Sigma J Eng & Nat Sci 34 (1), 2016, 125-133

Sigma Journal Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Research Article / Araştırma Makalesi

AUTOMATIC DISCOVERY OF SIMILAR WORDS BY SUBSTITUTE VECTORS

İnci DÜZENLİ, M. Fatih AMASYALI*

Yildiz Technical University, Computer Engineering Department, Esenler-ISTANBUL

Received/Geliş: 04.11.2013 Revised/Düzeltme: 23.02.2014 Accepted/Kabul: 28.10.2015

ABSTRACT

Patterns between words are generally used for automatic information extraction. However, the patterns can only find related words close to each other. In this study, a method based on substitute vectors can overcome of this difficulty. Firstly, the word sets having the same substitute vector are constructed. Then, similar word sets are obtained according to the number of co-occurring sets. In this sets, semantically relatedness ratio is above 70%. The proposed method is unsupervised. Because, it does not require any seed words manually labeled.
Keywords: Automatic information extraction, discovery of similar words, synonyms, near-synonyms, substitute vectors, natural language processing, artificial intelligence.

EŞ-BAĞLAMLILIK TABANLI BENZER ANLAMLI KELİME BULMA

ÖZ

Otomatik bilgi çıkarımında genelde kelimeler arası şablonlar kullanılır. Ancak, şablonlar birbirine yakın konumlardaki kavramlar arasındaki ilişkileri bulabilir. Bu çalışmada aralarındaki mesafeden bağımsız olarak ilişkili iki kavramı bulabilen eş-bağlamlılık tabanlı bir yaklaşım önerilmiştir. Yaklaşımda önce etrafında aynı kelimeler bulunan kelime kümeleri bulunmuş, ardından kelime ikililerinin bu kümelerde kaç kez beraber geçtiklerine göre de benzer anlamlı kelime kümeleri elde edilmiştir. Sonuçta %70'in üzerinde anlamsal birlik doğruluğuna sahip benzer anlamlı kelime kümelerini bulunmuştur. Yöntem eğiticisiz olup, şablonlar yönteminin gerektirdiği etiketli kavram ikililerine ihtiyaç duymamaktadır.
Anahtar Sözcükler: Otomatik bilgi çıkarımı, benzer kelimelerin otomatik bulunması, eşanlamlı kelimeler, yakın anlamlı kelimeler, bağlam vektörleri, doğal dil işleme, yapay zeka.

1. INTRODUCTION

Automatic knowledge extraction is transforming unstructured texts into the structured knowledge. For example, "location(teacher,school)" and "isa(teacher,staff)" relations can be extracted from the "The teachers and the other staff in the school are invited to the party" sentence.

The relation databases constructed from unstructured texts are using in the several natural

* Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: mfatih@ce.yildiz.edu.tr, tel: (212) 383 57 30

language processing applications such as dialogue systems [1, 2, 3], and customer management systems [4, 5].

Knowledge extraction from unstructured texts became popular, because the most information on the web is in the form of the unstructured text. Almost all data generated by the social network users, news, email bodies, twitter messages are unstructured texts.

The syntactic text patterns are the most popular way of automatic knowledge extraction [6]. Hearst used patterns to find Wordnet relations [7]. Amasyalı used patterns to construct Turkish Commonsense database [8]. Mitchell used patterns with an iterative way to read the entire Web [9]. Chang used html code patterns to extract knowledge from Web pages [10]. Yazıcı used manually constructed patterns to extract relations from the dictionary definitions [11].

The syntactic patterns method requires a set of word pairs which the relation between them is known if the patterns will be automatically constructed. Otherwise, manually constructed patterns should be given the system.

For the some relations, there are syntactic patterns between the words. For example, "isa(X,Y)" relation can be obtained by the "X and the other Y" and "X and similar Y" patterns. However, the patterns only reveal relations between very near words. Moreover, the finding syntactic pattern is not easy for the some relations. The words have the same or near meaning do not generally place in the same sentences. Even if they place, the syntactic patterns between the words would be "-", "," etc. But, these patterns are very general thus they can't be used for the extracting relations.

