Multi-Granularity Chinese Word Embedding

Rongchao Yin[†‡], Quan Wang[†‡] , Rui Li[†‡], Peng Li[†‡*], Bin Wang[†‡]
[†]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[‡]University of Chinese Academy of Sciences, Beijing 100049, China
{yinrongchao,wangquan,lirui,lipeng,wangbin}@iie.ac.cn

Abstract

This paper considers the problem of learning Chinese word embeddings. In contrast to English, a Chinese word is usually composed of characters, and most of the characters themselves can be further divided into components such as radicals. While characters and radicals contain rich information and are capable of indicating semantic meanings of words, they have not been fully exploited by existing word embedding methods. In this work, we propose *multi-granularity embedding* (MGE) for Chinese words. The key idea is to make full use of such word-character-radical composition, and enrich word embeddings by further incorporating finer-grained semantics from characters and radicals. Quantitative evaluation demonstrates the superiority of MGE in word similarity computation and analogical reasoning. Qualitative analysis further shows its capability to identify finer-grained semantic meanings of words.

1 Introduction

Word embedding, also known as distributed word representation, is to represent each word as a real-valued low-dimensional vector, through which the semantic meaning of the word can be encoded. Recent years have witnessed tremendous success of word embedding in various NLP tasks (Bengio et al., 2006; Mnih and Hinton, 2009; Collobert et al., 2011; Zou et al., 2013; Kim, 2014; Liu et al., 2015; Iyyer et al., 2015). The basic idea behind is to learn the distributed representation of a word using its context. Among existing approaches, the continuous bag-of-words model (CBOW) and Skip-Gram model are simple and effective, capable of learning word embeddings efficiently from large-scale text corpora (Mikolov et al., 2013a; Mikolov et al., 2013b).

Besides the success in English, word embedding has also been demonstrated to be extremely useful for Chinese language processing (Xu et al., 2015; Yu et al., 2015; Zhou et al., 2015; Zou et al., 2013). The work on Chinese generally follows the same idea as on English, i.e., to learn the embedding of a word on the basis of its context. However, in contrast to English where words are usually taken as basic semantic units, Chinese words may have a complicated composition structure of their semantic meanings. More specifically, a Chinese word is often composed of several characters, and most of the characters themselves can be further divided into components such as radicals (部首).[1] Both characters and radicals may suggest the semantic meaning of a word, regardless of its context. For example, the Chinese word "吃饭 (have a meal)" consists of two characters " 吃 (eat)" and "饭 (meal)", where " 吃 (eat)" has the radical of "口 (mouth)", and "饭 (meal)" the radical of " 饣 (food)". The semantic meaning of "吃饭" can be revealed by the constituent characters as well as their radicals.

Despite being the linguistic nature of Chinese and containing rich semantic information, such word-character-radical composition has not been fully exploited by existing approaches. Chen et al. (2015) introduced a character-enhanced word embedding model (CWE), which learns embeddings jointly for words and characters but ignores radicals. Sun et al. (2014) and Li et al. (2015) utilized radical information to learn better character embeddings. Similarly, Shi et al. (2015) split characters into small components based on the Wubi method,[2] and took into account those components during the learning process. In their work, however, embeddings are learned only for characters. For a word, the embedding is generated by simply combining the embeddings of the constituent characters. Since not all Chinese word-

*Corresponding author: Peng Li.

[1] https://en.wikipedia.org/wiki/Radical_(Chinese_characters)
[2] https://en.wikipedia.org/wiki/Wubi_method

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 981–986,
Austin, Texas, November 1-5, 2016.

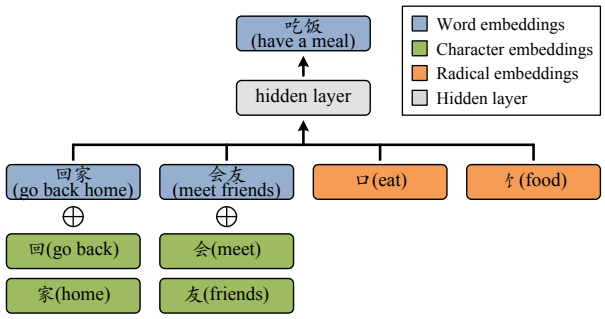

Figure 1: A simple illustration of MGE, where embeddings are learned jointly for words, characters, and radicals. Given a sequence of words {" 回家 (go back home)", " 吃饭 (have a meal)", " 会友 (meet friends)"}, MGE predicts the central word "吃饭" by using 1) the embedding composed by each context word and its constituent characters, and 2) the embedding associated with each radical detected in the target word.

s are semantically compositional (e.g., transliterated words such as "苏打 (soda)"), embeddings obtained in this way may be of low quality for these words.

