On a Dependency-based Semantic Space for Unsupervised Noun Sense Disambiguation with an Underlying Naïve Bayes Model

Florentina Hristea
University of Bucharest, Department of Computer Science
Academiei 14, Str., Bucharest, Sector 1, C.P. 010014
fhristea@fmi.unibuc.ro

Abstract

Recent studies refocus on usage of the Naïve Bayes model in unsupervised word sense disambiguation (WSD). They discuss the issue of feature selection for this statistical model, when used as clustering technique, and comment (Hristea, 2012) that it still holds a promise for unsupervised WSD. Within the various investigated types of feature selection, this ongoing research concentrates on syntactic dependency-based features, introduced in (Hristea and Colhon, 2012) with respect to adjectives only. We hereby extend the mentioned approach to the case of nouns and recommend the further investigation of this promising feature selection method.

1 Introduction

While the Naïve Bayes model has been widely and successfully used in supervised WSD (Navigli, 2009), its usage in unsupervised WSD has led to more modest disambiguation results and is less frequent. However, more recent studies (Hristea, 2012) state that this statistical model still holds a promise for unsupervised WSD.

The Naïve Bayes model needs to be fed knowledge (of various natures) in order to perform well as clustering technique for unsupervised WSD (Hristea, 2012). Three different sources of such knowledge have been predominantly examined and compared: WordNet (Hristea et al., 2008; Hristea, 2009; Hristea and Popescu, 2009), web N-grams (Preotiuc and Hristea, 2012) and dependency relations (Hristea and Colhon, 2012; Hristea, 2012). While most of these studies discuss all three major parts of speech (nouns, adjectives, verbs), the syntactic dependency-based feature selection method has been applied to adjectives only (Hristea and Colhon, 2012; Hristea, 2012). With the conclusion that the Naïve Bayes model reacts well in the presence of syntactic knowledge of this type and that dependency-based feature selection for the Naïve Bayes model is a reliable alternative to other existing ones. In fact, for the studied adjectives, this type of syntactic feature selection has provided the best disambiguation results (Hristea, 2012). Following the line of reasoning of the mentioned studies, we hereby extend the disambiguation method they propose to nouns, while exemplifying with tests concerning the nouns *line* and *interest*.

Although dependency-based semantic space models have been studied and discussed by several authors (Padó and Lapata, 2007; Năstase, 2008; Chen et al., 2009), to our knowledge, grammatical dependencies have been used in conjunction with the Naïve Bayes model only very recently (Hristea and Colhon, 2012; Hristea, 2012). The latter authors follow the line of reasoning of Padó and Lapata (2007) which they adapt to the particularities of the involved statistical model.

The present study investigates the usage of syntactic features provided by dependency relations as defined by the classical Dependency Grammar formalism (Tesnière, 1959) and as proposed in (Hristea and Colhon, 2012; Hristea, 2012). The semantic space we present to the Naïve Bayes model for unsupervised WSD will be based on dependency relations extracted from natural language texts via a syntactic parser. In order to ensure the same testing setup as the one used in the mentioned studies (Hristea and Colhon, 2012; Hristea, 2012), we shall be making use of a PCFG parser, namely the Stanford parser (Klein and Manning, 2003), for extracting syntactic dependency relations that will indicate the disambiguation vocabulary required by the Naïve Bayes model. When using dependency-based syntactic features this disambiguation vocabulary is formed by taking into account all words that participate in the considered dependencies. Also in order to ensure the same testing setup, we shall be estimating the model parameters using the Expectation-Maximization algorithm (Dempster et al., 1977). Our approach to feature selection is that of implementing a Naïve Bayes model that uses as features *the actual words* occurring in the context window of the target and decreases the existing number of features by selecting a restricted number of such words, as indicated by the chosen dependency relations. The size of the feature set must be reduced in order to decrease the number of parameters which are to be estimated by the EM algorithm for unsupervised WSD.

