GÜFBED/*GUSTIJ* (2018) CMES 2018 SEMPOZYUM EK SAYISI: 46-52
DOI: 10.17714/gumusfenbil.430034 Araştırma Makalesi / *Research Article*

Morphological Disambiguation of Turkish with Free-order Co-occurrence Statistics

Serbest Sırada Birliktelik İstatistiklerinin Kullanımıyla Türkçe'nin Biçimbirimsel Belirsizliği'nin Giderilmesi

Enis ARSLAN*[1,a], Umut ORHAN[1,b], B. Tahir TAHİROĞLU[2,c]
[1]*Çukurova Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, 01130, Adana*
[2]*Çukurova Üniversitesi, Türk Dili ve Edebiyatı Bölümü, 01130, Adana*

• Geliş tarihi / *Received*: 02.06.2018 • Düzeltilerek geliş tarihi / *Received in revised form*: 12.09.2018 • Kabul tarihi / *Accepted*: 09.10.2018

Abstract
In this article, a solution to the morphological ambiguity problem which occurs frequently in morphologically complex languages like Turkish is proposed. Generally, statistical methods are applicable for these tasks which maximize the information, obtained for a probable word order sequence in a sentence. The decision in selection of the method for calculation of the probabilities and the sequence selection method depends on the nature of the language. By using the co-occurrence statistics obtained from a semantic graph network which represents the lemmas of the sentences, the best word order sequence is selected from the alternatives. The non-ambiguous and free-word-order character of this network is helpful in determining the statistics independently. The probability values are obtained by using the Naive Bayes (NB) method and the selection of each word sequence is achieved by maximization, in the inspiration of the Viterbi algorithm.

Keywords: *Co-occurrence, Morphological ambiguity, Naive Bayes, the Viterbi algorithm*

Öz
Bu makalede, Türkçe gibi biçimsel olarak karmaşık yapıda olan dillerde sıklıkla karşılaşılan biçimbirimsel belirsizlik problemi için bir çözüm önerilmiştir. Genellikle, bu tipte bir problemin çözümü için bir cümledeki muhtemel kelime sıralarından uygun olanın seçilmesi için bilgiyi maksimuma çıkaran istatistiksel yöntemler uygulanmaktadır. Olasılıkların hesaplanması ve uygun sıranın seçilmesi için tercih edilecek metot uygulanacak dilin doğasına bağlıdır. Cümlelerde geçen kelimelerin madde başlarının oluşturduğu bir anlamsal çizgeden elde edilen birliktelik istatistikleri kullanılarak alternatifler arasından uygun olan kelime sıra dizilimi seçilmektedir. Bu çizge ağının belirsizlik içermeyen serbest sıralı karakteri istatistiklerin bağımsız olarak hesaplanmasında oldukça faydalıdır. Olasılıksal değerler Naive Bayes (NB) yöntemi kullanılarak elde edilmekte ve her kelime sıraları arasından uygun olanının, Viterbi algoritmasından esinlenilerek, maksimumu seçilmektedir.

Anahtar kelimeler: *Birliktelik, Biçimbirimsel belirsizlik, Naive Bayes, Viterbi algoritması*

*[a] Enis ARSLAN; enisarslan@gmail.com; Tel: (0322) 338 60 60; orcid.org/0000-0002-2609-3925
[b] orcid.org/0000-0003-1882-6567 [c] orcid.org/0000-0002-7956-3257

1. Introduction

Natural languages, ambiguity occurs according to the researched language's nature. In morphologically simple languages, especially like English, polysemous words can have different senses according to their usage purposes in texts. This kind of ambiguity problem is solved by Word Sense Disambiguation (WSD) methods. On the other hand, for the morphologically complex languages like Turkish, Finnish and Hungarian the output of the morphological analyser can cause more than one solution for a word which causes ambiguity. Lexical or syntactical disambiguation methods help in removing this kind of ambiguities. Recently graphs are getting more popular in the application of lexical disambiguation and WSD in many research (Sinha and Mihalcea, 2007; Minkov et al., 2006; Fan et al., 2011; Hessami et al., 2011).

