ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 508-514
doi: 10.18178/wcse.2019.06.075

Integrate Words Internal Information to Improve Word Embeddings

Chuanxiang Tang[1+], Yun Tang[1]

[1] School of Software, University of Science and Technology of China, China

Abstract. we propose a method of improving word embeddings by fusing the hidden information within words, which is different from the traditional method of directly using morphological information on the surface of words to train word embeddings. Based on the average principle and two attention mechanisms, we propose to use the hidden information inside words, which is called the implied meanings of morphemes of words in this paper, and propose six implied meaning embedding models. The comparative experiments are carried out on two basic Natural Language Processing tasks, which prove that our models have more advantages than the classical models represented by CBOW, Skip-Gram and GloVe in mining semantic information. In addition, exploring the relationship between the importance of synthetic implied meanings and the word itself.

Keywords: average principle, attention mechanism, word embedding, fusion.

1. Introduction

At present, the derivative word embeddings have been successfully applied to many downstream Natural Language Processing (NLP) tasks. Such as, named entity recognition [1], text classification [2], and question answering [3]. Among many embedded methods, the Continuous Bag-of-Word (CBOW) [4] model and the Skip-Gram [4] model and the Global Vectors (GloVe) [5] model, are recognized by the industry insiders for their practicability and efficiency. However, these word-embedded methods only learn semantic information at the word level, but ignore the morphemes within words. In recent years, there have been many models that use meaningful morphological structures inside words, and their effectiveness has been proved [6]. Unfortunately, these models only utilize the surface morphology of morphemes.

The new scheme we explored is to use the implied meanings of morphemes to train word embeddings. The traditional word embeddings model may not be able to shorten the distance between *"unseeable"* and *"invisible"* in the vector space. Because the morphemes composition of the two words is different, especially *"see"* and *"vis"*. However, by replacing morphemes with the implied meaning of words, it is obvious that the meaning of *"unseeable"* and *"invisible"* is the same.

In this paper, we use three strategies of integrating to combine implied meanings of morphemes and adopt two ways of fusing to train, so six simple and efficient models are proposed, which are collectively called Implied Meaning (IM) models. We directly cover the corresponding word embeddings in the vocabulary without adding extra embedding for generating and training of implied meanings. We only need to create a mapping table of words to describe the relationship between words and the implied meanings of their morphemes. We performed IM models and other classic models on the two tasks of word similarity and analogical reasoning, respectively. Experiments show that the performance of IM is more advantageous than all other classic models. In short, the contributions of this paper are as follows:

- We averagely distribute the weights of the implied meanings according to morphemes, and also introduce two types of attention mechanisms to assign the weights of the implied meanings, it perfects the strategy of weight allocation, and provides a new idea for the distribution of weights.

[+] Corresponding author. Tel.: + 86-0512-87161188; fax: +86-0512-87161100.
E-mail address: zhengtankung@foxmail.com

• We adopt two fusion methods to train and improve word embeddings. Six implied meaning models are proposed, and experiments are carried out on two well-recognized datasets and other two datasets, to verify the excellence of IM models.

• Exploring that the importance of synthetic implied meaning corresponding to morpheme set to predict target words depends on the degree of similarity between them and words themselves. It provides a reference for other researchers to further exploit the implied meanings of morphemes.

2. Related Work

2.1. Word Level

Generally, word embedding models are mainly divided into two types based on neural networks and matrix-based decomposition. CBOW [4] and Skip-gram [4] are widely used models based on neural networks, which are just opposite effects to each other. Skip-gram predicts context by using target words. The Latent Semantic Analysis (LSA) model [7] is a very classical matrix factorization model, the singular value decomposition of the word-document co-occurrence matrix is used to obtain the subject, word representation and document representation. In order to take advantage of these two kinds of advantages at the same time, Pennington et al. proposed the famous GloVe model [5], which is better than CBOW and Skip-gram in some specific tasks. These models can collect better semantic information at word level, but they do not mine morphological information of words.