Alternative approaches are needed to the syntactic patterns for the finding "same or near meaning" relations and finding relations between the words which are not near.

In a text, the words around a word refer to the context of a word and they are named as context vector of a word. A word can be used in several contexts. The "jaguar" word has different context vectors when its sense is an animal or a car. In other words, the sense of a word can only be understood by its context vector. This approach is generally used in the word sense disambiguation problems [12].

K-distanced context vector of a word is consists of the words whose maximum distances to a word is K. A word may have different context vectors. However, different words may have the same context vector. If we search the "w1 w2 * w3 w4" query, the result words (*) have the same context vector. These words are named as substitute vectors. They are used to each other. In this work, we investigate the usage of substitute vectors to find "same or near meaning" relations. The proposed method is unsupervised because it doesn't require the word pairs or seed relations.

In the second section, the proposed method is detailed. The similar works are given at the next section. The experimental results are given at the Section 4. At the last section, the results are discussed and the future studies are given.

2. FINDING WORD CLUSTERS OF SIMILAR MEANING BY THE SUBSTUTION VECTORS

Let the w_i represents the ith word in a text. K-distanced context vector of a w_i is $(w_{i-k},\ldots,w_{i-1}, w_{i+1},\ldots,w_{i+k})$. A substitution vector is consists of words having the same context vector. The proposed method uses these definitions to find word clusters of similar meaning. The method includes 4 steps:

Step 1: Finding K-distanced Context Vectors

Input: A text corpus

Output: N context vectors (CV) having (2**k*) length

For a text corpus having NN words, *NN-(2*k)* context vectors (CV) are found because the context vectors are not found for the first and last k words. CV_i for w_i is given at Equation 1.

$$CV_{i=(k+1):(NN-k+1)} = (w_{i-k},\ldots,w_{i-1}, w_{i+1},\ldots,w_{i+k}) \quad (1)$$

If the corpus consists of independent sentences, CVs of the first and last k words are not found. In this situation, *NN-(MM*2*k)* context vectors are found if the corpus includes MM sentences.

All the CVs have 2*k length. In these CV set, there are identical CVs. After the elimination of same CVs, unique N context vectors are obtained.

Step 2: Finding substitution vectors for each context vector

Input: *N* context vectors (*CV*)

Output: *NM* substitution vectors (*SV*)

For each CV, words (placed in the center of CV) and their frequencies (number of placing in the center of CV) are found. SV_i is the word set found for CV_i.

The frequency of a word w_j for a CV_i *($FR(CV_i,w_j)$)* is found with Equation 2.

$$FR(CV_i,w_j)= FF((w_{i-k},\ldots,w_{i-1}, \mathbf{w_j}, w_{i+1},\ldots,w_{i+k}),C) \quad (2)$$

In Equation 2, C is the text corpus. *FF(X,Y)* is the frequency of X in Y. After the calculating of $FR(CV_i,w_j)$ for each word in the corpus, SV_i is found with Equation 3.

$$EK_i = \bigcup_{w_j \in C \,\&\, FR(BV_i,w_j)\geq 1} w_j \quad (3)$$

Each *SV* may have different number of words. *SV*s having only one word are eliminated because there are useless. After the elimination, SVs having at least two words are obtained. The number of SVs is NM<N.

Step 3: Finding word pairs which are often located in the same substitution vectors

Input: *NM* substitution vectors (*SV*)

Output: *I word pairs (WP) locating with at least G same substitution vectors*

For any two words (w_i, w_j), the number of SVs having these words is calculated by Equation 4.

$$BG(w_i,w_j) = \sum_{k=1:NM} 1(w_i \in SV_k \,\&\, w_j \in SV_k) \quad (4)$$

In Equation 4, *1(x)* function returns 1 if x is true, otherwise it returns 0. The maximum and minimum outputs of *BG(.)* function are NM and 0 respectively. After the finding word pairs and their BGs, WP list is constructed from the word pairs having BG either bigger, or equal to G as Equation 5.

$$WP = \bigcup_{BG(w_i,w_j)\geq G} (w_i, w_j, BG(w_i, w_j)) \quad (5)$$

According to Equation 5, each element of WP list has a word pair and their BG. For the bigger values of G, the words in the pairs have more similar meaning.