The complex structure of the languages causes the new words to be included and the useless ones to be removed from usage. Sometimes, some relations between words are stronger and sometimes weaker as time goes by. Complex networks are ideal for modelling these languages according to the necessity. In general, the simplest design is created by using the co-occurrence relations between words in a text (Borge-Holthoefer and Arenas, 2010). It is possible to produce a co-occurrence graph by applying a fixed-width window in a sentence or document (Beliga et al., 2015). Co-occurrence graphs are used in Natural language processing (NLP) fields like text summarization (Mihalcea and Tarau, 2004), indexing (Matsuo et al., 2001), keyword and keyphrase extraction (Lahiri et al., 2014; Litvak et al., 2011), disambiguation (Duque et al., 2018; Martinez-Romo et al., 2011).

In this study, a directed and weighted graph is composed by ignoring the neighbouring sequences, in an unsupervised fashion. This graph is used in morphological disambiguation of Turkish sentences which have at least one ambiguous word. The second section of the study summarizes the related work for this field, the third section describes the methodology, the fourth section gives the experimental results and the fifth section concludes the study.

2. Related Work

Using co-occurrence of words in WSD takes place in many NLP studies. In (Niwa and Nitta, 1994), they have concluded that using co-occurrence vectors instead of distance vectors is advantageous. In machine translation field, query disambiguation tests provide better results when co-occurrence statistics is applied (Ballesteros and Croft, 1998). In (Duque et al., 2018), the paper abstracts of bio-medical science are used to construct a co-occurrence network of concepts for WSD and the study achieved a 10% improvement in accuracy. In another study (Martinez-Romo et al., 2011), a co-occurrence graph is established to cluster words with similar meanings and by assigning weight values of statistical significance using in WSD and Word Sense Induction (WSI). When morphological disambiguation of Turkish is considered, (Sak et al., 2007) provide the best reported accuracy value in their study. They collect the trigram statistics of the word features and roots in the sentences and select the n-best sequences by using the Viterbi algorithm.

3. Methodology

In this study, a semantic graph of Turkish lemmas is composed by using the co-occurrence relations existing in the sentence level. These relationship statistics are used as an input to the Naive Bayes (NB) method to be used in the selection of the correct sequence in an ambiguous sentence.

Lemma is the dictionary form of a word in a natural language. Lemmatization is used as a term to discover the lemmas of the words. In the first stage of the study, lemmas of Turkish Language Association (TLA) are added to a graph database to be used in lemmatization. For the lemmatization of words in corpus, we have used the finite state automatas which are described in (Eryiğit, 2012). In the second stage, the sentences in the corpus (Tahiroğlu et al., 2014) are lemmatized and lemmas are used as the main component of the semantic graph by establishing connections in-between. Co-occurrence provides frequency increases of the graph. As a corpus, 57,000 sentences of Turkish newspaper texts are processed.

3.1 *Composition of the Semantic Graph*

A semantic graph which has unambiguous character is composed by connecting the lemmas (If it is a lemma connect itself) of a sentence with a relation property named 'COOCCUR'. This property is used to calculate co-occurrence frequencies for further processing.

In the beginning, during the process of each sentence, a tokenization is applied. Following the

tokenization, all lemmas occurred in the sentence are collected as a list. The inflected words which have more than one lemma alternative (ambiguity) are ignored to obtain a non-ambiguous semantic graph. All the lemmas collected in the list are connected to each other with 'COOCCUR' relation type.

