Vector Space Models in Detection of Semantically Non-compositional Word Combinations in Turkish

Levent Tolga Eren and Senem Kumova Metin(✉)

Faculty of Engineering, İzmir University of Economics,
Sakarya C., No. 156 Balçova, İzmir, Turkey
{tolga.eren, senem.kumova}@ieu.edu.tr

Abstract. The semantic compositionality presents the relation between the meanings of word combinations and their components. Simply, in non-compositional expressions, the words combine to generate a different meaning. This is why, identification of non-compositional expressions (e.g. idioms) become important in natural language processing tasks such as machine translation and word sense disambiguation.

In this study, we explored the performance of vector space models in detection of non-compositional expressions in Turkish. A data set of 2229 uninterrupted two-word combinations that is built from six different Turkish corpora is utilized. Three sets of five different vector space models are employed in the experiments. The evaluation of models is performed using well-known accuracy and F-measures. The experimental results showed that the model that measures the similarity between the vectors of word combination and the second composing word produced higher average F-scores for all testing corpora.

Keywords: Semantic compositionality · Vector space model · Turkish

1 Introduction

The compositionality is described to be the degree to which the features of the constituents of a multiword expression combine to predict the feature of the whole [1]. In other words, it is the amount of information that is hold by the constituents that enables or eases the prediction of the expression, especially the meaning of the whole expression.

In this study, the notion of compositionality is limited to semantic compositionality of expressions that is composed of two consecutive words known as bigram. In this perspective, compositionality is the degree of relation between the meaning of expression and the individual meanings of its constituents. In compositional expressions, the meaning of expression can be predicted from the meanings of its composing words. For example, the two-word expression *trafik ışığı* (Eng. traffic light) is a compositional expression (to some degree). The regarding expression corresponds to signaling devices positioned at road intersections, pedestrian crossings etc. to control the flow of traffic. A person who knows the dictionary-based definitions/meanings of the words *trafik* (Eng.

W. M. P. van der Aalst et al. (Eds.): AIST 2018, LNCS 11179, pp. 53–63, 2018.
https://doi.org/10.1007/978-3-030-11027-7_6

traffic) and *ışığı* (Eng. light) may guess that the expression points to an object/item that includes a lighting item and is related somehow to traffic. On the other hand, in non-compositional expressions, the combined meaning of words is unrelated to individual meanings of its components. For instance, the two-word (idiomatic) expression *kanı bozuk* (Eng. corrupt or evil by nature) is a fully non-compositional expression. Even if a person is a native speaker of Turkish, he may not predict the meaning of multiword expression by the meanings – dictionary-based definitions-of *kanı* (Eng. blood) and *bozuk* (Eng. spoilt). Though it is almost impossible to discover such non-compositional multiword expressions utilizing dictionaries, distributional hypothesis/semantics where each word is represented by a vector of words enables the regarding discovery. Simply in distributional semantics, a word is expressed commonly by a vector of its neighboring words targeted with their occurrence frequency and the context of a target word is defined as the vector of its neighboring words in a fixed window size. As a result, one can tell that given two words are similar if they have a similar distribution of contexts.

The objective of this study is exploring the performance of vector space models (VSM) in detection of non-compositional expressions in Turkish. In line with this objective, a dataset of 2229 bigrams is constructed from 6 different Turkish corpora by the use of occurrence frequency methods (chi square, occurrence frequency counts, point-wise mutual information and t-test). This dataset is annotated by 4 human judges. The dataset is utilized in the experiments of vector space models that are previously proposed to measure the semantic compositionality/non-compositionality in different languages.

In following sections, we are going to explore the related work on semantic compositionality, and then we will present the data set, our proposed models and evaluation measures. Following the experimental results, we will conclude the paper.

