

Dual Embeddings and Metrics for Relational Similarity

Dandan Li

U.S. Army Research Laboratory
happydandan2016@gmail.com

Douglas Summers-Stay

U.S. Army Research Laboratory
douglas.a.summers-stay.civ@mail.mil

Abstract

Abstract. In this work, we study the problem of relational similarity by combining different word embeddings learned from different types of contexts. The word2vec model with linear bag-of-words contexts can capture more topical and less functional similarity, while the dependency-based word embeddings with syntactic contexts can capture more functional and less topical similarity. We explore topical space and functional space simultaneously by considering these two word embeddings and different metrics. We evaluate our model on relational similarity framework, and report state-of-the-art performance on standard test collections.

1 Introduction

Measuring relational similarity between two word pairs plays important roles in natural language processing (NLP). The techniques for solving this problem can be applied to a variety of NLP tasks, such as query expansion, word sense disambiguation, machine translation, information extraction and question answering. Previous work addressing the problem can be roughly classified into three categories: (1) learning word embeddings from large collections of text using variants of neural networks (Mikolov et al. (2013a); Mikolov et al. (2013b); Mikolov et al. (2013c); Levy and Goldberg (2014)) or global matrix factorization (Deerwester et al. (1990); Turney (2012)); (2) extracting knowledge from existing semantic networks, such as WordNet (Yang and Powers (2005); Alvarez and Lim (2007); Hughes and Ramage (2007)) and ConceptNet (Boteanu and Chernova (2015)); (3) combining the above two models by various ways (Agirre et al. (2009); Zhila et al. (2013); Iacobacci et al. (2015); Summers-Stay et al. (2016)).

The empirical evidence shows that the word representations learned from neural network models do an especially good job in capturing not only attributional similarities between words but also similarities between pairs of words (Mikolov et al. (2013c)). Levy and Goldberg (2014) generalize the skip-gram model with negative sampling to include arbitrary word contexts and present the dependency-based word embeddings, which are learned from syntactic contexts derived from dependency parse-trees. Qualitative and quantitative analysis demonstrates that the word2vec model with linear bag-of-words contexts can yield broad topical similarity while the dependency-based word embeddings with syntactic contexts can capture more functional similarity. Turney (2012) is the first, to the best of our knowledge, to raise the word vector representations in a dual space and unify semantic relations and compositions by the dual-space model. The dual-space model consists of a domain space and a function space, where the domain or topic of a word is characterized by the nouns that occur near it and the function or role of a word is characterized by the syntactic context that relates it to the verbs that occur near it.

We detail our main contributions as follows. (1) In this paper, we use the word2vec model with linear bag-of-words contexts to capture the domain of a word, and the dependency-based word embeddings with syntactic contexts to characterize the function of a word. The broad contexts used in our model can provide richer information for measuring domain similarity (i.e., topic, subject, or field similarity)

and function similarity (i.e. role, relationship, or usage similarity) than noun or verb-based patterns for contexts in Turney’s (2012) model. (2) The two existing models for measuring relational similarity are: the directional similarity model (Zhila et al. (2013)) and the dual-space model consisting of domain and function space (Turney (2012)). Both models suffer some drawbacks. The directional similarity model explores the difference of two relationships in multiple topicality dimensions in the vector space. However, it ignores the spatial distances between word vectors, which can reveal the function similarity of words in function space. The dual-space model can measure the domain similarity and function similarity between words. However, it only computes the domain similarity between two single words and places less emphasis on the domain similarity between two relations. In this work, we propose a new dual-space model for measuring relational similarity, which combines the advantages of the two existing models. (3) We evaluate our model on relational similarity framework and report state-of-the-art performance on SAT analogy questions.

2 Related Work

Vector space models have a long, rich history in the field of natural language processing, where each word is represented as a real-valued vector in a continuous vector space. All these models depend in some way or another on the distributional hypothesis which states that words that occur in similar contexts tend to have similar meanings (Harris, 1954; Firth, 1957). There are the two main model families for learning word vectors: (1) global matrix factorization methods, such as latent semantic analysis, which generates embeddings from term-document matrices by singular value decomposition (Deerwester et al. (1990)); (2) neural network models, such as the skip-gram and continuous bag of words models of (Mikolov et al. (2013a); Mikolov et al. (2013b); Mikolov et al. (2013c)), referred to as word2vec, which learn embeddings by training a network to predict neighboring words. Levy and Goldberg (2014) generalize the skip-gram model with negative sampling to include arbitrary word contexts, learn embeddings from syntactic contexts and demonstrate that the dependency-based embeddings can capture more functional similarity than the original skip-gram embeddings (word2vec).

