Exploring The Effect of Bag-of-Words and Bag-of-Bigram Features on Turkish Word Sense Disambiguation

Bahar İLGEN
Istanbul Technical University
ilgenb@itu.edu.tr

Eşref ADALI
Istanbul Technical University
adali@itu.edu.tr

Abastract

Feature selection in Word Sense Disambiguation (WSD) is as important as the selection of algorithm to remove sense ambiguity. Bag-of-word (BoW) features comprise the information of neighbors around the ambiguous target word without considering any relation between words. In this study, we investigate the effect of BoW features and Bag-of-bigrams (BoB) on Turkish WSD and compare the results with the collocational features. The results suggest that BoW features yield better accuracy for all the cases. According to the comparison results, collocational features are more effective than both BoW and the BoB features on disambiguation of word senses.

Key words: *Word Sense Disambiguation, feature selection, supervised methods, bag-of-word features.*

1 Introduction

The determination of proper sense label is required in almost all applications of Natural Language Processing (NLP) area. Machine Translation (MT), Information Retrieval (IR), Information Extraction (IE), Semantic Annotation (SA) and Question Answering (QA) are some of the NLP branches that benefit from WSD. The performance of these applications depends on the performance of WSD unit.

Gönderme ve kabul tarihi: 22.09.2014-25.10.2014

The basic approaches for WSD comprise the supervised, unsupervised and knowledge based methods. The selection of the proper method can be considered the application and the resources available. The knowledge based methods primarily rely on resources such as dictionaries, ontologies and thesaurus. These methods do not need to use corpus evidence. On the other hand, unsupervised methods utilize external information and work on raw corpora. Supervised methods use sense annotated data to train from. Although supervised methods yield superior results, the number of annotated corpora are too few for the majority of the natural languages. As a result, unsupervised methods have gained attention recently, since the annotation scheme is expensive and labor intensive. There is also one group of approaches of semi-supervised (or minimally supervised) methods which utilize a small amount of sense annotated data and expand the annotated part iteratively.

WSD can also be classified considering two variants: (1) Lexical sample task, and (2) all-words task. The first approach focuses on the disambiguation of the previously selected words. Machine Learning (ML) methods are usually preferred to handle these tasks since both the words and senses are limited. The labeled portion of the dataset is used the train classifier. Then the unlabeled portion of samples can be labeled using classifier. On the

other hand, all-words approach disambiguates all the words in a running text.

Knowledge is the central component to remove sense ambiguity of the words. It may be lexical or learned world knowledge. Sense frequency, concept trees, selectional restrictions and the POS information are some of the examples of lexical knowledge category. Learned knowledge category refers the information such as "Indicative words", "syntactic features" and "domain specific knowledge"[1]. Unsupervised methods usually utilize lexical knowledge sources while supervised methods use world knowledge. But in practice different combinations of the knowledge can be used in WSD systems.

There are two important decisions to be considered for a WSD system: the selection of learning algorithm and the set of features to be used. ML techniques can be used to automatically acquire disambiguation knowledge of the corpus-based WSD. And the several resources such as sense labeled corpora, dictionaries and other linguistic resources can be used for a typical WSD system. Supervised methods can be grouped into categories considering the induction strategy they use. These methods comprise probabilistic models, similarity based methods, linear classifiers and Kernel based methods and the methods based on some properties (i.e., one sense per collocation/discourse, attribute redundancy, decision lists/trees, rule combination etc.).

WSD introduces additional difficulties comparing to POS tagging or syntax parsing since each word is associated with unique meaning. That means a complete training set requires huge number of examples. This case is also known as language sparsity problem. This language sparsity problem can be handled with the selection of proper features in training algorithms.

In the scope of this study, we investigate the impact of bag-of-word and bag-of-bigram features on disambiguating senses. The rest of the paper is organized as follows. In section-2 related work has been summarized. Section-3 and Section-4 describe the dataset and features respectively. In Section-5, experimental results have been presented. Finally, Section-6 draws the conclusion.

2 Related Work

Feature selection has a critical importance in terms of correctly discriminating senses or categorizing them into proper labels. There are several studies to investigate the impact of feature selection strategies on WSD [2-7].

The impact of the features can be investigated by analyzing two aspects; feature type and the window size of the context. Selected features were classified as topical and local features in [8]. Topical features are usually extracted by checking the presence of keywords occurring anywhere in the sentence. The sentences around the ambiguous headword are taken as context. Local features comprise the information such as POS tagging, syntactic and semantic features for the neighbor words around headword.

In [9], main feature types have been grouped into local features, syntactic dependencies and global features. In total, six feature sets have been investigated including the bag-of-words, local collocations, bag-of-bigrams, syntactic dependencies, all features except bag-of-words and all features. They used different editions of Senseval [1] datasets in order to conduct experiments. The Lexical Sample data of the Senseval-2 has been used for parameter

[1] http://www.senseval.org

tuning. All-words and Lexical Sample datasets of Senseval-3 have been used for testing. It is reported that "all-features" set is the best single classifier for every method except one. It is also stated that local collocational features discriminate better than bag-of-word features for separate feature sets.