Co-occurrence graph is trained as follows:

1. Take a sentence from the sentence dataset
2. Select a token from the sentence.
   i. If the token is an inflected word lemmatize it. If there is only one lemma candidate of the inflected word (non-ambiguity) add this lemma to the sentence lemma list. If there is more than one candidate (ambiguity) do nothing.
   ii. If the token is a lemma add it to the lemma list
3. Connect all the lemmas in the lemma list with a 'COOCCUR' relation type in the graph. relFreq=1 (If this relation occurs only increment relFreq value)
4. Go to 1
5. End function

'relFreq' value of 'COOCCUR' relationship represents the lemma pairs which co-occur in a sentence. Each lemma node name in the list is checked with the others for a sentence and if there does not exist a relation between lemmas a 'COOCCUR' relation with type 'relFreq=1' is established. If there exists a relation 'relFreq' frequency value is incremented.

In Figure 1, other lemmas connected to the lemma node named 'festival' (fest in English) in the semantic graph can be seen. As seen in the Figure, nearly all other lemmas are semantically related to the lemma 'festival'.

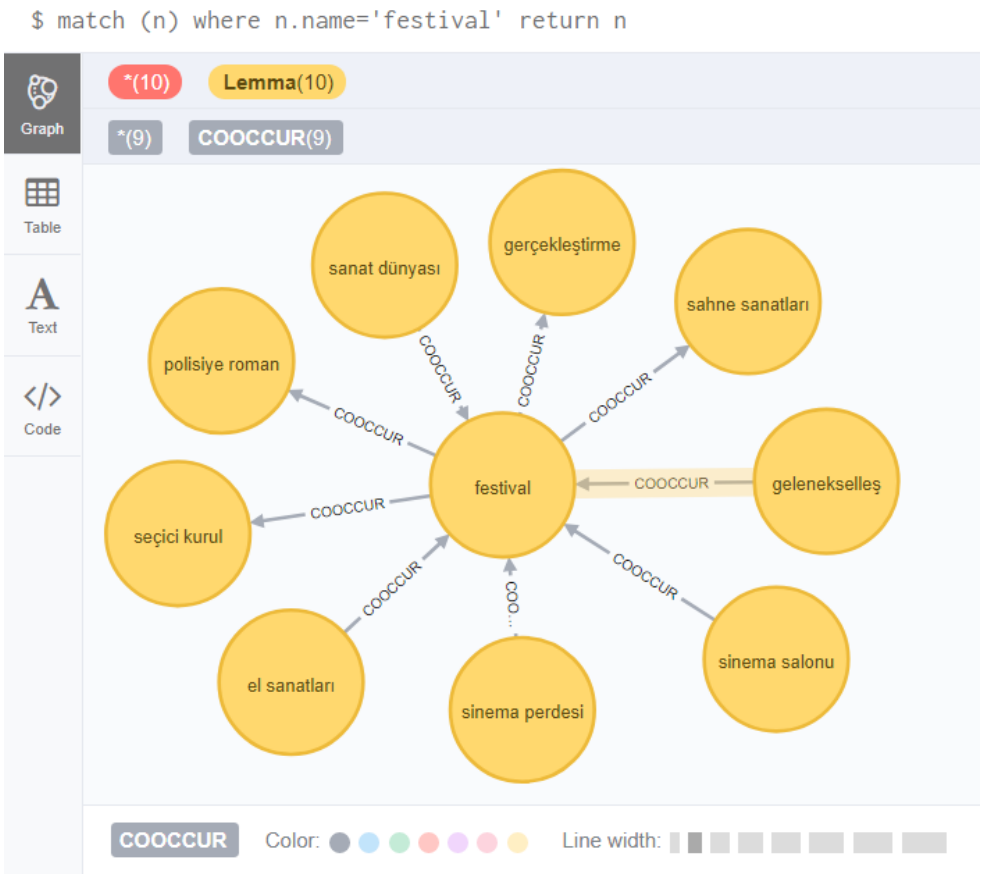

Figure 1. Lemma nodes connected to lemma node named 'festival' in semantic graph

3.2 *Disambiguation Methodology*

At the beginning of the disambiguation process, each test sentence is lemmatized and all alternative word sequences (permutation) are detected. An example sentence is introduced in Figure 2 which has 6 alternatives:

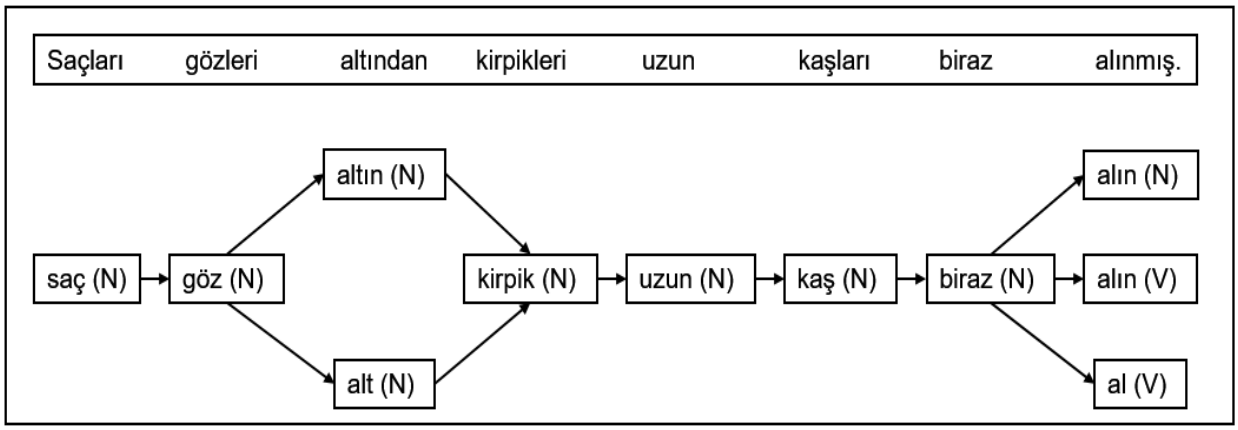

Figure 2. An example sentence of sequences

As seen in Figure 2 there are two lemma alternatives for the word 'altından' and three alternatives for the word 'alınmış'. When 3 is multiplied by 2 results is 6 representing the permutation count for this sentence.

Following the detection of the sequences, Naïve Bayes values for each lemma sequence are calculated using the lemma relation statistics obtained from the co-occurrence graph. The calculations are done as shown in Equations (1) and (2):

$$P(q_1^1, q_2^1) = \frac{f(q_1^1, q_2^1)}{f(q_1^1, q_2^1)} = 1 \qquad (q_1^1\text{=saç ve } q_2^1\text{=göz}) \tag{1}$$

$$P(q_1^1, q_3^1) = \frac{f(q_1^1, q_3^1)}{f(q_1^1, q_3^1) + f(q_1^1, q_3^2)} \qquad (q_1^1\text{=saç ve } q_3^1\text{=altın}) \tag{2}$$

In the preceding equations, f represents the frequency value (relFreq) of the relation between q_1^1 and q_2^1 in co-occur graph. All the permutations of the example sentence in Figure 2 can be seen in Figure 3.

saç (N)	göz (N)	altın (N)	kirpik (N)	uzun (N)	kaş (N)	biraz (N)	alın (N)
saç (N)	göz (N)	alt (N)	kirpik (N)	uzun (N)	kaş (N)	biraz (N)	alın (V)
saç (N)	göz (N)	altın (N)	kirpik (N)	uzun (N)	kaş (N)	biraz (N)	al (V)
saç (N)	göz (N)	alt (N)	kirpik (N)	uzun (N)	kaş (N)	biraz (N)	alın (N)
saç (N)	göz (N)	altın (N)	kirpik (N)	uzun (N)	kaş (N)	biraz (N)	alın (V)
saç (N)	göz (N)	alt (N)	kirpik (N)	uzun (N)	kaş (N)	biraz (N)	al (V)