Several algorithms have been proposed for solving SAT-style analogy questions. A good algorithm for recognizing analogies can also help to solve the classification problem of semantic relations, which has potential applications in machine translation, information extraction and word sense disambiguation. Quesada et al. (2004) propose LSA as a method to represent the relations between words and use a prediction to represent relation comparisons in the LSA semantic space. Veale (2004) considers the utility of the taxonomic structure of WordNet to the solution of SAT analogies. Turney and Littman (2005) show that the cosine metric in the Vector Space Model of information retrieval can be used to solve analogy questions and to classify semantic relations. Turney (2012) introduces a dual-space model, which consists of a space for measuring domain similarity and a space for measuring function similarity. The dual-space model has been applied to measuring relation similarity and compositional similarity.

3 Word Embeddings with Different Context Types

Word2vec (W2V) is one of the most popular word embedding methods that learn word vectors from raw text. It has two models for generating dense embeddings: the skip-gram model and the continuous bag-of-words (CBOW) model. The training objective of the skip-gram model is to predict each surrounding word in a context window of $2k$ words from the target word. So for $k=2$, the contexts of the target word w_t are $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ and we are predicting each of these from the word w_t . However, a context window with a smaller size k may miss some important contexts while including some accidental ones. Recently, Levy et al. propose the dependency-based word embeddings (DEP), which generalize the skip-gram model with negative sampling, and move from linear bag-of-words contexts to syntactic contexts that are derived from automatically produced dependency parse-trees. Embeddings produced from different kinds of contexts can induce different word similarities. The original skip-gram embed-

dings can yield broad topical similarities, while the dependency-based word embeddings can capture more functional similarities (Levy et al., 2014).

4 Relational Similarity

Relational similarity measures the degree of correspondence between two relations (Jurgens et al. (2012)). The task can be modeled as an analogy problem, where given two pairs of words, $A:B$ and $C:D$, the goal is to determine the degree to which the semantic relation between A and B is similar to that between C and D . We first introduce two types of existing models for relational similarity.

4.1 Directional Similarity Model

Zhila et al. (2013) propose a directional similarity model to evaluate the correspondence between relations. Given two pairs of words $A:B$ and $C:D$, suppose (v_A, v_B) and (v_C, v_D) are the corresponding vectors of these words. Relational similarity of these two word pairs is defined as the cosine function between the two directional vectors of $(v_A - v_B)$ and $(v_C - v_D)$:

$$\text{Similarity}(A:B :: C:D) = \frac{(v_A - v_B) \cdot (v_C - v_D)}{\|v_A - v_B\| \|v_C - v_D\|}.$$

In this model, the relationship between two words is represented by the difference of corresponding two word vectors, which reveals the change from one word to the other in terms of multiple topicality dimensions in the vector space. From the geometric point of view, if two relationship vectors are relatively parallel (i.e., share the same direction), then the two word pairs can be considered as they have similar relations. They apply the model to the problem of relation classification which determines whether two pairs share the same relation. The goal here is to design a model for answering the more difficult SAT analogy questions. However, sometimes this directional similarity model can not measure the analogy problem correctly.

4.2 A Dual-Space Model

Why does the directional similarity model fail on some analogy questions? One possible reason is that the model depends on direction alone and ignores spatial distance between word vectors. Turney (2012) presents a dual-space model to unify semantic relations and compositions. The dual-space model consists of a domain space for measuring domain similarity and a function space for measuring function similarity. He gives a good example to explain this model. Consider the analogy example, *traffic:street :: water:riverbed*. *Traffic* and *street* share the same domain, the domain of *transportation*. *Water* and *riverbed* share the same domain, the domain of *hydrology*. On the other hand, *traffic* and *water* share the same function, the function of *flow*, and *street* and *riverbed* share the same function, the function of *carry*. The model can recognize that the semantic relation between *traffic* and *street* is analogous to the relation between *water* and *riverbed* by considering the combination of domain and function similarity. A general explanation for the model is, given an analogy $A:B :: C:D$, A and B , C and D have relatively high domain similarity in their respective domains; A and C , B and D have relatively high function similarity in respective function spaces.