The determination of G value is a choice between obtaining a long WP list having less similarity and obtaining a short WP list having more similarity.

Step 4: Finding word clusters of similar meaning

Input: *I word pairs (WP) locating with at least G same substitution vectors*

Output: *F* word clusters of similar meaning (SM)

In this step, word clusters (SMs) are constructed from word pairs (WPs). The construction algorithm is given at Figure 1.

Choose c ($0 < c < 1$)
For each unique word (w_i)
Find all WPs including w_i (WP_{wi})
Sort descending WP_{wi} list according to $BG(w_i,w_j)$ values
Add two words in the first element of WP_{wi} to the cluster
threshold = max(BG)*c
For h=2 : size of WP_{wi}
Calculate mean(BG) of the cluster if we add h^{th} w_j to the cluster
if mean(BG) > threshold
    add w_j to the cluster
else break;

Figure 1. Construction of SMs from WPs

Given below is an example to show how the algorithm works. An example word pair list including "A" word (WP_A) is given at the Table 1. The elements of this list are sorted descending according to their BGs.

Table 1. An example word list including "A" word

word 1	word 2	BG(word 1,word 2)
A	B	40
C	A	30
A	E	20
A	D	5
A	F	3

The elements of word pair having the maximum BG are added to the cluster. In this step, cluster is {A, B}. Than, the words C, E, D, F are controlled respectively if they are added to the cluster or not. If the c value is 0.6, this control continue until the mean(BG) of the cluster is bigger than 24=(40*0.6).

For the word C, if it is added to the cluster, mean(BG) would be (40+30)/2=35 > 24. Than, word C is added. If the word E is added to the cluster, mean(BG) would be (40+30+20)/3=30 > 24. So, the word E is added. But, if the word D is added to the cluster, mean(BG) would be (40+30+20+5)/4=23.75 < 24. So, word D is not added to the cluster and the control is broken. At the end of this process a word cluster is obtained as {A, B, C, E}.

The substitution vectors obtained the second step are also word clusters of similar meaning. But, when we investigate them in the experiments, we saw that they may not have the same meanings. They may have 2 or more meanings. So, we added step 3 and step 4 to get a better word clusters. We can obtain more successful results thanks to these steps. In Section 4, the details of the experiments are given.

3. SIMILAR STUDIES

In literature, there are several works to finding word clusters of similar meaning. Chen clustered columns of the document-term matrix [13]. The document-term matrix contains rows corresponding to the documents and columns corresponding to the terms/words. The words having similar columns may have similar meanings. If the number of document is d, all columns are the vectors having d dimensions. Chen clustered these N dimensional vectors.

According to the Harris, if two words more co-occur in a context (text, sentence); they have more similar meaning [14]. Turney defines Harris's context as a web page [15]. He measures the

meaning similarity as the number of web pages including these two words. Turney applied this approach TOEFL tests and achieved 75% accuracy.

Kleinberg found similar web pages using their links to each other [16]. Senellart uses dictionary definitions to find similar words [17]. Senellart uses dictionary definitions as web pages He uses the words as links to the other definitions/words. He measures similarity with the number of these inlinks and outlinks. The number of common links increases the similarity of the words.

In literature, substitution vectors are also used for several aims. Yüret uses them for word sense disambiguation [12] and unsupervised syntactic categorization of the words [18].

4. EXPERIMENTAL RESULTS

The proposed system requires a big sized corpus because it depends on the frequencies. For this purpose, Turkish news texts in Bilcol corpus [19] are used. About two million sentences were extracted from Bilcol form the system's corpus.

Turkish is an agglutinative language. Morphologic analysis is applied to the corpus. All the words reduced into their stem form. In this way, the number of different words and concept vectors are decreased. Zemberek [20] is used for the morphologic analysis. Zemberek returns all possible morphologic analysis of the words. The first analysis is used. If a word is written wrong or it is not in the Zemberek's dictionary, it is used in the original form. After these preprocessing steps on the corpus, four steps of the system are applied.