2 Related Work

In recent years, there has been a growing awareness in the natural language processing field about the problems related to semantic compositionality/non-compositionality. Such that special interest workshops have been arranged and discussed issues like automatically acquiring semantic compositionality [2]. In Table 1, a set of studies presented in Distributional Semantics and Compositionality Workshop (DiSCo 2011) [2] that not only provide the inspiration but also directed our work are given. As the studies in Table 1 and the other works on distributional semantics are examined, it is clearly seen that the majority of the works on non-compositionality are performed on English and/or German corpora. In DiSCo shared task, the samples in data set are separated into multiple classes based on the compositionality score (low in compositional ($0 < \text{score} < 37$). Medium in compositional ($36 < \text{score} < 75$) and high in compositional ($74 < \text{score}$)). Though no clear winner emerged in the task, it was examined that the approaches based on distributional semantics seemed to outperform those based on statistical association measures [3].

Following the shared task. Krčmář et al. [12] evaluated various distributional semantic approaches in compositionality detection and showed that LSA-based models

Table 1. A set of studies presented in DiSCo 2011 workshop.

#	The study	Approaches
1	Identifying collocations to measure compositionality [4]	Statistical association measures. t-score and pmi
2	Measuring the compositionality of bigrams using statistical methodologies [5]	A mix of statistical association measures
3	Measuring the compositionality of collocations via word co-occurrence vectors [6]	Unsupervised WSM, cosine similarity
4	(Linear) maps of the impossible: capturing semantic anomalies in distributional space [7]	Cosine similarity
5	Frustratingly hard compositionality prediction [8]	Support vector regression with COALS-based endocentricity features
6	Exemplar-based word-space model for compositionality detection [9]	Exemplar-based WSM, prototype-based WSM
7	Distributed structures and distributional meaning [10]	Distributed tree vector, distributed kernel tree vector
8	Two multivariate generalizations of point-wise mutual information [11]	Multi-way co-occurrences

perform quite well. There also exist a few studies in literature that employ parallel corpora or other resources such as Wiktionary to detect the degree of compositionality/non-compositionality in word combinations, especially in multiword expressions (e.g. [13–15]).

3 Detection of Non-compositional Word Combinations in Turkish

In this section, we will present our work to measure the performance of vector space models in detection of non-compositional word combinations in Turkish texts. To the best of our knowledge, this is the first work for Turkish language in this scope.

In following subsections, firstly the dataset and the procedure to prepare the set will be explained. Secondly the vector space models utilized in this study will be presented. In third subsection, evaluation methods will be given.

3.1 Data Set

In this study, the experiments on non-compositionality/compositionality in Turkish are limited to bigrams that are known as uninterrupted two-word combinations in text. BilCol [16], Bilkent [17], Ege, Leipzig [18], Metu [19] and Muder [20] corpus are utilized to obtain bigrams that will be used in experiments. Briefly, punctuation marks are removed from each corpus and the text is tokenized to obtain bigrams. Table 2 gives total number of tokens (unigrams) and bigrams in regarding corpora.

Table 2. The corpora used in experiments

Corpus	Size (Number of tokens)	Number of unique tokens	Number of unique bigrams
BilCol [16]	42414743	984434	11759532
Bilkent [17]	706443	94552	507758
Ege	2465285	259196	1637055
Leipzig [18]	13389049	745446	7350443
Metu [19]	1987447	212853	1388722
Muder [20]	638547	82145	437826

In order to decrease the number of samples (bigrams), a set of occurrence frequency methods (chi-square, occurrence frequency counts, point-wise mutual information and the t-test) is applied and a sorted list of bigrams is built individually for each method from each corpus. Each method yielded 1200 distinct bigrams from 6 different corpora. The best scoring 200 bigrams of sorted lists are selected to construct the final data set of 2229 bigrams.