2 Design of the experiments

Our approach will take into account the final conclusions drawn in (Hristea, 2012) with respect to dependency-based feature selection for the Naïve Bayes model. According to this most recent study, several particularities determined by the involved statistical model stand out. When using the Stanford parser a projective[1] type analysis is recommended. This is in accordance with the classical dependency grammar theory and has previously (Hristea, 2012) improved disambiguation accuracy in the case of adjectives. According to the same study, directionality of the dependency relations counts and the head role of the target (word to be disambiguated) is essential. The type of the dependencies is equally of the essence. It seems sufficient to use first order dependencies (direct relationships between the target and other words). A small number of dependency types should be considered, preferably just one, in order to decrease the number of parameters that will be estimated by the EM algorithm. Some of these conclusions were determined specifically by the nature of the involved statistical model, others by the fact that the Naïve Bayes model is trained with the EM algorithm. For instance, contrary to other authors, who, when discussing the construction of a dependency-based semantic space in general, consider that “directed paths would limit the context too severely” (Padó and Lapata, 2007), Hristea and Colhon (2012) have taken into account both undirected and directed paths - with the latter providing the best test results. The Naïve Bayes model seemed to react strongly to the directionality of dependency relations and considering this directionality was essential when forming the disambiguation vocabulary.

Following this line of work, which is typical for the Naïve Bayes model, when disambiguating the nouns *line* and *interest*, we have considered a *single type* of *first order* dependencies having the target word as *head* and have collected all other words involved in these dependencies in order to form the disambiguation vocabulary.

2.1 Noun experiment

In the case of nouns we have used as test data the *line* corpus (Leacock et al., 1993; Mooney, 1996) and the *interest* corpus (Bruce and Wiebe, 1994). Within the present approach to disambiguation, the value of a feature is given by the number of occurrences of the corresponding word in the given context window (which is hereby represented by the entire sentence). Since the process of feature selection is based on the restriction of the disambiguation vocabulary, it is possible for certain instances not to contain any of the relevant (chosen) words forming this vocabulary. Such instances will have null values corresponding to all features. These instances do not contribute to the learning process. However, they have been taken into account in the evaluation stage of our experiments. Corresponding to these instances, the algorithm assigns the sense for which the value estimated by the EM algorithm is maximal. In order to enable comparison with the mentioned studies, performance is evaluated in terms of accuracy. Also in order to enable comparison with previous work, we have extracted the contexts corresponding to 3 chosen senses of the studied nouns, as shown in Table 1 (for *line*) and Table 2 (for *interest*), respectively. Another reason for performing this reduction to 3 senses was to verify to what extent the existence of a majority sense in the distribution of senses influences the performances of the discussed disambiguation method. Corresponding to the distribution of senses shown in Table 1 (for *line*) and in Table 2 (for *interest*) we have extracted all existing dependency relations using Stanford Parser.

In order to choose a specific type of dependency for the discussed disambiguation method, we have isolated all dependency relations having the target word as head and have classified them according to their frequency and their relevance. (Namely dependencies between the target and dependents which are not content words have been eliminated). The most frequent dependency

[1] Which does not allow the arches denoting the dependency relations to intersect.

relations thus obtained were *amod* (adjectival modifier) and *nn* (noun compound modifier)[2].

Sense	Count
Telephone connection	429 (37.33%)
Formation of people or things; queue	349 (30.37%)
A thin, flexible object; cord	371 (32.28%)
Total count	1149

Table 1 Distribution of the 3 chosen sense of *line*

Sense	Count
Money paid for the use of money	1252 (53%)
A share in a company or business	500 (21%)
Readiness to give attention	361 (15%)
Total count	2113

Table 2 Distribution of the 3 chosen senses of *interest*

We have started by taking into account both these relations since it is not presupposed that the most frequent dependency will provide the best disambiguation result. However, we are interested in frequent dependencies in order to minimize the number of instances having null values corresponding to all features (thus ensuring good corpus coverage). On the other hand, frequent dependencies will provide a greater number of features, resulting in a greater number of parameters that are to be estimated by the EM algorithm. These aspects, which, quite surprisingly, are not of linguistic nature, make the choice of the dependency type to be used in disambiguation a quite delicate one. The present study makes use of the mentioned *amod* and *nn* dependency relations. The disambiguation vocabulary was obtained by retaining all words that are dependents of the target within each of these relations, considered separately. Two distinct disambiguation vocabularies were thus created and tests have been performed corresponding to each of them. The number of contexts and features for each of the considered nouns and dependency relations can be seen in Table 3.