Step 1: The K parameter in the finding K-distanced concept vectors affects the number of concept vectors as previously explained. If K is set to 2, total 22 million concept vectors are found from 2 million sentences. The number of unique concept vectors is about 1.8 million.

The number of concept vectors decrease if we set K to 1. In this situation, the number of elements in the substitution vectors in step 2 would increase. But, the similarity between the words in these vectors would decrease. If we use bigger values of K, the number of elements in the substitution vectors would decrease and the similarity would increase. As understood, there is a trade of between the number of elements and the similarity quality in the vectors. In our experiments we tried several k values, and setting k to 2 gives better results.

Step 2: Substitution vectors were obtained for each concept vector out of 1.8 million. The number of substitution vectors having at least 2 words is 130748. The average number of the element is 2.3. Some examples of concept vectors and their substitution vectors are given at Table 2.

Table 2. Examples of concept vectors and their substitution vectors (The translation of concepts vector is not possible every time. So, only two concept vectors and their corresponding substitution vectors are translated into English.)

K-distanced Concept Vectors (K=2)	**Corresponding Substitution Vectors (the words used instead of * in the concept vector)**
meydan gel * kaza 1 (in the * crash)	trafik, tren, maden, son, zincir, feribot (traffic, train, mine, last, chain, ferry)
resmi internet * yap açık (official internet * of)	site, sayfa, portal (site, page, portal)
olay ilgi * sür bil	incele, sor, tahkikat
büyük bir * meydan gel	deprem, patla, sars, hasar, facia, olay
olay yer * polis ekip	gel, git, sevkedilen, bulun, takviye
şube müdür * büro amir	cinayet, gasp, ahlak, infaz, mali, narkotik, organize, yankesici, dolan, internet

Randomly selected 200 out of 130748 substitution vectors were used for measuring similarity quality of the substitution vectors. The frequency of concept vectors generate these selected substitution vectors is bigger than 4. At Table 3, the similarity quality is investigated according to these frequencies. Similarity control was done manually. If all the words in a substitution vector have similar meaning, that substitution vector is count as true.

Table 3. Similarity quality of the substitution vectors

Frequency (F) of concept vector	**Truth ratio of substitution vectors (based on randomly selected vectors)**	**Total number of substitution vectors**
F > 10	0.49	5870
10 ≥ F > 8	0.483	2379
8 ≥ F > 6	0.456	4692
6 ≥ F> 4	0.407	25257

According to the Table 3, there is a direct relation between frequency values and truth ratios. In other words, the more frequent concept vectors generate more successful substitution vectors.

Step 3: 56564 word pairs were obtained from 130748 substitution vectors by Equation 4 and 5. When the word pairs were investigated, it is seen that some pairs included stop words and functional words. “and”, “or”, “one”, “he”, “if” are the samples of these words. The pairs having one of these words were eliminated. After the elimination, 37969 word pairs were obtained.

Measuring similarity quality of word pairs, 200 out of 37969 randomly selected word pairs were used. BG value (Equation 4) of these selected pairs was bigger than 4. In Table 4, the similarity quality is investigated according to BG values. Similarity control was done manually. If the words have similar meaning, that word pair is count as true.

Table 4. Similarity quality of the word pairs

The number of common substitution vectors (BG)	**Truth ratio of word pairs (based on randomly selected pairs)**	**Total number of word pair**	**Some examples (word 1/word1 English-word 2/ word 2 English- BG)**
BG > 10	0.74	1669	(kanun/law-yasa/statute-350) (meclis/council-tbmm/parliament-219) (alan/area-saha/field-177)
10 ≥ BG > 8	0.66	703	(samsunspor / a football team-sivasspor / a football team-10) (can/life-hayat/lifetime-9) (heyelan/landslide-sel/flood-9)
8 ≥ BG > 6	0.67	1390	(eroin/heroin-kokain/cocaine-8) (grekoromen/Greco-Roman-serbest/freestyle-8) (grup/group company-holding/holding company-7)
6 ≥ BG > 4	0.64	3139	(Hollanda/Netherlands-KKTC/TRNC-6) (para/money-vergi/tax-6) (beklenti/expectation-umut/hope-5)

It can be seen easily that there is a direct relation between the BG values and the truth ratios. In other words, having more common substitution vectors yields more similar word pairs. Moreover, the truth ratios are bigger than the Table 3.

