Word Embedding Evaluation Datasets and Wikipedia Title Embedding for Chinese

Chi-Yen Chen, Wei-Yun Ma
Institute of Information Science, Academia Sinica
128 Academia Road, Section 2, Nankang, Taipei, Taiwan 115
{ccy9332, ma}@iis.sinica.edu.tw

Abstract

Distributed word representations are widely used in many NLP tasks, and there are lots of benchmarks to evaluate word embeddings in English. However there are barely evaluation sets with large enough amount of data for Chinese word embeddings. Therefore, in this paper, we create several evaluation sets for Chinese word embedding on both word similarity task and analogical task via translating some existing popular evaluation sets from English to Chinese. To assess the quality of translated datasets, we obtain human rating from both experts and Amazon Mechanical Turk workers. While translating the datasets, we find out that around 30 percents of word pairs in the benchmarks are Wikipedia titles. This motivate us to evaluate the performance of Wikipedia title embeddings on our new benchmarks. Thus, in this paper, not only the new benchmarks are tested but some new improved approaches of Wikipedia title embeddings are proposed. We perform training of embeddings of Wikipedia titles using not only their Wikipedia context but also their Wikipedia categories, most of categories are noun phrases, and we identify the head words of the noun phrases by a parser for further emphasizing their roles on the training of title embeddings. Experimental results and the comprehensive error analysis demonstrate that the benchmarks can precisely reflect the approaches' quality, and the effectiveness of our improved approaches on Wikipedia title embeddings are also verified and analyzed in detail.

Keywords: word embedding, word embedding benchmark, Wikipedia title embedding, Wikipedia category, knowledge base

1. Introduction

Word embeddings are widely used in various natural language processing (NLP) tasks. Researches evaluates word embeddings on some extrinsic, practical NLP tasks, and also some intrinsic tasks. The two most popular kinds of intrinsic tasks are word similarity and analogical reasoning, and there are an array of benchmarks for the two kind of tasks in English. Take word similarity for instance, the MEN-3k (Bruni et al., 2012) dataset consists of 3,000 word pairs. The MTurk-287 (Radinsky et al., 2011) dataset consists of 287 word pairs. The SimLex-999 (Hill et al., 2016) consists of 999 word pairs. For analogical reasoning tasks, the dataset is composed of analogous word pairs. Each word pair is a tuple of word relations that follow a common syntactic relation. For instance, the Google analogy (Mikolov et al., 2013) dataset consists of 19,544 question pairs.

Though there are lots of benchmarks to evaluate word embeddings in English, there are barely evaluation sets with large enough amount of data for Chinese word embeddings. For example, researchers at York University released a dataset[1] for Chinese word similarity; nevertheless, there are only 50 word pairs in the trial data and around 500 word pairs in the full dataset. Chen and Ma (2017) also created a small amount of datasets for Chinese word embedding, but the quantity and types are still far from English counterpart. Therefore, we generate several new evaluation sets for Chinese word embedding on both word similarity task and analogical task through translating some existing popular evaluation sets from English to Chinese. To assess the quality of translated datasets, we obtain human rating from both experts and Amazon Mechanical Turk workers.

While translating the datasets, we find out that around 30 percents of word pairs are Wikipedia titles. This motivate us to evaluate the performance of Wikipedia title embeddings on our new benchmarks. Wikipedia provides not only structural data, i.e., knowledge graphs via info-boxes, but also the nonstructural data, i.e., Wikipedia text, and semi-structural data, i.e., the title's categories. Most Wikipedia categories are long noun phrases other than noun words, so they are sometimes able to provide more complete information than info-boxes. Chen and Ma (2017) attempted to obtain title embedding of Wikipedia based on Wikipedia's content and categories; however, just a small amount of datasets for evaluation are used to evaluate the effectiveness and are not able to provide comprehensive evaluation results and error analysis. Thus, in this paper, we introduce the procedure of creating the new benchmarks and describe the difficulties we face along with the solutions we adopted. In addition, we extend (Chen and Ma 2017)'s work on training of embeddings of Wikipedia titles by considering the titles' Wikipedia content and categories, which syntactic heads are restricted to be a head of noun phrase, identified by a parser. We evaluate our new approaches on our new benchmarks, aiming to present a more complete and comprehensive error analysis to demonstrate the benchmarks' effects and also verify the performance of our improved approaches on Wikipedia title embeddings.

