Mahtab at SemEval-2017 Task 2: Combination of Corpus-based and Knowledge-based Methods to Measure Semantic Word Similarity

Niloofar Ranjbar
ni.ranjbar@Mail.sbu.ac.ir
Fatemeh Mashhadirajab
fa.mashhadi@gmail.com
Mehrnoush Shamsfard
m-shams@sbu.ac.ir
Rayehe Hosseini pour
Hosseinipour.r1@gmail.com
Aryan Vahid pour
aryanvahidpour@gmail.com
Faculty of Computer Science and Engineering / Shahid Beheshti University

Abstract

In this paper, we describe our proposed method for measuring semantic similarity for a given pair of words at SemEval-2017 monolingual semantic word similarity task. We use a combination of knowledge-based and corpus-based techniques. We use FarsNet, the Persian WordNet, besides deep learning techniques to extract the similarity of words. We evaluated our proposed approach on Persian (Farsi) test data at SemEval-2017. It outperformed the other participants and ranked the first in the challenge.

1 Introduction

Semantic similarity represents a special case of semantic relatedness: for example, cars and gasoline would seem to be more closely related than, say, cars and bicycles, but the latter pair are certainly more similar(Resnik et al., 1999). Semantic similarity has been used in many application in natural language processing. At SemEval-2017 monolingual semantic word similarity task, given a pair of words, we have to automatically measure their semantic similarity and score them according to a [0-4] similarity scale where 4 denotes that the two words are synonymous and 0 indicates that they are completely dissimilar(Camacho-Collados et al., 2017). In subtask 1 in which we participated, the two words in the pair belong to the same language. This subtask provides five monolingual word similarity datasets in English, German, Italian, Spanish and Farsi. The language whose dataset we used is Farsi.

The rest of the paper is organized as follows. Section 2 reviews published work related to the semantic word similarity task. Section 3 explains the proposed algorithm. The experimental results are discussed in Section 4 and the conclusion and future work are reported in Section 5.

2 Related Works

There are different methods for finding semantic similarity and relation between two words. Christoph(Christoph, 2015)generally divides similarity measurement techniques into two categories: knowledge-based and corpus-based techniques. Some of the available techniques use a combination of these two methods. In knowledge-based techniques, a taxonomy or ontology like WordNet(Miller, 1995) usually is used to extract taxonomic information like path length and depth in the hierarchy(Pilehvar and Navigli, 2015). For example, the proposed methods by Resnik(Resnik, 1995), Lin(Lin et al., 1998)and FaITH(Pirró and Euzenat, 2010) fall in this category. In corpus-based techniques, usually a large corpus is used to extract statistical information. Christoph(Christoph, 2015) also divides corpus-based methods into two categories. One category contains simple distributional approaches, which check co-occurrences of words like SemSim(Bollegala et al., 2007) and PMI(Church and Hanks, 1990) . Another category contains dense vector representations-based methods, which usually use dimensionally reduction techniques in vector representations. LSA(tefănescu et al., 2014) , SGNS(Mikolov et al., 2013), SVD(Levy and Goldberg, 2014)and GLOVE(Pennington et al., 2014) are some methods which fall in this category.

In the proposed method, we use a combination of both knowledge-based and corpus-based

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 256–260,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

techniques. On the one side, we have used FarsNet(Shamsfard et al., 2010a) ontology to enable knowledge-based techniques and on the other side we have used corpus-based techniques like Word2Vec(Mikolov et al., 2013) in order to improve results.

3 The Proposed Method

One of the corpus-based methods is continuous vector representation, also known as word embedding. For using this method, first we preprocess the Wikipedia corpus downloaded from Polyglot (Al-Rfou et al., 2013) and then we use Word2Vec toolkit, which is represented by deeplearnig4j library(Team, 2016) in java. Furthermore, we use some Lexical Resources such as FarsNet(Shamsfard et al., 2010a) (the Persian WordNet) and BabelNet(Navigli and Ponzetto, 2012) in order to measure similarity between pair of words. We explain this method in detail in sections 3.1 and 3.2.

3.1 Corpus-based Method

Preprocessing:Preprocessing includes four steps: stop-words removal, removing punctuations and numbers, stemming and normalizing multi word expressions by replacing all space characters with ”zero-width non-joining” character.

First, we remove all Persian stop-words according to ”ranks” website[1] , and then we remove punctuation marks and all English and Persian numbers in the text. After that, we replace plural words with their singular form and remove inflectional suffixes using STeP-1(Shamsfard et al., 2010b). Finally, we detect multi-word expressions, which appeared in corpus and contain only two words, by checking all bi-grams of corpus in FarsVaje Lexicon[2] and normalize them by replacing all space characters with ”zero-width non-joining” character. We also replace multi-word expressions, which appeared in test dataset, with their normalized form.