2.2. Morphological Structure-based

The fine-grained models are proposed by using the components that make up words, such as roots and affixes. Luong et al. [6] proposed a morphological recurrent neural network model to learn morphological perceptual word embedding, according to morphology the words were segmented to generate morphemes, and the generated morphemes were added to the training of word embedding. Kim et al. [8] integrated convolutional character information into words. This model, which can learn semantic information from the character-level, has proved to be an effective way to deal with morpheme-rich languages, but it takes an astonishing amount of time. Cotterell et al. [9] used a log-linear model to make words with similar morphologies close to each other. Cotterell et al. [10] constructed a model using Gauss graphs and used morphemes to infer the continuous representation of unknown words. However, these models can only collect information on the morphological surface of the word without digging deeper meanings. In contrast, the model we proposed not only uses morphological information, but more importantly, it digs deep implied meanings.

3. Our Models

We IM models are based on efficient CBOW with negative sampling. When fusing the implied meanings of morphemes, we adopt three strategies of integrating and two ways of fusing to train. Six models named Implied Meaning-Average-Average (IMA-A), Implied Meaning-Hard-Average (IMH-A), Implied Meaning-Soft-Average (IMS-A), Implied Meaning-Average-Weighting (IMA-W), Implied Meaning-Hard-Weighting (IMH-W), and Implied Meaning-Soft-Weighting (IMS-W), are proposed. The first three are mainly different in the strategy of combining implied meanings of morphemes. The main difference between the first three and the last three is that the way of fusion is different. Obviously, many words contain multiple affixes, this paper chooses the longest sequence of characters that can be matched as the final morpheme of the word. Then, we will describe each IM model in detail.

3.1. IMA-A

IMA series (including IMA-A and IMA-W) are based on the fact that words are composed of prefixes, roots and suffixes, so it is assumed that these three components contribute equally to the word. In a given corpus of $\mathrm{T} = \{t_1, t_2, \ldots, t_n\}$, the implicit meaning of $t_i \in T, i \in [1, n]$ morpheme is divided into three parts: P_i, R_i and S_i, which respectively denote the prefix implicit meaning set, the root implicit meaning set and the suffix implicit meaning set. Therefore, the modification of t_i in the input layer is as follows:

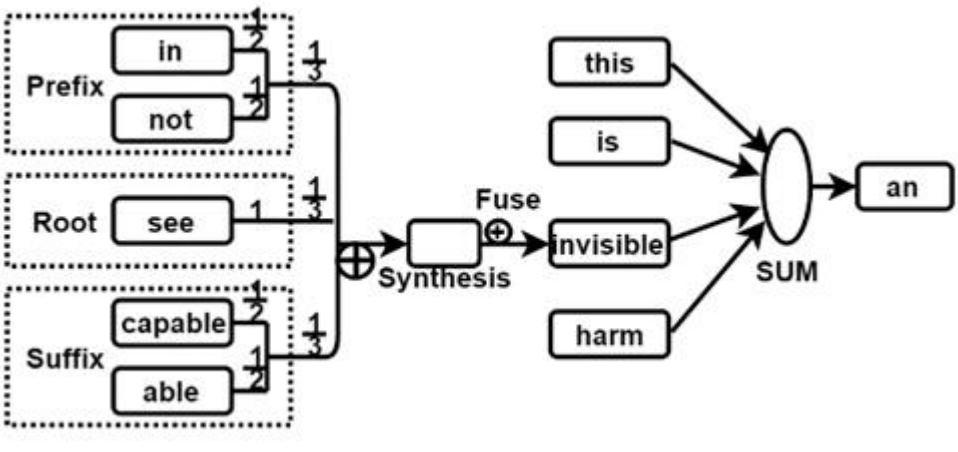

Fig. 1: IMA-A. We use the sentence "this is an invisible harm" as an example. When calculating the input vector of "invisible", we split the word into prefix, root and suffix, three parts have same weight. Implied meanings also gain the same weight inside each part.