The data set is annotated by four human judges and the interrater agreement among judges is measured by Fleiss kappa metric (kappa = $\sim$0.738). In this study, assuming that majority of multiword expressions (idioms. technical terms. named entities. some phrasal verbs) are non-compositional expressions as it was done in the study of Bu et al. [21]. Choueka [22] and Almi et al. [3], the human judges are asked to label the samples either as MWE (multiword expression) or non-MWE (non-multiword expression). A guide that directs judges to assign idiomatic expressions, named entities, technical terms, phrasal verbs and multi-word conjunctions in same class and all other word combinations in the other class is provided. Further information/details on the data set preparation may be found in [23]. In our experiments, all MWE labeled samples are accepted to be non-compositional and all non-MWE labeled samples are considered as compositional. As a result, we employed 1194 compositional bigrams (54% of the whole data set) and 1035 (46% of the whole set) non-compositional bigrams in our final set.

3.2 Proposed Method: Vector Space Models

In information retrieval systems, vector space models are commonly employed to represent the text documents as vectors of identifiers, such as, for example index terms. The vectors may be composed by occurrence frequencies of terms (term frequencies-tf), document frequencies (df), inverse document frequencies (idf) or a combination of these such as well-known tf.idf measure.

In this study, a vector of word occurrence frequencies will represent each bigram and/or its constituting words. The words that compose the vector are the ones that reside in same sentence with the target (bigram or a constituents of a bigram). For sentences that are relatively longer, a window size of 5 words is defined, in other words the preceding and/or following 5 words of the target are considered while building the regarding vector. To measure similarity between the vectors, we used cosine similarity as given below

$$sim\left(\overrightarrow{V_1} \cdot \overrightarrow{V_2}\right) = \frac{\overrightarrow{V_1} \cdot \overrightarrow{V_2}}{\left\|\overrightarrow{V_1}\right\|\left\|\overrightarrow{V_2}\right\|} \quad (1)$$

where V_i represents the i^{th} vector. In our experiments, we have used a normalized cosine similarity function that produces values in range [0.2] instead of range [−1.1] (0 indicates the exact similarity between vectors. 2 is vice versa).

The compositionality for a given bigram is measured by 5 different models that can be obtained from the following equation proposed by Reddy et al. [9] in DiSCo 2011 shared task [2]:

$$\begin{aligned}\alpha\left(\overrightarrow{w_1}.\overrightarrow{w_2}\right) = a + b * sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1}\right) + c * sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_2}\right) + d \\ * sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1} + \overrightarrow{w_2}\right) + e * sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1} * \overrightarrow{w_2}\right)\end{aligned} \quad (2)$$

where $\overrightarrow{w_1 w_2}$. $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$ represent the vectors of bigram, first and second words of the bigram respectively. In Reddy et al. [9], it is stated that the models that are derived from Eq. (2) outperform existing prototype-based models in DiSCo 2011 shared task [2].

In our experiments, we have employed 3 different sets of models that are derived from the Eq. 2. The brief definitions of model sets are given below.

Set 1 (S_1): This set includes 5 different models where the vectors include raw frequencies of neighboring words that reside in same window size with the target. The models are

$$\begin{aligned}&(S_1M_1) : sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1}\right)\\ &(S_1M_2) : sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_2}\right)\\ &(S_1M_3) : sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1} + \overrightarrow{w_2}\right)\\ &(S_1M_4) : sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1} * \overrightarrow{w_2}\right)\\ &(S_1M_5) : sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1}\right)\\ &\qquad + sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_2}\right)\\ &\qquad + sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1} + \overrightarrow{w_2}\right)\\ &\qquad + sim\left(\overrightarrow{w_1 w_2}.\overrightarrow{w_1} * \overrightarrow{w_2}\right)\end{aligned}$$

For example, in Model 1 (S_1M_1) the similarity of vectors that belong to bigram and the first word in bigram is measured. On the other hand, in Model 3, the similarity is measured between the vector of bigram and the vector that is obtained by summation of first and second word's vectors. Briefly, in models S_1M_1 and S_1M_2, the semantic similarity between the bigram and its constituents are measured. If the given bigram is non-compositional it is expected that the similarity score of these vectors will not be high. In models S_1M_3 and S_1M_4, the similarity between the bigram and a combined version of vectors (point-wise summation and multiplication) for the constituents is measured as in Mitchell and Lapata [24]. Finally in model S_1M_5, the results of the previous models are summed up.