2. Benchmarks for Chinese Word Embeddings Evaluation

In English, there are quite a lot datasets that are commonly used as benchmark for evaluating word embedding. However, there are barely evaluation sets with large enough amount of data for Chinese word embeddings. The two

[1] https://www.cs.york.ac.uk/semeval-2012/task6/index.php%3Fid=data.html

most popular kinds of benchmarks to evaluate word embedding are word similarity task and analogical reasoning task. Therefore, we create some benchmarks for Chinese word embedding on both kinds of tasks through translating some existing popular evaluation sets from English to Chinese and remains the scores of the original benchmarks. For word similarity task, we translate six datasets: **SimLex-999** (Hill et al., 2016), **MEN-3k** (Bruni et al., 2012), **MTurk-287** (Radinsky et al., 2011), **WordSim-353** (Finkelstein et al., 2001), also the partitioned datasets from WordSim-353, separated into two different relations, **WS353-Relatedness** and **WS353-Similarity** (Zesch et al., 2008; Agirre et al., 2009). For analogical reasoning task, we translate the **Google analogical dataset** (Mikolov et al., 2013).

We will illustrate the translation process and difficulties encountered during translation in the following sections. The datasets are available to download at the link[2].

Dataset		Original	Translated
Word Similarity	SimLex-999	999	999
	MEN-3k	3,000	3,000
	MTurk-287	287	287
	WordSim-353	353	353
	WS353-R	252	252
	WS353-S	203	203
Analogical Reasoning	Google	19,544	11,126

Table 1: Size of each dataset, including original version and translated version.

2.1. Translation Process

In this section, we will illustrate the translation process and policy we apply. First, to get appropriate translated Chinese words, we use authoritative online dictionaries as resources, including Cambridge Dictionary (English-Chinese), E-Hownet (Group and others, 2009; Huang et al., 2008; Chen et al., 2005), Merriam-Webster Dictionary, Oxford Dictionary, and Wiktionary[3]. To maintain the original semantic meaning in English, we disambiguate the appropriate Chinese words by using Cambridge Dictionary (English-Chinese) and E-Hownet. Some words in English have different meaning in different situations.

Take some word pairs from MEN-3k dataset, which consists of word pairs with similarity score in a 50 scale, for example, **(*palm*, *tree*) 37.0** and **(*hand*, *palm*) 44.0** both have relatively high similarity scores. According to the dictionaries, *palm* can either mean "手掌(the under part of the hand between the fingers and the wrist)" or "棕榈树(a tropical tree, shrub, or vine with a usually tall stem or trunk topped with large leaves that are shaped like feathers or fans)". Because the similarity scores are high for both **(*palm*, *tree*) 37.0** and **(*hand*, *palm*) 44.0** in the original dataset, we cannot just replace *palm* with a single meaning as we ought to maintain the original semantic meanings in English. Thus, we translate (*palm*, *tree*) to (棕榈树 *(palm)*, 树 *(tree)*) and (手 *(hand)*, 手掌 *(palm)*) to (*hand*, *palm*).

Furthermore, due to different language usages, in Google analogy dataset, we ignore the tuples based on certain English grammatical variations which do not exist on Chinese words, such as singular-plural nouns, verb tenses, third-person singular verb endings, and comparison of adjectives For instance, tuples like **(mouse, mice, goat, goats)**, **(describing, described, knowing, knew) (write, writes, decrease, decreases)**, and **(lucky, luckiest, wide, widest)** will be discarded during the translation since there are no proper mappings from English to Chinese for these grammar rules.

In addition, we ignore the appendix "的" and "地" of adjectives and adverbs relatively since the usage is relatively rare in daily Chinese language usage. For example, the tuple **(*immediate*, *immediately*, *rare*, *rarely*)** can be rigorously translated as *(立即的, 立即地, 稀有的, 稀有地)*; however, in daily language, it is uncommon for these phrases with "的" and "地". To exemplify, according to Cambridge Dictionary, the sentences: "We must make an immediate response." is translated as "我们必须立即作出反应。" and "We really ought to leave immediately." is translated as "我们真的应该马上就走。". None of "立即的 (immediate)" or "立即地 (immediately)" is used in daily language. After the translation process based on these considerations, the size of Google analogy dataset is reduced from 19544 words to 11126 words, shown in Table 1.

Even though we follow the translation policy, we still encounter some obstacles while translating the datasets. The difficulties will be elaborated in Section 2.2..

