012).

This is also very linguistically intuitive because verbs usually appear in very different contexts. This makes feature selection more difficult and is the main reason why most studies conclude that this is the hardest to disambiguate part-of-speech.

As we are shown from the Preoţiuc and Hristea (2012) results, corresponding to all parts of speech, we can restate the fact that, by taking more, less related words (increasing *t*), the accuracy drops, a fact which emphasizes the need for a "quality list of features". The presented feature selection method (Preoţiuc and Hristea 2012) obtains very high results compared to Pedersen and Bruce (1998) in all tests, good results compared to Hristea et al. (2008) and sometimes exceeds the most frequent sense baseline, which is a high baseline to achieve using the Naïve Bayes model.

5.3.3 Adding Knowledge from an External Knowledge Source

While noting that web N-gram feature selection has provided the best disambiguation results so far, we are now trying to "help" the Naïve Bayes model, when acting as clustering technique for unsupervised WSD, by combining the described features with other, additional ones, coming from an external knowledge source. For the purpose of the present discussion, the chosen knowledge source will be WordNet.

Disambiguation results provided by WN-based feature selection are shown and commented in Chap. 3 corresponding to all major parts of speech (nouns, adjectives, verbs). WN-based feature selection has provided more modest disambiguation accuracies than those obtained when using web N-gram features. It is therefore natural to hope for an increase in accuracy when combining the WN-based features with those that have led to the best disambiguation results. In order to test this assumption we have performed[8] a great number of experiments that combine WN-based and web N-gram features. For enabling comparison, we have attempted to disambiguate the same polysemous words that have been discussed so far: the noun *line*, the adjectives *common* and *public* and the verb *help*. The same corpora have been used corresponding to each of these polysemous words.

In the case of the noun *line* we have designed experiments which perform discrimination between the 6 senses listed in Table 5.1. We have started by combining the two sets of features which had provided the best disambiguation results for each of the considered feature selection methods. In the case of WN-based feature selection this is the disambiguation vocabulary formed with WN synonyms, content words of the associated synset glosses and example strings, and all nouns coming from all hyponym and meronym synsets (see Chap. 3). This disambiguation vocabulary had brought an accuracy of 47.8 % (see Sect. 3.4.2.1). In the case of web N-gram feature selection the best disambiguation accuracy (54.7 %) has been obtained with the feature set *5-line-100* (see Sect. 5.3.2.1). When combining these two feature sets accuracy drops to 43.0 % (obtained with 70.3 % corpus coverage). Numerous

[8] Together with Daniel Preoţiuc.

other tests have been performed, none of which have led to the improvement of the disambiguation accuracy. Our best result is represented by an accuracy of 48.7 % (obtained with 95.8 % corpus coverage). As far as WN-based feature selection is concerned, this best result is obtained when considering the disambiguation vocabulary formed with all WN-synonyms and content words of the associated synset glosses and example strings, all nouns of hyponym synsets plus all content words of the associated glosses and example strings, as well as all nouns coming from the meronym synsets, to which all content words of the corresponding glosses and example strings are added. As far as web N-gram feature selection is concerned, the best obtained accuracy resulted when using the feature set *5-line-200*.

This best obtained accuracy (48.7 %) slightly improves the one resulting as best when performing WN-based feature selection alone, and does not come close to the best one obtained with web N-gram feature selection. In the case of nouns, the Naïve Bayes model does not react well to the combination of web N-gram features and WN-based ones.

In the case of the adjective *common* we have designed experiments which perform discrimination again between the 3 senses listed in Table 5.2. Our best obtained result is an accuracy of 87.2 % (with corpus coverage 83.5 %). This is very close to the obtained web N-gram result (87.0 %) and significantly improves the best obtained WN result (77.5 %). As far as feature sets are concerned, it is obtained corresponding to the extended WN vocabulary (all relations) discussed in Chap. 3, but leaving out antonyms, and to the web N-gram feature set *2-common-100*.

In the case of the adjective *public* we have designed experiments which perform discrimination between the 3 senses listed in Table 5.3. Out best obtained result is an accuracy of 56.4 % (with corpus coverage 73.2 %). This is lower than the obtained web N-gram result (58.7 %) and very slightly improves the best obtained WN result (55.9 %). As far as feature sets are concerned, it is obtained corresponding to the same extended WN vocabulary (all relations, including antonymy) discussed in Chap. 3 and to the web N-gram feature set *3-public-100*.

In the case of the verb *help* we have designed experiments which perform discrimination between the 2 senses listed in Table 5.4. Our best obtained result is an accuracy of 70.3 % (with corpus coverage 61.8 %). This is lower than the obtained web N-gram result (73.1 %) and improves the best obtained WN result (67.1 %). As far as feature sets are concerned, it is obtained corresponding to the extended WN vocabulary (all relations) discussed in Chap. 3 and to the web N-gram feature set *3-help-100*.