$$\hat{v}_{t_i} = a_1 \cdot v_{t_i} + a_2 \cdot \frac{1}{3}\left(\frac{1}{N_{P_i}}\sum_{w_1 \in P_i} v_{w_1} + \frac{1}{N_{R_i}}\sum_{w_2 \in R_i} v_{w_2} + \frac{1}{N_{S_i}}\sum_{w_3 \in S_i} v_{w_3}\right) \tag{1}$$

Where v_{t_i} is the original word embedding of t_i. N_{P_i}, N_{R_i} and N_{S_i} denote the length of P_i, R_i and S_i, respectively. v_{w_1}, v_{w_2} and v_{w_3} represent prefix implied vector, root implied vector and suffix implied vector, respectively. Since IMA-A assume that the synthetic implied meaning of the morpheme is as important as the word itself for the prediction of the target word, that is, $a_1 = a_2 = 0.5$. Ultimately, v_{t_i} is replaced by $\hat{v}_{t_i}$ for CBOW training.

3.2. IMH-A

However, in fact, the root of a word has a decisive effect on the meaning of a word, while suffixes play a supplementary role in more cases, so unimportant implied meanings are excessively increased that can be opposite effect. Inspired by the Hard Attention mechanism, IMH series (including IMH-A and IMH-W) are proposed to focus only on the implied meaning closest to the word meaning, that is, selecting an implied meaning that is closest to the meaning of words from the implied meaning set of morphemes. For IMH-A, the embedding of t_i can be modified to:

$$\hat{v}_{t_i} = a_1 \cdot v_{t_i} + a_2 \cdot v_{max}^i \tag{2}$$

$$v_{max}^i = arg\max_w cos\left(v_{t_i}, v_w\right), w \in M_i \tag{3}$$

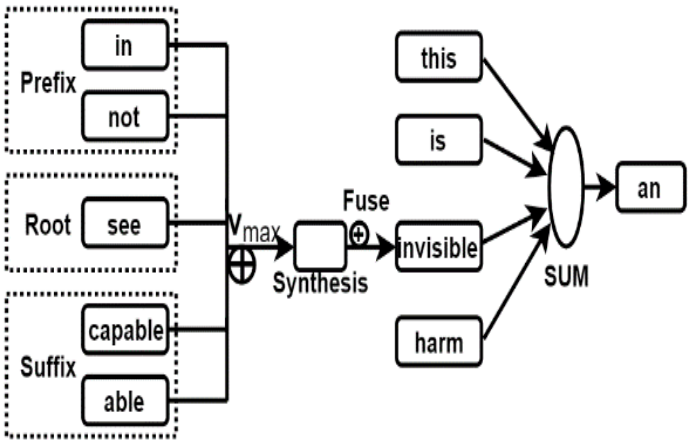

Fig. 2: IMH-A. Choosing only an implied meaning closest to the meaning of "invisible".

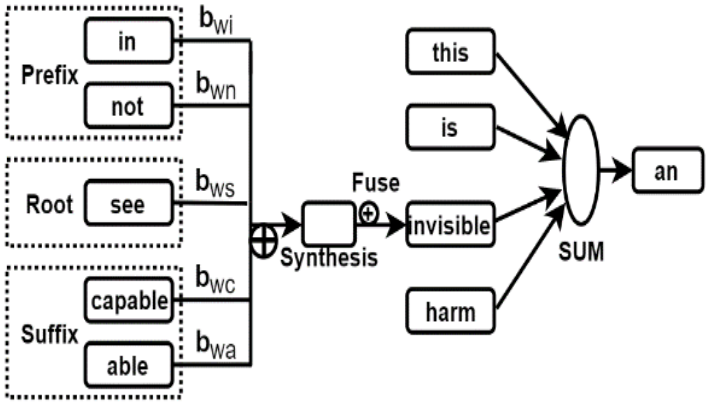

Fig. 3: IMS-A. All Implied meanings of morphemes of "invisible" obtain different weights.