In this paper, aiming at making full use of the semantic composition in Chinese, we propose *multi-granularity embedding* (MGE) which learns embeddings jointly for words, characters, and radicals. The framework of MGE is sketched in Figure 1. Given a word, we learn its embedding on the basis of 1) the context words (blue bars in the figure), 2) their constituent characters (green bars), and 3) the radicals found in the target word (orange bars). Compared to utilizing context words alone, MGE enriches the embeddings by further incorporating finer-grained semantics from characters and radicals. Similar ideas of adaptively using multiple levels of embeddings have also been investigated in English recently (Kazuma and Yoshimasa, 2016; Miyamoto and Cho, 2016).

We evaluate MGE with the benchmark tasks of word similarity computation and analogical reasoning, and demonstrate its superiority over state-of-the-art metods. A qualitative analysis further shows the capability of MGE to identify finer-grained semantic meanings of words.

2 Multi-Granularity Word Embedding

This section introduces MGE based on the continuous bag-of-words model (CBOW) (Mikolov et al., 2013b) and the character-enhanced word embedding model (CWE) (Chen et al., 2015).

MGE aims at improving word embedding by leveraging both characters and radicals. We denote the Chinese word vocabulary as $\mathcal{W}$, the character vocabulary as $\mathcal{C}$, and the radical vocabulary as $\mathcal{R}$. Each word $w_i \in \mathcal{W}$ is associated with a vector embedding $\mathbf{w}_i$, each character $c_i \in \mathcal{C}$ a vector embedding $\mathbf{c}_i$, and each radical $r_i \in \mathcal{R}$ a vector embedding $\mathbf{r}_i$. Given a sequence of words $\mathcal{D} = \{w_1, \cdots, w_N\}$, MGE predicts each word $w_i \in \mathcal{D}$ conditioned on 1) context words in a sliding window with size ℓ, denoted as $\mathcal{W}_i = \{w_{i-\ell}, ...w_{i-1}, w_{i+1}, ..., w_{i+\ell}\}$, 2) characters in each context word $w_j \in \mathcal{W}_i$, denoted as $\mathcal{C}_j$, and 3) radicals in the target word w_i, denoted as $\mathcal{R}_i$. See Figure 1 for a simple illustration.

More specifically, given the corpus $\mathcal{D}$, MGE maximizes the overall log likelihood as follows:

$$L(\mathcal{D}) = \sum_{w_i \in \mathcal{D}} \log p(\mathbf{w}_i | \mathbf{h}_i). \quad (1)$$

Here $\mathbf{h}_i$ is a hidden vector composed by the embeddings of context words, constituent characters, and radicals, defined as:

$$\mathbf{h}_i = \frac{1}{2}\Big[\frac{1}{|\mathcal{W}_i|}\sum_{w_j \in \mathcal{W}_i}\Big(\mathbf{w}_j \oplus \frac{1}{|\mathcal{C}_j|}\sum_{c_k \in \mathcal{C}_j}\mathbf{c}_k\Big) + \frac{1}{|\mathcal{R}_i|}\sum_{r_k \in \mathcal{R}_i}\mathbf{r}_k\Big]. \quad (2)$$

For each context word $w_j \in \mathcal{W}_i$, a word-character composition $(\mathbf{w}_j \oplus \frac{1}{|\mathcal{C}_j|}\sum_{c \in \mathcal{C}_j}\mathbf{c})$ is first generated by the embeddings of w_j and its constituent characters $\mathcal{C}_j$. These word-character compositions are then combined with the radical embeddings in $\mathcal{R}_i$ to predict the target word. $|\mathcal{W}_i|$/$|\mathcal{R}_i|$/$|\mathcal{C}_j|$ is the cardinality of $\mathcal{W}_i$/$\mathcal{R}_i$/$\mathcal{C}_j$, and $\oplus$ is the composition operation.[3] Given $\mathbf{h}_i$, the conditional probability $p(\mathbf{w}_i|\mathbf{h}_i)$ is defined by a softmax function:

$$p(\mathbf{w}_i|\mathbf{h}_i) = \frac{\exp(\mathbf{h}_i^\top \mathbf{w}_i)}{\sum_{w_{i'} \in \mathcal{W}} \exp(\mathbf{h}_i^\top \mathbf{w}_{i'})}. \quad (3)$$

We use negative sampling and stochastic gradient descent to solve the optimization problem.