Set 2 (S_2): This set includes models where the vectors of component words are refined. In this set, the refined vector for each constituent is built by the sentences that include the constituent but not the bigram. As a result the refined vector of $\overrightarrow{\mathrm{w}_1}$ is $\overrightarrow{\mathrm{w}_1'} = \overrightarrow{\mathrm{w}_1} - \overrightarrow{\mathrm{w}_1\mathrm{w}_2}$ and refined vector of $\overrightarrow{\mathrm{w}_2}$ is $\overrightarrow{\mathrm{w}_2'} = \overrightarrow{\mathrm{w}_2} - \overrightarrow{\mathrm{w}_1\mathrm{w}_2}$. The models in set 2 are listed as below

$$
\begin{aligned}
&(S_2M_1) : sim\left(\overrightarrow{w_1w_2}.\,\overrightarrow{w_1'}\right)\\
&(S_2M_2) : sim\left(\overrightarrow{w_1w_2}.\,\overrightarrow{w_2'}\right)\\
&(S_2M_3) : sim\left(\overrightarrow{w_1w_2}.\,\overrightarrow{w_1'} + \overrightarrow{w_2'}\right)\\
&(S_2M_4) : sim\left(\overrightarrow{w_1w_2}.\,\overrightarrow{w_1'} * \overrightarrow{w_2'}\right)\\
&(S_2M_5) : sim\left(\overrightarrow{w_1w_2}.\,\overrightarrow{w_1'}\right)\\
&\qquad + sim\left(\overrightarrow{w_1w_2}.\,\overrightarrow{w_2'}\right)\\
&\qquad + sim\left(\overrightarrow{w_1w_2}.\,\overrightarrow{w_1'} + \overrightarrow{w_2'}\right)\\
&\qquad + sim\left(\overrightarrow{w_1w_2}.\,\overrightarrow{w_1'} * \overrightarrow{w_2'}\right)
\end{aligned}
$$

Set 3 (S_3): In Set 3, while building the vectors of constituents irrelevant sentences in the corpus are removed. For example, building the vector of w_1, only the sentences that includes both w_1 and a word that is semantically related to w_2 are considered and the other sentences that includes only w_1 are removed.

In [9], it is stated that the composing words of a bigram may be used in a different context that may be unrelated to the regarding bigram. Reddy et al. [9] exemplified this by the bigram *traffic light*. The composing word *light* may occur in different context in corpus. And some of the occurrences may be unrelated to notion of *traffic light*. These unrelated occurrences tend to decrease the semantic relation between the composing words; *light* and *traffic*. In order to decide relevant occurrences of *light*, a group of words that appears in similar context of *traffic* is defined. This group of words will be named as context words from now on. While building the vector of *light,* the sentences where both *light* and at least one of the context words of *traffic* are selected, the other sentences where only *light* is observed are accepted to be in a context that is unrelated to *traffic light*.

In this study, for each composing word a group of context words is determined. The context words group includes the words that are most frequently co-occurring words with the regarding word. Simply, for each word, the most frequently co-occurring words are listed in the corpus, the stop words are removed from the list and finally the first five words are assigned as context words. The list of stop words that is given in [25] is used. Following models in Set 3 are determined:

$$
\begin{aligned}
(S_3M_1)&: sim\left(\overrightarrow{w_1w_2}.\overrightarrow{w_1}^r\right)\\
(S_3M_2)&: sim\left(\overrightarrow{w_1w_2}.\overrightarrow{w_2}^r\right)\\
(S_3M_3)&: sim\left(\overrightarrow{w_1w_2}.\overrightarrow{w_1}^r+\overrightarrow{w_2}^r\right)\\
(S_3M_4)&: sim\left(\overrightarrow{w_1w_2}.\overrightarrow{w_1}^r*\overrightarrow{w_2}^r\right)\\
(S_3M_5)&: sim\left(\overrightarrow{w_1w_2}.\overrightarrow{w_1}^r\right)\\
&\quad + sim\left(\overrightarrow{w_1w_2}.\overrightarrow{w_2}^r\right)\\
&\quad + sim\left(\overrightarrow{w_1w_2}.\overrightarrow{w_1}^r+\overrightarrow{w_2}^r\right)\\
&\quad + sim\left(\overrightarrow{w_1w_2}.\overrightarrow{w_1}^r*\overrightarrow{w_2}^r\right)
\end{aligned}
$$

where $\overrightarrow{w_k}^r$ represents the revised vector of kth word in bigram.

3.3 Evaluation

The evaluation of proposed models is performed in 3 steps. For each proposed model below steps are followed:

1. Bigrams are sorted according to the similarity score that is produced by the model in decreasing order. It is accepted that if the similarity score of a bigram is low, then it is non-compositional.
2. F1-score and accuracy are measured in a point-wise manner. Simply they are measured for set size N where N is increased from 1 to the total set size. The resulting N number of regarding evaluation values (F1 or accuracy) are summed up and divided by N to obtain average evaluation score.
3. Average F1-score and accuracy are compared to respectively F1 and accuracy scores of other models.

The F1 measure is given as

$$\mathrm{F1} = \frac{2TP}{2TP + FN + FP} \tag{2}$$

where *TP* is the number of true positives (samples that are both expected and predicted to belong to the same class (non-compositional or compositional)). *FN* is the number of false negatives (type 2 error). *FP* is the number of false positives (type1 error).

Accuracy (*A*) is measured as follows

$$A = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

where *TN* is the number of true negatives.

4 Experimental Results

In experiments, three sets of models are tested for 3 corpora of different sizes: Bilkent [17], Muder [20] and Metu [19] corpus. Table 3 presents accuracy and F1 scores obtained from the regarding corpora. In Table 3, bold cells represent the highest scores obtained from each corpus for the regarding evaluation measure.

Table 3. Average F1-measure and accuracy values obtained from Bilkent [17]. Muder [20] and Metu [19] corpora

Model	Bilkent		Muder		Metu	
	A	*F1*	*A*	*F1*	*A*	*F1*
S_1M_1	0.626	0.501	0.576	0.521	0.656	0.519
S_1M_2	0.615	0.490	0.565	0.513	0.650	0.513
S_1M_3	0.600	0.476	0.549	0.500	0.633	0.499
S_1M_4	0.591	0.469	0.551	0.503	0.624	0.492
S_1M_5	0.602	0.479	0.560	0.509	0.639	0.503
S_2M_1	0.694	0.564	0.580	0.527	0.709	0.570
S_2M_2	**0.703**	**0.570**	**0.595**	**0.537**	0.710	0.569
S_2M_3	0.694	0.565	0.576	0.524	0.704	0.565
S_2M_4	0.686	0.557	0.573	0.521	0.707	0.568
S_2M_5	0.697	0.567	0.583	0.529	**0.711**	**0.571**
S_3M_1	0.627	0.503	0.585	0.530	0.660	0.523
S_3M_2	0.628	0.504	0.589	0.533	0.662	0.524
S_3M_3	0.613	0.489	0.570	0.517	0.647	0.511
S_3M_4	0.615	0.493	0.576	0.522	0.650	0.516
S_3M_5	0.620	0.496	0.583	0.528	0.657	0.520

The number of bigrams that reside both in data set and Bilkent corpus is 957 in which 63.32% of bigrams is annotated as non-compositional. The highest averaged F1 and accuracy values are obtained by S_2M_2. Considering average F1 values (average $F_{Set1} = 0.483$, $F_{Set2} = 0.565$, $F_{Set3} = 0.497$) it is observed that in Bilkent corpus, Set 2 outperforms the other sets of models. As a result, it is possible to state that refined vectors of composing words; the vectors that are built by the sentences that include the constituents but not the bigram; are better representatives to detect non-compositionality.