We use $\cos(v_a, v_b)$ to represent the cosine similarity between v_a and v_b, where v_{max}^{i} is variable value when this function gets the maximum value. v_w indicates the word embedding of implied meaning w. M_i is all the implied meanings of t_i morphemes, which consists of three parts: P_i, R_i and S_i. The same reason $a_1 = a_2 = 0.5$.

3.3. IMS-A

We found that IMS-A model might be able to cause other cases where useful implications are missing, such as "$invisible$", which is easily understood to the opposite mean in the case of only choosing the implied meaning "see", and interferes with word embedding. Therefore, considering the Soft Attention mechanism, we put forward the IMS series (including IMS-A and IMS-W), and believe that Implied meanings of all morphemes contribute to the meaning of words, and the weight is assigned according to their contribution. The more the contribution of the implied meaning, the more weight it gains, so as to enhance its importance. The same reason $a_1 = a_2 = 0.5$. The embedding of t_i is obtained by the following equation:

$$\hat{v}_{t_i} = a_1 \cdot v_{t_i} + a_2 \cdot \sum_{w \epsilon M_i} b_w \cdot v_w \tag{4}$$

$$b_w = cos(v_{t_i}, v_w) \Big/ \sum_{x \epsilon M_i} cos(v_{t_i}, v_x) \tag{5}$$

3.4. Differences between IMX-A and IMX-W

The difference between IMA-A and IMA-W is that IMA-A believes that the synthetic implied meanings of morphemes of a word is as important as the word itself for predicting the target word, so the above(1),(2),(4) equation $a_1 = a_2 = 0.5$, while IMA-W considers that the contribution of the synthetic implied meaning of the morphemes to predicting the target word depends on the degree of similarity between it and the word itself, so the synthesized implied meaning and the word are fused by weighted averaging. The same is true of the differences between IMH-A, IMS-A and IMH-W, and IMS-W. IMX-A Contains IMA-A IMH-A and IMS-A, and IMX-W Contains IMA-W IMH-W, and IMS-W. Equations (6) (7) for a_1 and a_2 values in IMX-W.

$$a_1 = \frac{\cos(v_{t_i}, v_{t_i})}{\cos(v_{t_i}, v_{t_i}) + \cos(v_{t_i}, v_s)} = \frac{1}{1 + cos(v_{t_i}, v_s)} \tag{6}$$

$$a_2 = \frac{\cos(v_{t_i}, v_s)}{\cos(v_{t_i}, v_{t_i}) + \cos(v_{t_i}, v_s)} = \frac{cos(v_{t_i}, v_s)}{1 + cos(v_{t_i}, v_s)} \tag{7}$$

Where $\cos(v_{t_i}, v_{t_i}) = 1$

IMA-W modified part of the formula:

$$v_s = \frac{1}{3}\left(\frac{1}{N_{P_i}}\sum_{w_1 \in P_i} v_{w_1} + \frac{1}{N_{R_i}}\sum_{w_2 \in R_i} v_{w_2} + \frac{1}{N_{S_i}}\sum_{w_3 \in S_i} v_{w_3}\right) \tag{8}$$

IMH-W modified part of the formula:

$$v_s = v_{max}^{i} \tag{9}$$

IMS-W modified part of the formula:

$$v_s = \sum_{w \in M_i} b_w \cdot v_w \tag{10}$$

4. Experimental Settings

4.1. Corpus and Morpheme Mapping Table

In this paper, we use corpus which is originates from the website of ACL Machine Translation[1] Seminar in 2013 [8]. We chose the news corpus of 2009, which is about 1.8GB in size. It contains more than 0.5billion tokens and more than 0.6 million words. We filter out all the numbers and punctuation symbols in the corpus to achieve better quality of embedded words.