Note that 1) Not all Chinese words are semantically compositional, e.g., transliterated words and entity names. For such words we use neither characters nor radicals. 2) A Chinese character usually plays

[3]There are a variety of options for $\oplus$, e.g., addition and concatenation. This paper follows (Chen et al., 2015) and uses the addition operation.

different roles when it appears at different positions within a word. We follow (Chen et al., 2015) and design a position-based MGE model (MGE+P). The key idea of MGE+P is to keep three embeddings for each character, corresponding to its appearance at the positions of "begin", "middle", and "end". For details, please refer to (Chen et al., 2015).

3 Experiments

We evaluate MGE with the tasks of word similarity computation and analogical reasoning.

3.1 Experimental Setups

We select the Chinese Wikipedia Dump[4] for embedding learning. In preprocessing, we use the THULAC tool[5] to segment the corpus. Pure digit words, non-Chinese words, and words whose frequencies are less than 5 in the corpus are removed. We further crawl from an online Chinese dictionary[6] and build a character-radical index with 20,847 characters and 269 radicals. We use this index to detect the radical of each character in the corpus. As such, we get a training set with 72,602,549 words, 277,200 unique words, 8,410 unique characters, and 256 unique radicals. Finally, we use THULAC to perform Chinese POS tagging on the training set and identify all entity names. For these entity names, neither characters nor radicals are considered during learning. Actually, Chen et al. (2015) categorized non-compositional Chinese words into three groups, i.e., transliterated words, single-morpheme multi-character words, and entity names. In their work, they used a human-annotated corpus, manually determining each word to be split or not. Since human annotation could be time-consuming and labor intensive, we just consider automatically identified entity names.

We compare MGE with CBOW (Mikolov et al., 2013b)[7] and CWE (Chen et al., 2015)[8]. Both CWE and MGE are extensions of CBOW, with the former taking into account characters and the latter further incorporating radical information. We further consider position-based CWE and MGE, denoted as CWE+P and MGE+P, respectively.We follow (Chen et al., 2015) and use the same hyperparameter setting. For all the methods, we set the context window size to 3, and select the embedding dimension k in $\{100, 200\}$. During optimization, we use 10-word negative sampling and fix the initial learning rate to 0.025.

Method	WordSim-239		WordSim-293	
	$k=100$	$k=200$	$k=100$	$k=200$
CBOW	0.4917	0.4971	**0.5667**	**0.5723**
CWE	0.5121	0.5197	0.5511	0.5655
CWE+P	0.4989	0.5026	0.5427	0.5545
MGE	**0.5670**	**0.5769**	0.5555	0.5659
MGE+P	0.5511	0.5572	0.5530	0.5692

Table 1: Results on word similarity computation.

3.2 Word Similarity Computation

This task is to evaluate the effectiveness of embeddings in preserving semantic relatedness between two words. We use the WordSim-240 and WordSim-296 datasets[9] provided by Chen et al. (2015) for evaluation, both containing Chinese word pairs with human-labeled similarity scores. On WordSim-240 there is a pair containing new words (i.e., words that have not appeared in the training set), and on WordSim-296 there are 3 such pairs. We remove these pairs from both datasets, and accordingly get WordSim-239 and WordSim-293.

We compute the Spearman correlation coefficient (Myers et al., 2010) between the similarity scores given by the embedding models and those given by human annotators. For the embedding models, the similarity score between two words is calculated as the cosine similarity between their embeddings. The Spearman correlation coefficient is a nonparametric measure of rank correlation, assessing how well the relationship between two variables can be described. The results are shown in Table 1.

From the results, we can see that 1) On WordSim-239, MGE(+P) performs significantly better than CWE(+P), which in turn outperforms CBOW. This observation demonstrates the superiority of incorporating finer-grained semantics, particularly from radicals. For example, MGE performs much better on word pairs such as "银行 (bank)" and " 钱 (money)", in which the two words share the same radical of "钅 (gold)". 2) On WordSim-293, MGE(+P)

[4] http://download.wikipedia.com/zhwiki
[5] http://thulac.thunlp.org/
[6] http://zd.diyifanwen.com/zidian/bs/
[7] https://code.google.com/p/word2vec/
[8] https://github.com/Leonard-Xu/CWE
[9] https://github.com/Leonard-Xu/CWE/tree/master/data

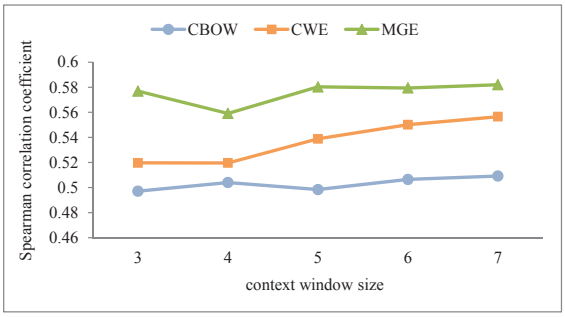

Figure 2: Word similarity computation results with different context window sizes on WordSim-239 ($k = 200$).