Word2Vec: ”Word2Vec is a two-layer neural network that processes text. Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. Deeplearning4j(Team, 2016) implements a distributed form of Word2vec for Java”[3].

To measure the similarity of two words first we measure the similarity of their corresponding synonyms. If there are n synonyms for the first word and m synonyms for the second, we will calculate the similarity for n*m pairs which are made of the Cartesian product of two synonym sets extracted from FarsNet (Word2Vec gives us cosine similarity and we consider it as similarity between a pair of words). At last, we choose maximum value of similarities as score of a pair of words. We use default tokenizer factory for tokenizing the corpus, which tokenizes the text by spaces and is useful for our purpose. We set Word2Vec parameters as below: minWord-Frequency= 1, iterations=1, layer Size=100, seed=42 and windowSize=5.

3.2 Knowledge-based Methods

FarsNet: FarsNet is a Lexical ontology for Persian language. This ontology is designed to contain a Persian WordNet with about 42000 synset in the last version, which we use to measure similarities. In this section, we will explain the approach we use to measure semantic distance between two words using some rules and then introduce a function, which map the measured distance to similarity score. First, we explain some strict rules introduced by(Rychalska et al., 2016) and then we explain some less strict rules:

1. If two words are exactly the same or are two different writing forms of one word or belong to the same synset, the distance will be zero (D(x,y)=0).
2. If two words have more than four common senses in their corresponding synsets, the distance will be one (D(x, y) =1).
3. If there is a direct or two-level hypernym relation between the corresponding synsets of words, the distance will be two (D(x, y) =2).
4. If two words share any common sense, the distance will be three (D(x, y) =3).
5. If two words are derivationally related, the distance will be four (D(x, y) =4).

If none of these rules met, we use the following rules, which are less strict:

1. If there is any relation except hypernym between synsets of two words, the distance will be three (D(x, y) =3).

[1] http://www.ranks.nl/stopwords/persian

[2] A lexicon developed at NLP lab of Shahid beheshti University and is used in Negar (Shamsfard et al., 2016)

[3] https://deeplearning4j.org/word2vec

2. If there is any two-links relation except hypernym between synsets of two words, the distance will be four (D(x, y) =4).

3. If there is any three-links relation between synsets of two words, the distance will be five (D(x, y) =5).

After all, if no relation is found between a pair of word to measure the distance between them, the distance will set to -1 and then we calculate similarity score using equation 1 introduced by(Rychalska et al., 2016):

$$A(x.y) = \begin{cases} \beta e^{-\alpha D(x,y)} & if D(x,y) \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We set α to 0.25 and β to 1 as these values seemed to yield the best results.

BabelNet: BabelNet is a very large, wide coverage multilingual semantic network. We use the version of Babel Net which was available on September 2016.

Semantic distance D is measured using the following rules for these pairs of words:

1. If words are exactly the same or one of them is main sense for another, the distance will be zero (D(x, y) =0).

2. If there is a direct named-relation between pairs, the distance will be one (Un-named relations are filtered e.g. semantically related) (D(x, y) =1).

3. If words share more than four common sense the distance will be two (D(x, y) =2).

4. If words share any common sense, the distance will be three (D(x, y) =3).

5. If their synsets share any important domain, the distance will be four (domains like media which are too general to be considered as a similarity measure are filtered) (D(x, y) =4).

6. If the main gloss of one of the words contains the other one the distance will be five (D(x, y) =5).

7. If there is a 2-link (indirect) n amed-relation between them, the distance will be six (D(x, y) =6).

If none of these rules met, D will set to -1 then we calculate similarity score using equation 1.

Gloss: We also use gloss of words extracted from FarsNet and BabelNet to measure similarity of a pair of words. A combination of the following methods is used (note that in all methods finally we calculate sum of intersections)

1. Gloss-Gloss: In this method, the intersection between glosses of both words is calculated.

2. Hyper-Hyper: In this method, the intersection between glosses of Hypernyms of both words is calculated.

3. Hypo-Hypo: In this method, the intersection between glosses of Hyponyms of both words is calculated.

4. Gloss-Hyper: In this method, we calculate the intersection between glosses of Hypernyms of the first word and glosses of the second word and vice versa (the intersection between glosses of Hypernyms of the second word and glosses of the first word) and finally we calculate sum of both intersections.