Step 4: 2406 word clusters of similar meaning were obtained from 37969 word pairs using the algorithm in Figure 1. The c value in the algorithm is set to 0.6.

Measuring similarity quality of word clusters 200 out of 2406 randomly selected word clusters were used. The number of words in these clusters is between 3 and 10. In Table 5, the similarity quality is investigated according to number of words in the clusters. To get more confidential truth ratios, similarity control were done by different 3 referees. If all the words in a word cluster have similar meaning, that word cluster is count as true. The majority vote of 3 referees is the final decision. The consistency ratio of the referee decisions was also reported.

Table 5. Similarity quality of the word clusters

Number of words in the clusters (KS)	**Average Truth ratio of word clusters**	**Total number of word clusters**	**Consistency ratio between referees**	**Some word clusters**
KS = 3	0.76	366	0.9	{asker,jandarma,polis} {soldier, gendarme, police}
KS = 4	0.67	257	0.85	{ankara,hatay,istanbul, izmir} {some cities in Turkey]
KS = 5	0.56	181	0.76	{cami,camii,lokanta, restaurant,restoran} {mosque, diner, restaurant}
$10 \geq KS > 5$	0.55	512	0.67	{denizlispor,galatasaray, konyaspor,malatyaspor, sakaryaspor,samsunspor, trabzonspor} {some football clubs in Turkey}

In Table 5, it can be seen easily that the truth ratios decrease with the number of the words in the clusters. In other words, the smaller clusters have more similar words in it. Moreover, the truth ratios are bigger than Table 3 and 4.

As expected, the diversity of referee decisions increase with the number of the words in the clusters.

When the Table 3, 4, and 5 are investigated together, the applied processes (step 3 and step 4) increased the truth ratios but the number of extracted information. In other words, the more refined clusters yield the less number of clusters.

5. RESULTS AND FUTURE STUDIES

In this study, a new algorithm is proposed to find word clusters of similar meaning. The algorithm uses substitution vectors and refines them according to the co-occurrence of the words. In the experiments, we showed the higher performance of the system. The findings obtained in our experiments are listed below:

- The system can find the relations between near and far words, whereas the pattern matching method can only find relations between near words.
- The system is unsupervised because it doesn't require neither patterns nor seed word lists.
- Whereas, the experiments are based on a Turkish corpus the system is language independent.

It doesn't use any specific structure of Turkish language. Moreover, if the corpus of language is not agglutinative, the concept vectors can be obtained more accurately. So, all the system can work better.

- The system is based on the frequencies. If we have bigger sized corpus, the number of obtained word clusters would be more and the words in the clusters would be more similar.
- The usage of only substitution vectors does not provide sufficient semantic consistency of word clusters. The system also uses co-occurrence of the words in the clusters (step 3 and 4). In this way, the system obtains word clusters of more similar meaning than the substitution vectors.
- The system is implemented for single word concepts. But it can be expanded to the multi-word concepts.
- The system requires a big sized corpus. The complexity of the used algorithm is also high. So, its processing time is long.
- In the system, detecting words having similar meaning is required having the same concept vectors. Moreover, the probability of having same concept vectors is very low. Only about 2406 word clusters can be obtained with 2 million sentences. The effectiveness of the system can be seen as low. In the corpus, there would be more word clusters can not be found by the system. If the system is used with other pattern based systems, the results would be more sufficient. Whereas the pattern base system detects the similarity between near words, the proposed system detects similarity between far words.

Usage of other corpus on the other languages, reducing complexity of the algorithm, using also for multi-word concepts are planned as the future works.