4.3 A New Dual-Space Model: Refining the Directional Similarity Model and the Dual-Space Model

The directional similarity model (Zhila et al. (2013)) explores the difference of two relationships in multiple topicality dimensions in the vector space. However, it ignores the spatial distance between word vectors, which can reveal the function similarity of words in function space. The dual-space model (Turney (2012)) can measure both domain similarity and function similarity between words. However, it only computes the domain similarity between two single word vectors and places less emphasis on

the domain similarity between two relations. Moreover, Turney’s (2012) model uses only noun or verb-based patterns for contexts to model domain or function space. In this paper, we propose a novel dual-space model, which combines the advantages of the above two existing models. We use the word2vec-based directional similarity model to measure the similarity of two relations in domain space, and the dependency-based word embeddings to represent the word vector in function space. The two embeddings used in our model consider broader contexts than those in the original dual-space model, which can provide richer information for measuring domain similarity and function similarity.

A mathematical description of our dual-space model is given as follows:

Given two pairs of words A:B and C:D, suppose $V_{W2V(A)}$, $V_{W2V(B)}$, $V_{W2V(C)}$, $V_{W2V(D)}$ are the vectors of these words in domain space and $V_{DEP(A)}$, $V_{DEP(B)}$, $V_{DEP(C)}$, $V_{DEP(D)}$ are the vectors of these words in function space. Based on the directional similarity model (Zhila et al., 2013), we define the domain similarity of two pairs of words as follows:

$$\text{Similarity}_D(A: B :: C: D) = \frac{(V_{W2V(A)} - V_{W2V(B)}) \cdot (V_{W2V(C)} - V_{W2V(D)})}{\|V_{W2V(A)} - V_{W2V(B)}\| \|V_{W2V(C)} - V_{W2V(D)}\|}.$$

The function similarity of A and C and the function similarity of B and D are defined respectively as follows:

$$\text{Similarity}_{F1}(A: C) = \frac{V_{DEP(A)} \cdot V_{DEP(C)}}{\|V_{DEP(A)}\| \|V_{DEP(C)}\|};$$

$$\text{Similarity}_{F2}(B: D) = \frac{V_{DEP(B)} \cdot V_{DEP(D)}}{\|V_{DEP(B)}\| \|V_{DEP(D)}\|}.$$

We design the method Similarity_{ADD} to combine the above similarities for relational similarity as follows, which satisfies that the combined similarity is high when the component similarities are high:

$$\text{Similarity}_{ADD}(A: B :: C: D) = \frac{1}{3}(\text{Similarity}_D(A: B :: C: D) + \text{Similarity}_{F1}(A: C) + \text{Similarity}_{F2}(B: D))$$

We give an SAT question in Table 1 to demonstrate how our compositional similarity works. An SAT analogy question consists of a target pair of words and five option pairs of words. The task is to select the option pair that “best expresses a relationship similar to that expressed in the original pair”, as stated in the test’s directions. We use the 300-dimensional W2V and DEP vectors available for downloading ¹. For the target pair *bruise:skin* in Table 1, our method Similarity_{ADD} (= 0.406) can recognize the correct answer *stain:fabric* although Similarity_D (= 0.189), Similarity_{F1} (= 0.539) and Similarity_{F2} (= 0.491) for the correct answer are all ranked second among five options.

bruise:skin	Similarity_D	Similarity_{F1}	Similarity_{F2}	Similarity_{ADD}
muscle:bone	0.023	0.457	0.622	0.367
smudge:blemish	-0.135	0.540	0.320	0.242
rash:allergy	0.208	0.358	0.370	0.312
layer:veneer	-0.091	0.273	0.468	0.216
stain:fabric	0.189	0.539	0.491	0.406

Table 1: Experimental results of four models on SAT question *bruise:skin*

5 Experiments and Evaluation

5.1 Relational Similarity Experiment

In the following experiments, we evaluate our approaches to solving analogies by a set of 374 SAT analogy questions, which is the same set of questions as was used in Turney’s Dual-Space mode (Turney

¹<http://u.cs.biu.ac.il/~nlp/resources/downloads/embeddings-contexts/>

(2012)). Precision and Recall are two standard performance measurements used for evaluation. The definitions of precision and recall are specified by (Turney and Littman (2005)). In all experiments, we use the 300-dimensional W2V and DEP vectors pretrained on a concatenation of three large, diverse English corpora, and those vectors are available for downloading ¹.

Table 2 shows the experimental results of four approaches presented in Section 4.3 on the set of 374 analogy questions. Two questions are skipped because the vector for the target pair is not available in the collection. Since there are five options for each target pair of an SAT analogy question, random guessing would yield a recall of 20%. Domain similarity Similarity_D , function similarity Similarity_{F1} and Similarity_{F2} all perform much better than random guessing. Our compositional similarity model Similarity_{ADD} in the dual space clearly outperforms Similarity_D , Similarity_{F1} and Similarity_{F2} in single domain or function space.