In order to calculate intersections, we use following method:

First, we remove stop-words from sentences and extract words, after that we choose longest common subsequence in each iteration and calculate square of its length as its score in that iteration. For example, suppose that we have two following sequences:

1,2,3,4,5 and 1,2,7,4.

In the first iteration 1, 2 is LCS and its length is 2 so the score in this iteration will be 4.

In the second iteration, we have two following sequences: 3,4,5 and 7,4 and 4 is LCS of these sequences with length of 1 so the final score will be 1+4=5.

After measuring all scores, we normalize all of them between 0 and 1 using following equations:

$$mean = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (2)$$

$$variance = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - mean)^2 \quad (3)$$

$$y = \frac{x_i - mean}{2 * \sqrt{variance}} \quad (4)$$

$$x = \frac{1}{1 + e^{-y}} \quad (5)$$

3.3 Combination of methods

To obtain the final similarity score for a pair of words, we calculate their normalized weighted sum using equations 6 and 7 :

$$score_{knowledge-based} = score_{Farsnet} + 0.3 * score_{Babelnet} + 0.15 * score_{Gloss} \tag{6}$$

$$FinalScore = \frac{score_{knowledge-based} + 0.75 * score_{W2V}}{2.2} \tag{7}$$

4 Experimental Results

We evaluated our proposed approach at SemEval 2017 and ranked first among the participants in task2, subtask1 for Farsi.Table 1 shows the results of the submitted runs on test data from SemEval 2017 task2-subtask1 for Farsi. Test data at SemEval 2017 task2 is available at :

alt.qcri.org/semeval2017/task2/data/uploads/semeval17_task2_test.zip

The test data for Farsi language contains 500 pairs of Farsi words that we have to measure their semantic similarity.

Farsi Test Data	Pearson	Spearman	Final
Mahtab	0.719	0.711	0.715
hhu_run2	0.606	0.601	0.604
hhu_run1	0.541	0.585	0.562
Luminoso_run2	0.507	0.498	0.503
Luminoso_run1	0.506	0.496	0.501
HCCL_run1	0.424	0.45	0.436
NASARI(baseline)	0.412	0.398	0.405
SEW_run1	0.383	0.404	0.393
RUFINO_run1	0.378	0.344	0.36
RUFINO_run2	0.25	0.262	0.256
hjpwhuer_run1	0.002	-0.003	0.0

Table 1: The results of the submitted runs on Farsi test data at SemEval 2017 task2_subtask1.

5 Conclusions and Future Work

This paper described our proposed method, a combination of corpus-based techniques like Word2Vec and knowledge-based techniques using FarsNet to measure semantic similarity between given pairs of words. The results show that our method achieved good results, better than other participants in the challenge. Future work will focus on enhancing the similarity measures besides using other corpus-based techniques like GloVe and LSA.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662* .

Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Measuring semantic similarity between words using web search engines. *www* 7:757–766.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proceedings of SemEval*. Vancouver, Canada.

LOFI Christoph. 2015. Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. *Information and Media Technologies* 10(3):493–501.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1):22–29.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*. pages 2177–2185.

Dekang Lin et al. 1998. An information-theoretic definition of similarity. In *ICML*. Citeseer, volume 98, pages 296–304.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Mohammad Taher Pilehvar and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence* 228:95–128.

Giuseppe Pirró and Jérôme Euzenat. 2010. A feature and information theoretic framework for semantic similarity and relatedness. In *International semantic web conference*. Springer, pages 615–630.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007* .

Philip Resnik et al. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)* 11:95–130.

Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung poland nlp team at semeval-2016 task 1: Necessity for diversity; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*.

M Shamsfard, H Fadaiei, and H.S Jafari. 2016. Negar: Persian Standardization System. Technical report, Shahid Beheshti University ,NLP Research Lab.

Mehrnoush Shamsfard, Akbar Hesabi, Hakimeh Fadaei, Niloofar Mansoory, Ali Famian, Somayeh Bagherbeigi, Elham Fekri, Maliheh Monshizadeh, and S Mostafa Assi. 2010a. Semi automatic development of farsnet; the persian wordnet. In *Proceedings of 5th Global WordNet Conference*. volume 29.

Mehrnoush Shamsfard, Hoda Sadat Jafari, and Mahdi Ilbeygi. 2010b. Step-1: A set of fundamental tools for persian text processing. In *LREC*.

Deeplearning4j Development Team. 2016. Deeplearning4j: Open-source distributed deep learning for the jvm. *Apache Software Foundation License* 2.0. http://deeplearning4j.org.

Dan tefănescu, Rajendra Banjade, and Vasile Rus. 2014. Latent semantic analysis models on wikipedia and tasa. In *Language Resources Evaluation Conference (LREC)*.