Figure 3. Six alternative sequences of the example sentence

Six Naive Bayes values are calculated for each sequence in Figure 3 as:

$$P(Q_1) = \prod_{i=1}^{N-1} \prod_{j=i+1}^{N} P(q_i, q_j) \tag{3}$$

4. Experimental Results

4.1. Training of the Semantic Graph

Statistics of the semantic graph built by training all the sentences in the corpus are listed in Table 1:

Table 1. Statistics of the semantic graph

Lemma node count	79364
Connected Lemma count	18929
Co-occur relationship count	2121864
Trained sentence count	57K

The 79K TLA lemmas listed in Table 1 are used for lemmatization. 18929 of these lemmas are connected to each other with at least one 'COOCCUR' relation type after training.

4.2. Preparation of the Test Sentences

To obtain test sentences, at first, all ambiguous sentences from the sentence dataset are selected. The sentences which have fully inter-connected (detected by using the co-occurrence graph) lemmas are captured from the ambiguous sentences. The resulting sentences are the input to the disambiguation function as test sentences. The statistics for the test sentences are shown in Table 2:

Table 2. Test sentence statistics

Sentence length < 150	
Total sentence count	30000
Ambiguous sentence count	13725
Semantically connected sentence count	180
Ambiguity rate	45.75 %

As seen in Table 2, 30K portion of the 57K sentences is used to obtain test sentences. A sentence size of minimum 150 characters is applied as a filter to be able to select the sentences with at least a few words. Semantically connected sentences represent the sentences which include lemmas full-connected with at least one 'COOCCUR' relation in the semantic graph.

4.3. Test Results

180 ambiguous sentences are subject to the disambiguation process. Since there are errors due to lemmatization nearly half of the test sentences are useless. The remaining sentences are considered in accuracy calculation, as shown in Table 3:

Table 3. Disambiguation test results

Semantically connected sentence count	180
Permutations with equal value (ignored)	72
Incorrectly lemmatized sentences	70
Correctly disambiguated sentences	26
Success ratio	68.42 %

The test results are checked in a supervised fashion. The success ratio is 68.42% which can increase in value with more training. When the same test set is applied to a graph trained with 73K sentences there is one more correctly disambiguated sentence as shown in Figure 4:

```
1
2  Test output of a sentence trained with 57 K. sentences:
3  günün (günü) en (en) yüksek (yüksek) sıcaklıkları (sıcaklık)
4
5  Test output of a sentence trained with 73 K. sentences:
6  günün (gün) en (en) yüksek (yüksek) sıcaklıkları (sıcaklık)
```

Figure 4. Correct disambiguation with a larger train-set

As seen in Figure 4, the word 'günün' is selected as 'günü' lemma alternative (red frame). This erroneous selection is automatically corrected after the training with a larger data-set as the word 'gün'. This is because of the frequency increase between some re-occurring relations between some lemmas in the co-occurrence graph as training goes by.

For the disambiguation of Turkish language, (Sak et al., 2007)'s study improves the baseline study (93.61%) by providing an accuracy value of 96.80%. Statistics for the trigram models are provided by Markov method and Perceptron algorithm is used for training and ranking.

5. Conclusion

In this study, a disambiguation model is implemented on an unsupervised lemma graph database by using co-occurrence relations in-between. Co-occurrence frequency statistics are the main calculation parameter. A sentence with ambiguous words may have many alternative word sequences. Co-occurrence statistics are helpful in solving this kind of ambiguity.

When we consider the disambiguation of Turkish texts, features of words with their syntactical tags and root relationships are the main source of information in use. This study is different from (Sak et al., 2007)'s study, which is nearest in statistical means. We take the advantage of lemma-to-lemma relationships in a global context instead of trigram statistics of root-to-root relationships of words. Also we cannot profit from the syntactic features of words because of the nature of our algorithm. In our knowledge, there is not any work for morphological disambiguation of Turkish, exactly as the same of our methodology, which only relies on co-occurrence relationship of words in semantical means.