Corpus	line	interest
No. of contexts	1150	2112
No. of senses	3	3
No. of *nn* features	104	65
No. of *amod* features	101	102

Table 3 Corpora features

3 Test results

Performance is evaluated in terms of accuracy, as in (Hristea and Colhon, 2012; Hristea, 2012). In the case of unsupervised disambiguation defining accuracy is not as straightforward as in the supervised cased. The objective is to divide the given instances of the ambiguous word into a specified number of sense groups, which are in no way connected to the sense tags existing in the corpus. These sense groups are then mapped to the sense tags of the annotated corpus. The mapping that results in the highest classification accuracy is chosen. The discussed test results will represent the average accuracy and standard deviation obtained by the learning procedure over 1000 random trials while using the entire sentence as context window and a threshold ε having the value 10^{-9}. As in (Hristea and Colhon, 2012; Hristea, 2012), apart from accuracy, the following type of information is also provided: number of features resulting in the experiment and percentage of instances having only null features.

At the first stage of our experiment, we have performed 100 random trials, both for *line* and for *interest*, corresponding to the *nn* and the *amod* relations, respectively. We have analyzed the obtained results after 10% of the intended tests in order to observe the differences between the two involved dependency relations. These results are presented in Table 4.

After the first 100 random trials, the differences between results obtained with the two considered relations have become visible.

Target word	Relation	No. of features	Accuracy
line	*amod*	101	.544±.08
	nn	104	.579±.08
interest	*amod*	102	.684±.08
	nn	65	.686±.07

Table 4 Test results for *line* and *interest* after 100 random trials

Corresponding to both nouns the obtained accuracy is higher in the case of the *nn* dependency relation. For *line* the "*nn* accuracy" is significantly higher. This has determined us to perform the remaining 900 random trials using the *nn* relation, in the case of both nouns. The obtained test results are shown in Table 5.

Let us note that the *nn* and *amod* relations have a similar frequency in the *line* corpus, while

[2] For both of which see the Stanford Parser Manual (de Marneffe and Manning, 2012).

the frequency of the *amod* relation is significantly higher within the *interest* corpus[3].

Target word	No. of features	Percentage of instances having only null features	Accuracy
line	104	15.7	.584±.09
interest	65	38.2	.683±.07

Table 5 Disambiguation accuracy corresponding to the *nn* dependency relation after 1000 random trials

In spite of this, a higher disambiguation accuracy seems to be obtained using the *nn* dependency relation. In the case of *interest* this can be expected since the number of resulting features is smaller, minimizing the number of parameters that are to be estimated by the EM algorithm. In the case of *line* this observation does not hold, but the difference between the number of resulting features is not significant (see Table 4). The final obtained disambiguation results clearly show that the dependency relation which is most frequently occurring in a corpus is not necessarily the most relevant one for unsupervised WSD of this type.

3.1 Further analysis of the results

We have compared the disambiguation accuracy obtained when performing syntactic dependency-based feature selection with that resulting when using other types of features, proposed by the relatively recent literature: semantic WordNet (WN) features (Hristea et al., 2008) and N-gram features (Preotiuc and Hristea, 2012). These authors report test results for the noun *line*.

In the case of the three chosen senses of *line*, the best reported accuracy when using WN features was 0.591 ± .06, obtained with 229 features and with only 15.1% instances having only null features. The N-grams feature selection method reports as highest accuracy 0.547%, obtained for a context window of size 5 and for the 5-*line*-100 feature set[4]. As shown in Table 5, the best obtained dependency-based accuracy is 0.584 ± .09, a result which, at first glance, would encourage us to prefer semantic WN-based feature sets.

We have further performed tests for the three chosen senses of *interest* using both mentioned feature selection methods and within the same testing setup.

In the case of *interest*, WN feature selection results in a maximum accuracy of 0.587 ± 3.3 when using 18 features that ensure 15.9% corpus coverage. Corresponding to N-gram feature selection we have performed tests with the set of features that had provided the best result for *line*. The obtained accuracy was 44.15% ± 1.97%. With respect to *interest* dependency-based feature selection clearly outperforms both these methods (see Table 5).