2.2. Difficulties

2.2.1. Word Similarity Datasets Translation

In word similarity datasets, some English word pairs have slight differences which are even unnoticeable in Chinese. Both words in a word pair have similar meanings but not exactly the same. Take word pairs from MEK-3k for instance, **(*stairs*, *staircase*) 49.0**, though they are different in English, both words mean "楼梯 (stairs)" in Chinese. The solution to this case is that we look the words up in WikiDiff[4] to figure out the slight differences between words and then select a different but appropriate meanings in Chinese, i.e., (*stairs*, *staircase*) is translated as (阶梯 *(stairs)*, 楼梯 *(staircase)*).

There are also words that have exact the same meaning but with different expressions; hence, the word pair's similarity score is not full mark. For example, the word pair **(*bicycle*, *bike*) 45.0**, though both words mean "a two-wheeled vehicle that you sit on and move by turning the two pedals", the similarity score is not 50. The solution to this case is that we look these words up in Cambridge Dictionary and get the meanings in different Chinese expressions separately. That is, we find explanations for both ***bicycle*** and ***bike*** are "脚踏车, 单车, 自行车", and then we choose different Chinese words for ***bicycle*** and ***bike*** relatively. Also, we look up the word "脚踏车" in E-Hownet to get its synonyms in Chinese. Thus, (*bicycle*, *bike*) is translated as (脚踏车 *(bicycle)*, 自行车 *(bike)*).

[2]http://ckipsvr.iis.sinica.edu.tw/ecemb/reg.php

[3]https://en.wiktionary.org/wiki/Wiktionary

[4]http://wikidiff.com/

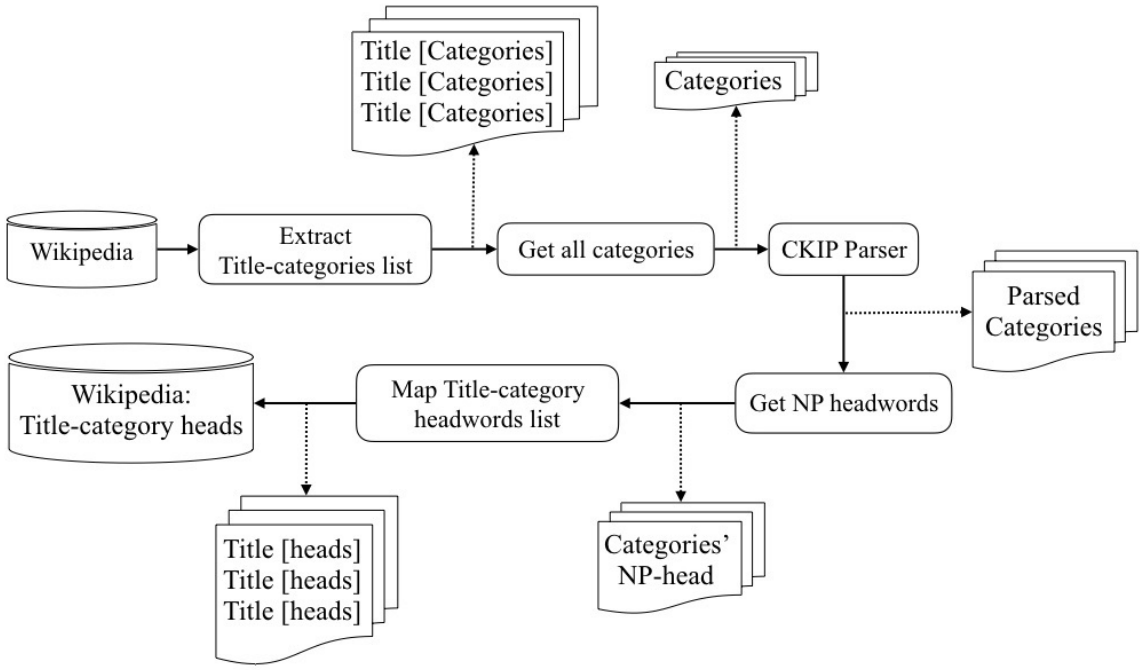

Figure 1: The work flow of getting categories' noun-phrase headwords mentioned in Section 3.3.2.

2.2.2. Analogical Reasoning Dataset Translation

In analogical reasoning questions set, some English words are difficult to find an appropriate mapping in Chinese words because they have more than one meanings. To solve this kind of problems, if a word has more than one forms as noun and others, we primarily consider its meaning as a noun, then as an adjective, and finally other forms because the original dataset mainly consists of nouns.