performs equally well as CWE(+P), but both are slightly worse than CBOW. The reason may be that WordSim-293 contains a great many of word pairs in which the two words belonging to different domains, e.g., "公鸡 (rooster)" and "航程 (flying range)". These pairs usually get low human-labeled similarity scores. However, splitting the words in such pairs into characters, and further the characters into radicals will not help to effectively identify the *dissimilarity* between them.[10]

We further investigate the influence of the context window size in word similarity computation. Figure 2 gives the results of CBOW, CWE, and MGE on WordSim-239, with the context window size set in $\{3, 4, 5, 6, 7\}$. The results indicate that MGE performs consistently better than CBOW and CWE on this dataset, unaffected by varying the context window size.

3.3 Word Analogical Reasoning

This task evaluates the effectiveness of embeddings in capturing linguistic regularities between pairs of words, in the form of " 伦敦 (London) : 英国 (England) ≈ 巴黎 (Paris) : 法国 (France)". We use the dataset provided by Chen et al. (2015) for evaluation. It contains 1,124 analogies categorized into 3 types: 1) capitals of countries (677 groups); 2) states/provinces of cities (175 groups); and 3) family relations (272 groups). All the words in this dataset are covered by the training set.

Method	Total	Capital	State	Family
CBOW	0.7498	0.8109	0.8400	**0.5294**
CWE	0.7248	0.8375	0.8541	0.3566
CWE+P	0.7391	0.8065	0.8114	0.5147
MGE	0.7524	**0.8804**	0.8686	0.3529
MGE+P	**0.7720**	0.8685	**0.8857**	0.4485

Table 2: Results on word analogical reasoning ($k = 200$).

For each analogy "$a : b \approx c : d$", we create a question "$a : b \approx c :$? ", and predict the answer as: $d^* = \arg\max_{w \in \mathcal{W}} \cos(\mathbf{b} - \mathbf{a} + \mathbf{c}, \mathbf{w})$. Here $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{w}$ are the word embeddings, and $\cos(\cdot, \cdot)$ the cosine similarity. The question is considered to be correctly answered if $d^* = d$. We use accuracy as the evaluation metric, and report the results in Table 2.

The results indicate that 1) MGE(+P) substantially outperforms the baseline methods on almost all types of analogies (except for the Family type). This again demonstrates the superiority of incorporating radical information. 2) For the Capital and State types, all the words are entity names for which neither characters nor radicals are used. MGE(+P) still outperforms the baselines on these two types, showing its capability to learn better embeddings even for non-compositional words. 3) On the Family type, both MGE(+P) and CWE(+P) perform worse than CBOW. This may be caused by the inappropriate decomposition of family words into characters. Consider, for example, the question " 叔叔 (uncle) : 阿姨 (aunt) ≈ 王子 (prince) : ? ". If we split " 王子" into "王 (king)" and "子 (son)", we will more likely to predict " 女王 (queen)" rather than the correct answer "公主 (princess)", since "女王" contains the character " 女 (daughter)" which is usually the antonym of " 子 (son)".

3.4 Case Study

Besides quantitative evaluation, this section further provides qualitative analysis to show in what manner the semantic meaning of a radical, character and word can be captured by their embeddings.

Take the word "游泳 (swimming)" as an example. Table 3 presents the words that are most similar to it (with the highest cosine similarity between their embeddings), discovered by MGE, CWE, and CBOW. The results show that 1) By incorporating the character information, MGE and CWE are capable of

[10] This observation is inconsistent with that reported in (Chen et al., 2015), which shows that CWE outperforms CBOW on WordSim-296. The reason may be that Chen et al. (2015) used a human-annotated corpus for embedding learning, and manually determined each word to be split or not. In contrast, we use the publicly available Chinese Wikipedia data, and automatically segment the corpus and identify entity names (words that are not to be split), without human annotation.

MGE	潜泳(underwater swimming), 畅泳(swimming happily) 爬泳(front crawl swimming), 泳手(swimmer) 泳术(swimming skill), 冬泳(winter swimming) 裸泳(swimming skill), 田径(track and field)
CWE	潜泳(underwater swimming), 畅泳(swimming happily) 爬泳(front crawl swimming), 田径(track and field) 泳手(swimmer), 习泳(learn to swim) 冬泳(winter swimming), 泳术(swimming skill)
CBOW	田径(track and field), 跳高(high jump) 跳水(diving), 跳绳(rope skipping) 划船(boating), 撑竿跳(pole vaulting) 皮划艇(canoeing), 体操(gymnastics)

Table 3: The most similar words to "游泳 (swimming)".