	Similarity_D '13	Similarity_{F1}	Similarity_{F2}	Similarity_{ADD}
Correct	170	139	144	185
Incorrect	202	233	228	187
Precision%	45.7	37.4	38.7	49.7
Recall%	45.5	37.2	38.5	49.5

Table 2: Experimental results on the set of 374 analogy questions

Table 3 splits out the results for different parts of speech of the 374 SAT questions, which include noun:noun, adjective:adjective, verb:verb, noun:adjective or adjective:noun, noun:verb or verb:noun, verb:adjective/adverb or adjective/adverb:verb. We compare the number of correct guesses of our model Similarity_{ADD} with two existing models Similarity_D (Zhila et al. (2013)) and Dual-Space model (Turney (2012))². Similarity_{ADD} works best on the questions with the labels noun:adjective and adjective:adjective. For the questions with the label verb:verb, Similarity_D achieves the best performance. Dual-Space'12 outperforms other two models on the noun:noun, noun:verb and verb:adj/adv questions.

Parts of speech	Total	Similarity_{ADD}	Similarity_D '13	Dual-Space'12
noun:noun	192	95	93	97
noun:adj or adj:noun	66	38	28	35
noun:verb or verb:noun	54	23	20	27
adj:adj	24	10	8	9
verb:adj/adv or adj/adv:verb	21	10	9	12
verb:verb	17	9	12	11

Table 3: Experimental results on the 374 SAT questions labeled with different parts of speech

6 Conclusion and Future Work

In this work, we explore domain space and function space simultaneously by considering two kinds of word embeddings and different metrics. Word embeddings can capture topical and functional information of a word by using different types of contexts, however they are unable to model the words with multiple meanings accurately because a word is represented as just a single vector which carries a weighted average of different meanings. Existing lexical and knowledge databases, such as WordNet, ConceptNet and Cyc, can be modeled as graphs in which words are represented as the nodes and the relations between words are signified by the edges. These databases have more accurate information although their coverage of words and types of relations is usually limited. We plan to look at ways to combine our dual-space model with existing databases to improve the performance of our current system.

²In the paper (Turney (2012)), the number of questions labeled by noun:noun is 191, and the number of questions labeled by noun:verb or verb:noun is 55.

References

- Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado.
- Alvarez, M. A. and Lim, S. J. (2007). A graph modeling of semantic similarity between words. In *Proceedings of the Conference on Semantic Computing*, pages 355–362.
- Boteanu, A. and Chernova, S. (2015). Solving and explaining analogy questions using semantic networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1460–1466.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantics analysis. *Journal of the Association for Information Science and Technology*, 41(6):391–407.
- Hughes, T. and Ramage, D. (2007). Lexical semantic relatedness with random graph walks. In *Proceedings of EMNLP-CoNLL-2007*, pages 581–589.
- Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). Sensembed: learning sense embeddings for word and relational similarity. In *ACL-IJCNLP 2015 : The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 95–105, Beijing, China.
- Jurgens, D. A., Mohammad, S. M., Turney, P. D., and Holyoak, K. J. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 356–364, Montreal, Canada.
- Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (short papers)*, pages 302–308.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *arXiv preprint arXiv: 1301.3781*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, Nevada, USA.
- Mikolov, T., Yih, W.-T., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, USA.
- Quesada, J., Kintsch, W., and Mangalath, P. (2004). Analogy-making as prediction using relational information and lsa vectors. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, page 1623, Austin, TX.
- Summers-Stay, D., Voss, C., and Cassidy, T. (2016). Using a distributional semantic vector space with a knowledge base for reasoning in uncertain conditions. *Biologically Inspired Cognitive Architectures*, 16:34–44.
- Turney, P. D. (2012). Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research (JAIR)*, 44:533–585.

- Turney, P. D. and Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278.
- Veale, T. (2004). Wordnet sits the sat: A knowledge-based approach to lexical analogy. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, pages 606–612, Valencia, Spain.
- Yang, D. and Powers, D. M. (2005). Measuring semantic similarity in the taxonomy of wordnet. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science*, pages 315–322.
- Zhila, A., Yih, W.-T., and Meek, C. (2013). Combining heterogeneous models for measuring relational similarity. In *Proceedings of NAACL-HLT*, pages 1000–1009.