REFERENCES / KAYNAKLAR

[1] Tarau, Paul, ve Elizabeth Figa. "Knowledge-based conversational agents and virtual storytelling", Proceedings of the 2004 ACM symposium on Applied computing, ACM, 2004

[2] Kathy Panton, Cynthia Matuszek , Douglas Lenat , Dave Schneider, Michael Witbrock, Nick Siegel, Blake Shepard, "Common sense reasoning–from Cyc to intelligent assistant", Ambient Intelligence in Everyday Life. Springer Berlin Heidelberg, 2006.

[3] Al-Zubaide, Hadeel, ve Ayman A. Issa. "Ontbot: Ontology based chatbot", Innovation in Information & Communication Technology (ISIICT), 2011 Fourth International Symposium on. IEEE, 2011.

[4] Cambria, Erik, Catherine Havasi, ve Amir Hussain. "SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis", FLAIRS Conference. 2012.

[5] Cambria, E., Song, Y., Wang, H., ve Howard, N., "Semantic multi-dimensional scaling for open-domain sentiment analysis", In IEEE Intelligent Systems, DOI: 10.1109/MIS.2012.118, 2013.

[6] Chia-Hui Chang, Kayed Mohammed, Girgis, M.R., Shaalan, K.F., "A Survey of Web Information Extraction Systems",Knowledge and Data Engineering, IEEE Transactions on, Vol:18(10), pp.1411 - 1428, 2006.

[7] Hearst, M., "Automated Discovery of WordNet Relations in WordNet: An Electronic Lexical Database", Christiane Fellbaum (ed.), MIT Press, 1998.

[8] M.Fatih Amasyalı, "Automatic Construction of Turkish Wordnet", SIU 2005.

[9] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., Tom M. Mitchell, "Toward an Architecture for Never-Ending Language Learning", AAAI Publications, Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.

[10] Chia-Hui Chang, Chun-Nan Hsu, Shao-Chen Lui, "Automatic Information Extraction from Semi-Structured Web Pages By Pattern Discovery", Decision Support Systems J., vol. 35, no. 1, pp. 129-147, 2003.

[11] Emre Yazıcı, M.Fatih Amasyalı, "Automatic Extraction of Semantic Relationships Using Turkish Dictionary Definitions", EMO Bilimsel Dergi, Vol. 1, No. 1, pp. 1-13, 2011.

[12] Deniz Yüret, "Word sense disambiguation by substitution", SemEval-2007, pages 207-214, Prague, Czech Republic.
[13] H. Chen, K.J. Lynch, "Automatic construction of networks of concepts characterizing document databases", IEEE Transactions on Systems, Man and Cybernetics, 22(5):885–902, 1992.
[14] Z.S., Haris, "Mathematical structures of language", Wiley, s.12, 1968.
[15] P. D. Turney, "Mining the Web for synonyms: PMI-IR versus LSA on TOEFL", In Proceedings of the European Conference on Machine Learning, p. 491–502, 2001.
[16] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 46(5):604–632, 1999.
[17] Pierre P. Senellart, Vincent D. Blondel, "Automatic Discovery of Similar Words", Survey of Text Mining Clustering, Classification, and Retrieval Berry, Michael W. (Ed.), pp.2-44, 2004.
[18] Yatbaz, A. Y., Sert E., Yuret D., "Learning Sysntactic Categories Using Paradigmatic Representations of Word Context", EMNLP-CoNLL 2012, Jeju Island, Korea.
[19] Can, F., Koçberber, S., Bağlıoğlu, O., Kardaş, S., Öcalan, H.C., Uyar, E., "Türkçe haberlerde yeni olay bulma ve izleme: Bir deney derleminin oluşturulması", Akademik Bilişim Sempozyumu, 2009.
[20] Ahmet Afşin Akın, Mehmet Dündar Akın, Zemberek, an open source NLP framework for Turkic Languages, Yayınlanmamış çalışma, 2007, http://zemberek.googlecode.com/files/zemberek_makale.pdf