First, we use Morfessor [11] to perform unsupervised morphological segmentation of words in the vocabulary. Then, the matching between segmentation results and implied meaning is made in the implied meaning table, in which the final morpheme is selected according to the rules of morpheme matching, and then is further replaced by its implied meaning. Because this article focuses on verifying that IM models use the implied meaning of morphemes to improve word embedding is more superior, and it is just a simple common sense of a language, so this article uses artificially created mapping tables and only contains 102 prefixes, 403 roots and 86 suffixes.

4.2. Contrast Models

In order to make the experimental results more convincing, we chose the most well-known word embedding models of three word-levels: CBOW [4], Skip-gram [4] and GloVe [5]. We also implemented a Morpheme Enhance Word Embedding (MEWE) model, its structure is a variant of a morphological-based recurrent neural network model [6]. The structure of MEWE model is similar to our IMA-A model, but it uses morpheme to embedding directly. The source code for our training CBOW and Skip-gram is word2vec[2]. Glove is trained using Pennington et al. open source code[3] [5]. We modified the original word2vec code to train IM models and MEWE.

4.3. Parameter Settings

Because parameter setting directly affects the performance of word embeddings [12], all models are trained with the same parameters to ensure fairness and justice. We used negative sampling techniques to speed up the training process. According to the size of the corpus used in this paper, we choose to set it to 20 [13]. The dimension of word embedding is set to 200 [14]. Setting the context window size to 5 [13].

4.4. Word Similarity

This experiment is to test the ability of word embeddings to extract semantic information from corpus. If the related words are more similar in the vector space of a certain model, then the model is considered to have better semantic mining ability. This paper uses four public datasets, including two recognized standard

[1] http://www.statmt.org/wmt13/translation-task.html
[2] https://github.com/dav/word2vec
[3] https://github.com/maciejkula/glove-python

datasets Wordsim-353 [15] and RG-65 [16]. In order to avoid contingency, the data set used at the same time is Rare-Word (RW) [9] and Men-3k [17]. In this paper, the distance between two words be measured by using cosine similarity [13], [5], and the correlation is detected by Spearman rank sum coefficient (ρ). The higher the ρ value, the better the performance.

4.5. Analogical Reasoning

According to the fact that two objects are identical or similar in some attributes, the reasoning process that they also are the same in other attributes is inferred by comparison. For example: amazing, amazingly, apparent, apparently, namely, "$a \longrightarrow b,\ c \longrightarrow d$". Where d is unknown, let v_a, v_b, v_c, v_d denote the words a, b, c, d, respectively. To obtain d, we first calculate $\hat{v}_d = v_b - v_a + v_c$. Then, we find that the maximum cosine distance from $\hat{v}_d$ is the word $\hat{d}$. Therefore, set d to $\hat{d}$. In this article, we use the MRAR (Microsoft Research Analogical Reasoning) datasets. This 8,000-size data set was created by Mikolov [18].

5. Experimental Results

5.1. Word Similarity Results

In Table 1, it can be seen that the IM models surpass all classical models on four datasets. In particular, our models' performance on the two datasets of recognized standards (Wordsim-353 and RG-65) is approximately 7% and 10% higher than the traditional CBOW model, respectively. On the Men-3k, IMS-W achieves the results of 71.53%. The obvious contrast confirms the superiority of our models. By fusing morphemes, MEWE also performs better than other classic models, but there is still a certain gap compared to the performance of our IM models. In fact, MEWE only makes the words with similar morphemes distribute more closer in the vector space, and does not excavate deeper semantic information, because it simply adds morphological information to word embedding. By comparing the results of IMX-A and IMX-W, it is easy to find that the weighted fusion of synthetic implied meanings and words can achieve better effects. Especially, IMH-W is 3.75% ahead of IMH-A on RG-65 datasets.

Table 1. Results (%) on word similarity and analogical reasoning (AR). The bold numbers indicate the highest values.