In[10], the impact of collocational features have been investigated on Turkish. The root forms and the POS information of the target word and its' neighbors have been used at encoding grammatical local lexical features. These features have been extracted from the text which is segmented into POS tagged units. The target word itself, the words within ±4 positions of the target word and the corresponding POS tags have been used in the study. Turkish Lexical Sample dataset (TLSD) have been used in the experiments. Figure-1 shows the sample window scope for the collocational features.

```
kriz: (Noun)(A3pl)(Pnon)(Nom)
sonra: (Noun)(Zero)(A3sg)(P3sg)(Loc)
büyük: (Adj)
şirket: (Noun)(A3pl)(Pnon)(Gen)
<HEAD-SENSE='baş'-SENSE_TDK_NO="2"…
baş: (Noun)(A3sg)(P3sg)(Loc)
</HEAD>
bulun: (Adj)(PresPart)
yönetici: (Noun)(A3pl)(Pnon)(Gen)
görev: (Noun)(A3sg)(Pnon)(Nom)
değişim: (Noun)(A3pl)(P3sg)(Nom)
```

Figure-1. Window scope for the colllocational features.

As being a member of agglutinative languages, Turkish is based on suffixation. And grammatical functions of the language are generated adding proper suffixes to the stems. As a result, number of POS features may be excessive. Because of the agglutinative property with inflectional and derivational suffixes in Turkish, two tools have been utilized. Firstly, a finite-state two level Turkish morphological analyzer has been used for morphological decomposition [11]. Then a disambiguation tool has been used since the output of the morphological analyzer is ambiguous [12].

3 Dataset

TLSD has been used in the experiments of this study. This dataset has been gathered to conduct our previous studies on Turkish WSD. TLSD comprises the highly ambiguous noun and verb samples of Turkish. These words were selected by considering the polysemous Turkish words in [13] and the polysemy degree of the words in Dictionary of Turkish Language Association (TLA) [14]. The results of our simple analysis on dictionary of TLA show that the average polysemy degree for Turkish is 3.53. The polysemy degrees of TLSD are calculated as 10.67 and 26.53 for noun and verb sets respectively. Both noun and verb groups in the dataset include 15 ambiguous words each of which has at least 100 samples. The samples have been gathered from Turkish websites on health, education, sports and news. We follow "one sense per sample principle" and each sample has only one sense of the ambiguous word. The ambiguous words in TLSD noun and verb sets are shown below (Table-1).

Table-1. Ambiguous noun and verb sets of TLSD.

Nouns:	*Açık, baskı, baş,derece, dünya,el, göz,hat, hava, kaynak, kök, kör, ocak, yaş, yüz*
Verbs:	*Aç, al, at, bak, çevir, çık, geç, gel, gir, gör, kal, ol, sür, ver, yap*

In the scope of the work, we also investigated the effective number of bag-of-word features and determined the most frequent content words as features. The most frequent 100, 75, 50 and 25 content words have been taken as features. We used vectors for the corresponding sizes and repeated the experiments (Figure-2). These vectors are initialized by assigning "0" to each cell. Then the values are incremented by "1" if the feature

exists in the lexical sample. Our findings suggest that the most frequent 75 and 100 content words yielded better accuracy for noun and verb sets respectively.

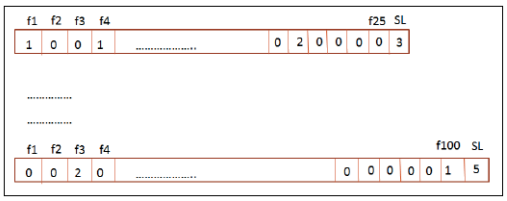

Figure-2. Bag-of-words feature vectors.

4 Features

4.1 Bag-of-word Features

Bag-of-word features comprise words around target word without considering their relations, grammar and even word order. Unordered set of words serve as features. The value of any feature is determined by counting the number of times that they occur for a given context. The context is fixed window with the ambiguous word as center of other words. For the experiments of this study, we followed the given steps:

- *Removing stopwords from samples*
- *Morphological analysis of dataset.*
- *Removing ambiguity after morphological analysis.*
- *Determination of features and encoding samples using them.*
- *Applying algorithms which we utilized for other features.*

4.2 Bag-of-bigram Features

We gathered bag-of-bigram features by following the similar steps with the bag-of-words features. After eliminating the stopwords, bigram words of the lexical samples have been exracted. Then we obtained most frequent bigram words. The only difference for BoB features is that we took the more features than the bag-of-word features since the features are more sparse. The number of features have been chosen considering the observation frequency of the bigrams and taken as approximately between 350 and 500 for each ambiguous word.