Take the tuple **(*croatia*, *croatian*, *thailand*, *thai*)** from Google analogical dataset, for instance. According to the dictionaries, *croatian* can be referred to "克罗埃西亚人 (a person from Croatia)", "克罗埃西亚语 (the language spoken in Croatia)", or "克罗埃西亚的 (belonging to or relating to Croatia, its people, or its language)". The solution to this case is that we choose "克罗埃西亚人 (a person from Croatia)" over other possible explanations because "克罗埃西亚人 (a person from Croatia)" is the first noun explanation in most dictionaries. While the same translation policy is applied, the translation of *thai* is "泰国人 (a person from Thailand)". That is, **(*croatia*, *croatian*, *thailand*, *thai*)** is translated as (克罗埃西亚, 克罗埃西亚人, 泰国, 泰国人).

3. Wikipedia Title Embedding

3.1. Wikipedia Data

While translating the datasets, we find out that around 30 percents of word pairs are Wikipedia titles. Wikipedia provides data in three structural extents: nonstructural, i.e., content, semi-structural, i.e., categories, and structural data, i.e., info-boxes. Categories are usually long noun phrases and provide more information than info-boxes. The page *Albert Einstein*, for example, is in categories of *ETH Zurich alumni*, *ETH Zurich faculty* and 20 more. The categories provide information that Einstein was both alumni and faculty of ETH Zurich while in info-box, only Einstein was related to ETH Zurich is shown.

We follow the work of (Chen and Ma 2017) and download the Chinese version of Wikipedia dump file in September 2016, which compromised 1,243,319 articles at a time. There are 244,430,247 words in the Chinese Wikipedia corpus and each title has 2.16 categories in average.

Chen and Ma (2017) obtained Wikipedia title embedding by linear combining context embedding and categories embedding. They generated context embedding by skipgram model and proposed several methods to generate categories embedding. In this paper, we use the same context embedding but purpose improved method to acquire categories embedding. We use CKIP Chinese parser (Hsieh et al., 2012; Yang et al., 2008; Hsieh et al., 2007) to parse each category and get the rigorously parsed noun-phrase (NP) headword.

3.2. Categories Embedding

Wikipedia categories can partially represent the corresponding title. Chen and Ma (2017) proposed several approaches of acquiring Wikipedia categories embedding, categories embedding for short. In this section, we first briefly introduce (Chen and Ma 2017)'s work, followed by description of our extension on the new approach to extract the headword of a category, elaborated in Section 3.3.2..

Formally, given a title t and its corresponding categories from C^1 to C^n, each category C^i has been word-segmented as K words from W_1^i to W_K^i and W_H^i is the headword of C^i. We use c^i and w_j^i to represent embedding of category C^i and word W_j^i.

3.2.1. Average of Category Words

We acquire categories embedding $e_{category}$ by averaging every category of a single title. Considering the complete-

WS Evaluation Sets	SimLex999		Men-3k		MTurk-287		ws353		ws353r		ws353s	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
skipgram	18.20	**40.30**	67.10	64.60	**50.40**	59.40	60.70	61.10	52.10	57.70	63.60	67.60
Avg(words)	18.00	39.00	**67.80**	**65.20**	50.20	59.90	**61.60**	**62.00**	**53.20**	**58.90**	**64.40**	**67.80**
Avg(headwords)	18.00	39.00	67.70	**65.20**	50.10	**60.20**	**61.20**	61.90	52.90	58.80	**64.40**	**67.80**
Avg(NP-heads)	18.10	**40.30**	67.30	64.90	49.80	59.30	60.50	60.80	51.60	57.60	63.40	67.10
wsc(d=1)	18.20	39.80	67.50	65.10	50.10	59.40	60.40	61.40	51.90	58.20	63.50	66.90
wsc(d=2)	18.20	39.80	67.50	65.10	50.10	59.50	60.40	61.30	51.90	58.30	63.60	66.80
wsc(headwords)	**18.30**	39.80	67.50	65.10	49.90	59.50	60.40	61.40	52.40	58.20	63.60	66.90
wsc(d=1, NP-heads)	18.10	**40.30**	67.30	64.90	49.80	59.30	60.50	60.80	51.60	57.60	63.40	67.10
wsc(d=2, NP-heads)	18.10	**40.30**	67.30	64.90	49.80	59.30	60.50	60.80	51.60	57.60	63.40	67.10
wsc(NP-heads)	18.10	**40.30**	67.30	64.90	49.80	59.30	60.50	60.80	51.60	57.60	63.40	67.10

Table 2: Spearman correlation on word similarity task. All embedding are 300 dimensions.