Muder corpus includes 798 bigrams (56% non-compositional, 44% compositional) of the data set. The maximum F1 and accuracy values are obtained in model S_2M_2 similar to Bilkent corpus. In addition, though the difference in average F-values is not as high as the values in Bilkent corpus, it is observed that still Set 2 models generate higher F1 performance scores compared to other sets (average $F_{Set1} = 0.509$, $F_{Set2} = 0.528$, $F_{Set3} = 0.526$).

The third corpus in our experiments, Metu, includes 1129 of bigrams in data set. In Metu corpus, it is examined that the models in Set 2 are performing better compared to the other models (average $F_{Set1} = 0.505$, $F_{Set2} = 0.569$, $F_{Set3} = 0.519$), supporting the results in previous corpora. The best performing model is observed to be S_2M_5.

Based on the overall results of 3 corpora, it is examined that models in Set 2 (especially Model 2 – S_2M_2) are succeeding in Turkish corpora in this experimental set up. Though the size of the corpus changes the evaluation scores, the best set of models do not differ according to the corpus size.

To the best of our knowledge, since this is the first study that employs vector space models in detection of non-compositional word combinations in Turkish, there exist no other evaluation results for alternative vector space models. Nevertheless it may be stated that the range of highest average F1 scores ([0.537 0.571]) is promising compared to the performance scores reported in previous frequency-based studies employing different Turkish data sets. For example, in [26] where Google search engine is used as an additional resource, it is reported that considering web-based frequency metrics, highest F scores are observed to be in range [0.570 0.585] and the highest F score when corpus-based frequency is utilized, is given as 0.57.

5 Conclusion

In this study, we analyzed the semantic compositionality/non-compositionality in Turkish by vector space models. We introduced three sets of 5 different VSMs that assess the non-compositionality in Turkish. VSMs of Set 2; the models where the vector of composing words are built by ignoring the sentences that hold the word combination; are observed to provide better performance results compared to other models. It is also examined that as the size of the corpus increases, the difference in performances of successful and unsuccessful methods becomes more significant.

Due to the high time and space complexity of the algorithms that are used to implement models, we were unable to work on larger corpus. As a future work, we are planning to repeat our experiments in larger corpora and with different settings (e.g. windows size. stemmed/surface formed corpus. binary/weighted vectors. unigrams/bigrams/trigrams). Moreover, we plan to evaluate the performance of word-embedding models (e.g. word2vec, skip-gram models) in Turkish data sets [27].

References

1. Baldwin, T.: Compositionality and multiword expressions: six of one, half a dozen of the other? In: COLING/ACL 2006 Workshop on MWEs, Invited Speech (2006)
2. Biemann, C., Giesbrecht, E.: Distributional semantics and compositionality 2011: shared task description and results. In: Proceedings of the Workshop on Distributional Semantics and Compositionality, pp. 21–28 (2011)
3. Almi, P., Snajder, J.: Determining the semantic compositionality of croatian multiword expressions. In: 9th Language Technologies Conference Information Society, IS 2014 (2014)
4. Pedersen, T.: Identifying collocations to measure compositionality: shared task system description. In: Proceedings of the Workshop on Distributional Semantics and Compositionality, pp. 33–37 (2011)