Radical	疒(illness)
Closest characters	佝(rickets) 痫(chronic disease) 偻(bending one's back) 疬(epidemic disease) 痨(tuberculosis) 淬(quenching) 疥(scabies) 痔(hemorrhoids)
Closest words	佝偻(rickets) 癣疥(ringworm scabies) 痘疤(pock) 瘴疬(communicable subtropical disease) 疮痍(traumata) 疮疤(scar) 痲疹(measles) 疱疮(pemphigus)

Table 4: The most similar characters/words to "疒 (illness)".

capturing finer-grained semantics that are more specific to the word. The top words discovered by them are semantically related to "游泳 (swimming)" itself, e.g., "潜泳 (underwater swimming)" and "爬泳 (front crawl swimming)". But the top words discovered by CBOW are just other types of sports in parallel with " 游泳 (swimming)", e.g., " 跳高 (high jump)" and " 跳水 (diving)". 2) MGE performs even better than CWE by further incorporating the radical information. The less relevant word " 田径 (track and field)" is ranked 4th by CWE. But after introducing the radical "氵 (water)", MGE can successfully rank "泳手 (swimmer)", "泳术 (swimming skill)", and " 冬泳 (winter swimming)" before it. All these words contain the radical "氵 (water)" and are more relevant to "游泳 (swimming)".

We further take the radical "疒 (illness)" as an example, and list the most similar characters and words discovered by MGE in Table 4. The similarity between a radical and a character/word is also defined as the cosine similarity between their embeddings. From the results, we can see that almost all the characters and words are disease-related, e.g., "佝 (rickets)", "痨 (tuberculosis)", and "癣疥 (ringworm scabies)", and most of them share the same radical "疒 (illness)". This observation demonstrates the rationality of embedding Chinese words, characters, and radicals into the same vector space, and measuring their similarities directly in that space. Note that this operation might be problematic for English. For example, it could be hard to figure out what kind of similarity there is between the character "i" and the word "ill". But for Chinese, this problem might be alleviated since characters and radicals themselves contain rich semantic information.

4 Conclusion and Future Work

In this paper we propose a new approach to Chinese word embedding, referred to as *multi-granularity embedding* (MGE). MGE improves word embedding by further leveraging both characters and radicals, and hence makes full use of the word-character-radical semantic composition. Experimental results on word similarity computation and analogical reasoning demonstrate the superiority of MGE over state-of-the-art methods. A qualitative analysis further shows that by incorporating radical information MGE can identify finer-grained semantic meanings of words.

As future work, we would like to 1) Investigate more complicate composition manners among radicals, characters, and words, e.g., a hierarchical structure of them. 2) Explore the semantic composition of higher level language units such as phrases, sentences, and even documents.

5 Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments and suggestions. This research is supported by the National Natural Science Foundation of China (grant No. 61402465 and No. 61402466) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant No. XDA06030200).

References

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint learning of character and

word embeddings. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1236–1242.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1681–1691.

Hashimoto Kazuma and Tsuruoka Yoshimasa. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. 2015. Component-enhanced chinese character embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 829–834.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. 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.

Yasumasa Miyamoto and Kyunghyun Cho. 2016. Gated word-character recurrent language model. *arXiv preprint arXiv:1606.01700*.

Andriy Mnih and Geoffrey E. Hinton. 2009. A scalable hierarchical distributed language model. In *Advances in Neural Information Processing Systems 21*, pages 1081–1088.

Jerome L Myers, Arnold Well, and Robert Frederick Lorch. 2010. *Research design and statistical analysis*. Routledge.

Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. 2015. Radical embedding: Delving deeper to chinese radicals. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 594–598.

Yaming Sun, Lei Lin, Nan Yang, Zhenzhou Ji, and Xiaolong Wang, 2014. *Radical-Enhanced Chinese Character Embedding*, chapter Proceedings of the 21st International Conference on Neural Information Processing, pages 279–286.

Ruifeng Xu, Tao Chen, Yunqing Xia, Qin Lu, Bin Liu, and Xuan Wang. 2015. Word embedding composition for data imbalances in sentiment and emotion classification. *Cognitive Computation*, 7(2):226–240.

Mo Yu, Matthew R. Gormley, and Mark Dredze. 2015. Combining word embeddings and feature embeddings for fine-grained relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1374–1379.

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 250–259.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.