A dense training graph, composed of hundred thousands of words is needed to obtain more accurate results in the application of this methodology. More training will provide more clues in sentence disambiguation with the increased co-occurrence frequency values.

Acknowledgements

This study is a part of the research programme with project number 212E256, which is financed by the Scientific and Technological Research Council of Turkey (TUBITAK).

References

Ballesteros, L. and Croft, W.B., 1998. August. Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 64-71). ACM.

Beluga, S., Meštrović, A. and Martinčić-Ipšić, S., 2015. An overview of graph-based keyword extraction methods and approaches. Journal of information and organizational sciences, 39 (1), pp.1-20.

Borge-Holthoefer, J. and Arenas, A., 2010. Semantic networks: Structure and dynamics. Entropy, 12 (5), pp.1264-1302.

Duque, A., Stevenson, M., Martinez-Romo, J. and Araujo, L., 2018. Co-occurrence graphs for word sense disambiguation in the biomedical domain. Artificial intelligence in medicine.

Eryiğit, G., 2012. Biçimbilimsel Çözümleme. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5 (2).

Fan, X., Wang, J., Pu, X., Zhou, L. and Lv, B., 2011. On graph-based name disambiguation. Journal of Data and Information Quality (JDIQ), 2 (2), p.10.

Hessami, E., Mahmoudi, J. and Jadidinejad, A.H., 2011. Unsupervised graph-based word sense disambiguation using the lexical relation of WordNet. Int. J. Comput. Sci. Issues (IJCSI).

Lahiri, S., Choudhury, S.R. and Caragea, C., 2014. Keyword and keyphrase extraction using centrality measures on collocation networks. arXiv preprint arXiv:1401.6571.

Litvak, M., Last, M., Aizenman, H., Gobits, I. and Kandel, A., 2011. DegExt—A language-independent graph-based keyphrase extractor. In Advances in Intelligent Web Mastering–3 (pp. 121-130). Springer, Berlin, Heidelberg.

Martinez-Romo, J., Araujo, L., Borge-Holthoefer, J., Arenas, A., Capitán, J.A. and Cuesta, J.A., 2011. Disentangling categorical relationships through a graph of co-occurrences. Physical Review E, 84 (4), p.046108.

Matsuo, Y., Ohsawa, Y. and Ishizuka, M., 2001. November. Keyword: Extracting keywords from document s small world. In International Conference on Discovery Science (pp. 271-281). Springer, Berlin, Heidelberg.

Mihalcea, R. and Tarau, P., 2004. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

Minkov, E., Cohen, W.W. and Ng, A.Y., 2006. August. Contextual search and name disambiguation in email using graphs. In Proceedings of the 29th annual international ACM SIGIR conference on Research and

development in information retrieval (pp. 27-34). ACM.

Niwa, Y. and Nitta, Y., 1994. August. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In Proceedings of the 15th conference on Computational linguistics-Volume 1 (pp. 304-309). Association for Computational Linguistics.

Sak, H., Güngör, T. and Saraçlar, M., 2007, February. Morphological disambiguation of Turkish text with perceptron algorithm. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 107-118). Springer, Berlin, Heidelberg.

Sinha, R. and Mihalcea, R., 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In Semantic Computing, 2007. ICSC 2007. International Conference on (pp. 363-369). IEEE.

Tahiroğlu B.T. 2014, Türkçe Çevrim İçi Haber Metinlerinde Yeni Sözlerin (Neolojizm) Otomatik Çıkarımı. Derlem Dilbilim Uygulamaları, Özkan, B., Tahiroğlu, B. Tahir ve Özkan Ayşe Eda (Ed.), Karahan Kitabevi Yayınları, Adana, ss.1-22.