Our conclusion is that it is not worth combining these features of totally different natures, but it is recommendable to rather use web N-gram features alone.

5.4 Conclusions

This chapter has examined web N-gram feature selection for unsupervised word sense disambiguation with an underlying Naïve Bayes model.

The disambiguation method using N-gram features that we have presented here is unsupervised and uses counts collected from the web in a simple way, in order to rank candidates. It creates features from unlabeled data, a strategy which is part of a growing trend in natural language processing, together with exploiting the vast amount of data on the web. Thus, the method does not rely on sense definitions or inventories. It is knowledge-lean in the sense that it just requires the existence or the possibility to estimate N-gram counts for the target language corresponding to which the disambiguation process takes place. No information regarding the actual word senses is used at any stage of the process.

Comparisons have been performed with previous approaches that rely on completely different feature sets. In the case of all studied parts of speech, test results were better, by a wide margin, than those obtained when using local-type features (Pedersen and Bruce 1998). They have also indicated a superior alternative to WordNet feature selection for the Naïve Bayes model (see Chap. 3). Strictly as far as adjectives are concerned, results are more or less similar to those obtained when feeding the Naïve Bayes model syntactic knowledge of the studied type (see Chap. 4). Web N-gram feature selection seems a reliable alternative to projective dependency-based feature selection as well.

The experiments conducted for all three major parts of speech (nouns, adjectives, verbs) have provided very different results, depending on the feature sets that were used. These results are in agreement with the linguistic intuitions and indicate the necessity of taking into consideration feature sets that are adapted to the part of speech which is to be disambiguated.

Another conclusion we have come to, in the present study, is that, when using the Naïve Bayes model as clustering technique for unsupervised WSD, it is not recommended to combine features created from unlabeled data with those coming from an external knowledge source (such as WordNet).

Last but not least, the presented method has once again proven that a basic, simple knowledge-lean disambiguation algorithm, hereby represented by the Naïve Bayes model, can perform quite well when provided knowledge in an appropriate way.

References

Bergsma, S., Lin, D., Goebel, R.: Web-scale N-gram models for lexical disambiguation. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence, pp. 1507–1512. Pasadena, California (2009)

Bergsma, S., Pitler, E., Lin, D.: Creating robust supervised classifiers via web-scale N-gram data. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10), pp. 865–874. Uppsala, Sweden (2010)

Brants, T., Franz, A.: Web 1T 5-gram corpus version 1.1. Technical Report, Google Research (2006)

Brants, T., Franz, A.: Web 1T 5-gram, 10 European languages version 1. Technical Report, Linguistic Data Consortium, Philadelphia (2009)

Bruce, R., Wiebe, J., Pedersen, T.: The Measure of a Model, CoRR, cmp-lg/9604018 (1996)

Chang, C.Y., Clark, S.: Linguistic steganography using automatically generated paraphrases. In: Human Language Technologies: The Annual Conference of the North American Chapter of the

Association for Computational Linguistics (HLT '10), pp. 591–599. Los Angeles, California (2010)

Hristea, F.: Recent advances concerning the usage of the Naïve Bayes model in unsupervised word sense disambiguation. Int. Rev. Comput. Softw. **4**(1), 58–67 (2009)

Hristea, F., Popescu, M., Dumitrescu, M.: Performing word sense disambiguation at the border between unsupervised and knowledge-based techniques. Artif. Intell. Rev. **30**(1), 67–86 (2008)

Hristea, F., Popescu, M.: Adjective sense disambiguation at the border between unsupervised and knowledge-based techniques. Fundam. Inform. **91**(3–4), 547–562 (2009)

Islam, A., Inkpen, D.: Real-word spelling correction using Google Web IT 3-grams. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '09), pp. 1241–1249. Singapore (2009)

Leacock, C., Towell, G., Voorhees, E.: Corpus-based statistical sense resolution. In: Proceedings of the ARPA Workshop on Human Language Technology, pp. 260–265. Princeton, New Jersey (1993)

Pedersen, T., Bruce, R.: Knowledge lean word-sense disambiguation. In: Proceedings of the 15th National Conference on Artificial Intelligence, pp. 800–805. Madison, Wisconsin (1998)

Preoţiuc-Pietro, D., Hristea, F.: Unsupervised word sense disambiguation with N-gram features. Artif. Intell. Rev. doi:10.1007/s10462-011-9306-y (2012)

Yuret, D.: KU: Word sense disambiguation by substitution. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07), pp. 207–214. Prague (2007)