5 Experiments

After determining the number of effective bag-of-word features, we investigated the optimal window size to consider. We adjust the samples to take different values of ±n words (preceding and following words for values; ±30, ±15, ±10 and ±5) around target word. Our experiment on varying window sizes show that the best window size for noun and verb sets is 5. We kept this setting for the BoW features but took whole samples for the BoB features since the features are more sparse. Naïve Bayes, IBk, Support Vector Machines and tree based methods (J48 and FT algorithms) have been used in the experiments. Figure-3 shows the accuracy results for BoW and BoB features of Turkish nouns. Figure-4 displays the similar results for Turkish verb set. MFB represent the most frequent baseline.

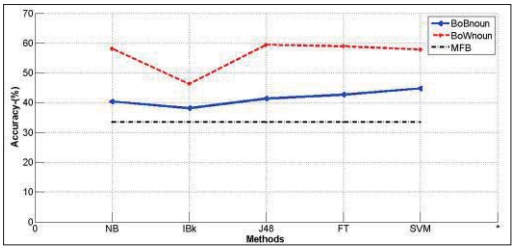

Figure-3. Accuracy(%) results of noun set.

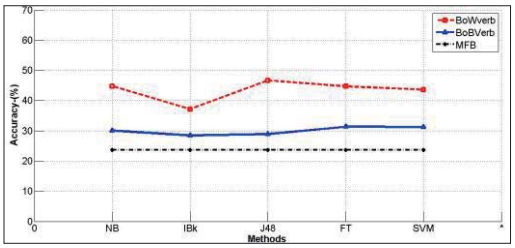

Figure-4. Accuracy(%) results of verb set

Table-2 and Table-3 summarizes the accuracy (%) results on TLSD for three feature sets. The results of the noun and verb sets are presented in Table-2 and Table-3 respectively. MFB values for noun and the verb sets are 33.47(%) and 27.60(%).

Table-2. Comparison of feature sets on Turkish nouns.

Feature	NB	IBk	J48	FT	SVM
BoW	58.1	46.3	59.4	58.9	57.8
BoB	40.4	38.1	41.3	42.6	44.8
Colloc	60.6	53.9	61.0	73.5	69.0

Table-3. Comparison of feature sets on Turkish verbs.

Feature	NB	IBk	J48	FT	SVM
BoW	44.8	37.2	46.7	44.7	43.6
BoB	30.1	28.4	28.9	31.3	31.2
Colloc	46.5	43.1	66.0	67.3	58.6

6 Conclusion

It is known that the features extracting from context words play important role on isolating senses. And there are many features to consider that can contribute the meaning of a given word. In this study, we investigated the impact of bag-of-word and bag-of-bigram features on Turkish WSD systems. Then we compare the results of these two groups with the results of collocational features. Our findings suggest that bag-of-word features yielded better results than bag-of-bigrams. The results also show that the collocational features are more efficient than both the bag-of-words and bag-of-bigram features. It is thought that the results of the bag-of-word and bag-of-bigram features can be improved by combining diverse set of features.

7 References

1. Zhou, X. and H. Han. *Survey of Word Sense Disambiguation Approaches*. in *FLAIRS Conference*. 2005.
2. Orhan, Z. and Z. Altan. *Effective Features for Disambiguation of Turkish Verbs*. in *IEC (Prague)*. 2005.
3. ORHAN, Z. and Z. Altan, *Determining Effective Features for Word Sense Disambiguation in Turkish.* IU-Journal of Electrical & Electronics Engineering, 2011. **5**(2): p. 1341-1352.
4. Agirre, E., O.L. de Lacalle, and D. Martınez. *Exploring feature spaces with svd and unlabeled data for Word Sense Disambiguation*. in *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP'05)*. 2005.
5. Turdakov, D.Y., *Word sense disambiguation methods.* Programming and Computer Software, 2010. **36**(6): p. 309-326.
6. Suárez, A. and M. Palomar, *Feature selection analysis for maximum entropy-based wsd*, in *Computational Linguistics and Intelligent Text Processing*. 2002, Springer. p. 146-155.
7. Dang, H.T., et al. *Simple features for Chinese word sense disambiguation*. in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. 2002. Association for Computational Linguistics.
8. Dang, H.T. and M. Palmer. *Combining contextual features for word sense*

disambiguation. in *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*. 2002. Association for Computational Linguistics.

9. Agirre, E., O.L. de Lacalle, and D. Martínez. *Exploring feature set combinations for WSD*. in *Proc. of the SEPLN*. 2006.
10. Ilgen, B., E. Adali, and A. Tantug. *The impact of collocational features in Turkish Word Sense Disambiguation*. in *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on*. 2012. IEEE.
11. Oflazer, K., *Two-level description of Turkish morphology*. Literary and linguistic computing, 1994. **9**(2): p. 137-148.
12. Yuret, D. and F. Türe. *Learning morphological disambiguation rules for Turkish*. in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. 2006. Association for Computational Linguistics.
13. Göz, İ., *Yazılı türkçenin kelime sıklığı sözlüğü*. Vol. 823. 2003: Türk Dil Kurumu.
14. Sözlük, G.T., *Türk Dil Kurumu,[çevrimiçi]*. Elektronik adres: http://tdk. org. tr/tdksozluk/sozbul. ASP, 2005.