In fact, we can state that this type of syntactic feature selection is recommended in the case of both studied nouns. Since corresponding to *line* the number of features used in disambiguation by WN feature selection is much greater (more than double) than the one provided by dependency relations. Which makes us believe that, when moving to 6 senses of line, namely to more fine-grained disambiguation, accuracy will drop severely if using this method. In the case of *interest*, where the number of resulting features is low, one should notice the very low corpus coverage. This is probably due to the fact that the synsets corresponding to the three chosen senses of *interest* do not have many semantic relations in WordNet. Due to possible very reduced corpus coverage, we cannot recommend a feature selection method relying solely on the number of WN relations corresponding to a specific synset.

4. Conclusions and future work

So far, syntactic dependency-based feature selection for unsupervised WSD with an underlying Naïve Bayes model seems a reliable alternative to other existing ones. It has already been recommended for adjectives (Hristea, 2012). Concerning nouns, our next step will be to use it for more fine-grained sense disambiguation, namely in the case of all 6 senses of *line* and of *interest*. Using other test data is also intended. The choice of the dependency type to be used in noun disambiguation should also be subject to further investigation, especially in establishing if a connection exists between the frequency of occurrence of a dependency type and disambiguation accuracy. Augmenting the role of linguistic knowledge in informing the construction of this semantic space is also a future goal.

[3] In the subcorpus corresponding to the three chosen senses of *line* the *amod* relation occurs 1638 times while the *nn* relation occurs 1657 times. In the *interest* subcorpus the *amod* relation occurs 5410 times while the *nn* relation occurs 4634 times.

[4] Preotiuc and Hristea (2012) use the following notation: *n-w-t* represents the set containing the top *t* words occurring in N-grams together with the word *w*.

References

Bruce, Rebecca and Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* pages 139-145, Las Cruces, New Mexico.

Chen, Ping, Wei Ding, Chris Bowes, and David Brown. 2009. A fully unsupervised word sense disambiguation method using dependency knowledge. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28-36, Boulder, Colorado.

Dempster, Arthur, Nan Laird and Donald Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1-38.

Hristea, Florentina, Marius Popescu, and Monica Dumitrescu. 2008. Performing Word Sense Disambiguation at the Border Between Unsupervised and Knowledge-based Techniques. *Artificial Intelligence Review*, 30(1):67-86.

Hristea, Florentina. 2009. Recent Advances Concerning the Usage of the Naïve Bayes Model in Unsupervised Word Sense Disambiguation. *International Review on Computers and Software*, 4(1):58-67.

Hristea, Florentina and Marius Popescu. 2009. Adjective Sense Disambiguation at the Border Between Unsupervised and Knowledge-based Techniques. *Fundamenta Informaticae*, 91(3-4):547-562.

Hristea, Florentina and Mihaela Colhon. 2012. Feeding Syntactic Versus Semantic Knowledge to a Knowledge-lean Unsupervised Word Sense Disambiguation Algorithm with an Underlying Naïve Bayes Model. *Fundamenta Informaticae*, 119(1):61-86.

Hristea, Florentina. 2012. *The Naïve Bayes Model for Unsupervised Word Sense Disambiguation. Aspects Concerning Feature Selection.* SpringerBriefs in Statistics Series, Springer.

Klein, Dan and Christopher Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423-430, Sapporo, Japan.

Leacock, Claudia, Geoffrey Towell, and Ellen Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA workshop on Human Language Technology,* pages 260-265, Princeton, New Jersey.

de Marneffe, Marie-Catherine and Christopher Manning. 2012. Stanford typed dependencies manual. Technical Report, Stanford University.

Mooney, Raymond. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 82-91, Philadelphia, PA.

Nastase, Vivi. 2008. Unsupervised All-words Word Sense Disambiguation with Grammatical Dependencies. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 757-762, Hyderabad, India.

Navigli, Roberto. 2009. Word Sense Disambiguation: a Survey. *ACM Computing Surveys*, 41(2):1-69.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161-199.

Preotiuc-Pietro, Daniel and Florentina Hristea. 2012. Unsupervised Word Sense Disambiguation with N-Gram Features. *Artificial Intelligence Review*, doi:10.1007/s10462-011-9306-y.

Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale.* Klincksieck, Paris.