	CBOW	**Skip-g**	**GloVe**	**MEWE**	**IMA-A**	**IMH-A**	**IMS-A**	**IMA-W**	**IMH-W**	**IMS-W**
Wordsim-353	58.57	61.63	49.41	60.05	62.25	61.56	63.15	64.45	62.69	**65.56**
RG-65	56.46	62.76	59.87	60.82	62.77	63.01	62.53	63.02	**66.76**	64.56
RW	40.32	36.27	33.38	40.86	43.43	40.69	42.16	**45.45**	41.98	42.96
Men-3k	68.03	66.25	60.46	66.79	66.39	64.68	69.37	67.56	65.63	**71.53**
AR	13.34	13.09	13.74	17.36	20.56	18.35	17.6	**22.57**	19.23	18.58

5.2. Results of Analogical Reasoning

As shown in Table 1, our IM models have a greater advantage than other classic models, leading the classic CBOW model by 9.23%. According to our scheme, words with similar morphemes will move closer together, and there is a tendency to group near the implied meaning of the corresponding morphemes. This makes our model have obvious advantages in dealing with the problem of analogical reasoning. Analogical reasoning is actually a semantically related task, and "c" and "d" have similar attribute characteristics. The IM models have the ability to mine deeper semantic information, and therefore achieve better results than EMEW. It is worth noting that IMA-W is 2.01% higher than IMA-A, which confirms the above conclusion again.

6. Conclusion and Future Work

In this paper, we propose a set of new learning word embeddings schemes, which mine the deep information inside words and integrate it into word embeddings, instead of simply using morphological components. By introducing attention mechanisms, three strategies of combining implied meanings of morphemes and two ways of integration are created, and propose six implied meaning models, named IMA-A, IMH-A, IMS-A, IMA-W, IMH-W and IMS-W respectively. We choose three classical word embedding models for comparison, meanwhile，and implement a model that directly utilizes morphemes. Testing in two natural language processing tasks, the results show that the IM models have more advantages

than all the classic models on the four word similarity datasets, and in analogical reasoning task, our models still perform better than all classical models, even if WEME performance also is slightly inferior. By comparing the effects of different fusion methods, it is found that the weighted fusion of synthetic implied meanings and words can achieve better results. In the future, we hope to further explore the use of supervised deep learning models with implied meanings of morphemes to mine semantic information.

7. Acknowledgements

We are grateful to the reviewers for valuable opinions. Finally, we also would like to thank Yang Xu and Yuncheng Song for their help in writing thesis and guiding experiment.

8. Reference

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*. 2011, pp. 2493–2537.

[2] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical word embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[3] G. Zhou, T. He, J. Zhao, and P. Hu. Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of ACL*. 2015, pp. 250–259.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 2013.

[5] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*. 2014, 14: 1532–43.

[6] T. Luong, R. Socher, and C. D. Manning. Better word representations with recursive neural networks for morphology. In *Conference*. 2013, pp. 104–113.

[7] D. Scott, et al. Indexing by latent semantic analysis. *Journal of the American society for information science.* 1990, pp. 391-407.

[8] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush. Character-aware neural language models. *Computer Science*. 2015.

[9] R. Cotterell and H. Schutze. Morphological word-embeddings. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 1287–1292.

[10] R. Cotterell, H. Schutze, and J. Eisner. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.* 2016, 1: 1651–1660.

[11] M. Creutz and K. Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*. 2007.

[12] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics.* 2015,3: 211–225.

[13] T. Mikolov and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*. 2013.

[14] P. S. Dhillon, D. P. Foster, and L. H. Ungar. Eigenwords: Spectral word embeddings. *The Journal of Machine Learning Research*. 2015,16(1): 3035–3078.

[15] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. *ACM Transactions on information systems*. 2002, 20(1): 116-131.

[16] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*. 1965, 8(10): 627–633.

[17] E. Bruni, N.-K. Tran, and M. Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res. (JAIR)*. 2014, 49(1): 47.

[18] T. Mikolov, W. T. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*. 2013.