5. Chakraborty, T., Pal, S., Mondal, T., Saikh, T.: Shared task system description: measuring the compositionality of bigrams using statistical methodologies, pp. 38–42. Association for Computational Linguistics (2011)
6. Maldonado-guerra, A., Emms, M.: Measuring the compositionality of collocations via word co-occurrence vectors: shared task system description, pp. 48–53. Association for Computational Linguistics (2011)
7. Vecchi, E.M., Baroni, M., Zamparelli, R.: (Linear) maps of the impossible: capturing semantic anomalies in distributional space. In: Proceedings of the Workshop on Distributional Semantics and Compositionality, pp. 1–9 (2011)
8. Johannsen, A., Alonso, H.M., Rishøj, C., Søgaard, A.: Shared task system description: frustratingly hard compositionality prediction. In: Proceedings of DiSCo, pp. 29–32 (2011)
9. Reddy, S., McCarthy, D., Manandhar, S., Gella, S.: Exemplar-based word-space model for compositionality detection: shared task system description. In: Proceedings of the Workshop on Distributional Semantics and Compositionality, pp. 54–60. Association for Computational Linguistics (2011)
10. Zanzotto, F.M., Dell'Arciprete, L.: Distributed structures and distributional meaning. In: Proceedings of the Workshop on Distributional Semantics and Compositionality, DiSCo 2011, Portland, Oregon, pp. 10–15 (2011)
11. Van De Cruys, T.: Two multivariate generalizations of pointwise mutual information. In: Proceedings of the Workshop on Distributional Semantics and Compositionality, pp. 16–20 (2011)
12. Krčmář, L., Ježek, K., Pecina, P.: Determining compositionality of word expressions using various word space models and measures. In: Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, Sofia, Bulgaria, 9 August 2013, pp. 64–73 (2013)
13. Salehi, B., Cook, P., Baldwin, T.: Detecting non-compositional MWE components using Wiktionary. In: EMNLP (2014)
14. Salehi, B., Askarian, N., Fazly, A.: Automatic identification of Persian light verb constructions. In: Gelbukh, A. (ed.) CICLing 2012. LNCS, vol. 7181, pp. 201–210. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28604-9_17
15. de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., Villavicencio, A.: Alignment-based extraction of multiword expressions. Lang. Resour. Eval. **44**, 59–77 (2010)
16. Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C., Uyar, E.: New event detection and topic tracking in Turkish. J. Am. Soc. Inf. Sci. Technol. **61**, 802–819 (2010)
17. Tur, G., Hakkani-Tur, D., Oflazer, K.: A statistical information extraction system for Turkish. Nat. Lang. Eng. **9**, 181–210 (2003)
18. Quasthoff, U., Richter, M., Biemann, C.: Corpus portal for search in monolingual corpora. In: Proceedings of 5th International Conference on Language Resources and Evaluation, pp. 1799–1802 (2006)
19. Say, B., Zeyrek, D., Oflazer, K., Umut, Ö.: Development of a corpus and a treebank for present-day written Turkish. In: Proceedings of the Eleventh International Conference of Turkish Linguistics (2002)
20. Dinçer, B.T.: Türkçe için istatistiksel bir bilgi geri-getirim sistemi (2004)
21. Bu, F., Zhu, X., Li, M.: Measuring the non-compositionality of multiword expressions. In: Proceedings of the 23rd International Conference on Computational Linguistics, Coling 2010, pp. 116–124 (2010)
22. Choueka, Y.: Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In: RIAO, pp. 609–624 (1988)
23. Metin, S.K., Taze, M.: A procedure to build multiword expression data set. In: 2nd International Conference on Computer and Communication Systems, pp. 46–49 (2017)

24. Mitchell, J., Lapata, M.: Vector-based models of semantic composition, pp. 236–244. Association for Computational Linguistics (2008)
25. Turkish Stopwords. http://www.ranks.nl/stopwords/turkish
26. Aka Uymaz, H., Metin, S.K.: A comprehensive analysis of web-based frequency in multiword expression detection. Int. J. Intell. Syst. Appl. Eng. (2017, in print)
27. Salehi, B., Cook, P., Baldwin, T.: A word embedding approach to predicting the compositionality of multiword expressions. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 977–983 (2015)