WS Evaluation Sets	SimLex999		Men-3k		MTurk-287		ws353		ws353r		ws353s	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Title coverage(%)	25.00	16.63	33.53	31.80	27.27	33.33	27.68	32.95	22.22	34.13	36.27	27.72
skipgram	9.40	49.40	69.40	73.20	**61.10**	55.90	62.40	78.10	**71.90**	76.10	75.40	79.80
Avg(words)	**10.80**	50.40	**70.60**	**74.50**	59.10	56.60	**63.40**	**78.20**	71.10	77.10	75.20	80.80
Avg(headwords)	10.70	50.60	70.50	**74.50**	58.50	**57.50**	62.90	78.00	71.00	**77.40**	75.00	**81.40**
Avg(NP-heads)	9.00	**50.90**	69.50	73.70	57.90	53.80	60.20	77.40	67.70	75.00	**75.60**	79.00
wsc(d=1)	10.30	50.20	70.20	**74.50**	59.80	55.90	62.00	77.60	71.30	76.80	**76.40**	80.20
wsc(d=2)	10.10	50.10	70.20	74.40	59.80	56.00	62.20	77.70	71.00	77.00	**76.40**	79.50
wsc(headwords)	10.40	50.50	70.30	**74.50**	59.70	55.80	61.70	77.70	**71.90**	76.40	**76.30**	79.70
wsc(d=1, NP-heads)	9.00	50.80	69.50	73.60	58.20	54.20	60.40	77.50	67.70	75.00	74.80	79.00
wsc(d=2, NP-heads)	9.00	50.80	69.50	73.60	58.20	54.20	60.40	77.50	67.70	75.00	74.80	79.00
wsc(NP-heads)	9.00	50.80	69.50	73.60	58.20	54.20	60.40	77.50	67.70	75.00	74.80	79.00

Table 3: Spearman correlation on word similarity task but only cope with word pairs that are Wikipedia titles. Wikipedia title coverage of each dataset is shown in the table. All embedding are 300 dimensions.

ness of category information, we obtain a category embedding by averaging all words in the category. Process of computing $e_{category}$ is shown as following:

$$e_{category} = \frac{1}{n}\sum_{i=1}^{n} c^i, \quad \text{where } c^i = \frac{1}{K}\sum_{j=1}^{K} w_j^i. \qquad (1)$$

3.2.2. Average of Category headwords

By observing Wikipedia data, we find out that most categories are noun-phrases and generally the headword of a noun-phrase contains more information than other words. We presume that sometimes other words besides the headword may bring some noisy information; thus, we acquire $e_{category}$ by averaging only every category headword of a single title and do not consider any other words. Computing process is shown as following:

$$e_{category} = \frac{1}{n}\sum_{i=1}^{n} c^i, \qquad \text{where } c^i = w_H^i. \qquad (2)$$

3.2.3. Weighted Sum of Categories (WSC)

We assume that each category of a title has different degree of representation and its representation depends on its headword's occurrence in context of the title. Therefore, we assert that categories should be treated distinctively according to their representation degrees. In this section, we acquire $e_{category}$ by summing up categories of a single title with the occurrence of category headword in context. Considering category information completeness, we obtain a category embedding by averaging all words' embeddings in the category but apply a different weight d to the headword. Computing process is shown as following:

$$e_{category} = \sum_{i=1}^{n} a_i c^i, \qquad (3)$$

where a_i is the category headword frequency in context with normalization and

$$c^i = \frac{1}{K+d-1}\left(\sum_{j=1}^{K} w_j^i + (d-1)\, w_H^i\right), \qquad (4)$$

where d is the weight added to the headword and will be adjusted during the experiments.

3.3. Headwords

In this section, we will introduce two different approaches we applied to extract headwords of categories: rule-based method and parsing-based method.

3.3.1. Rule-based Headwords Extraction

By observing Chinese linguistic structure, generally, the headword of a noun-phrase is the last word. Therefore, for each category C^i, we use its last word W_K^i as its headword. We denote the headwords obtained by this method as *headwords* during the evaluation section.

3.3.2. Phrasing-based Headwords Extraction

In this section, we illustrate the method of getting each category's headword by using CKIP Chinese Parser (Hsieh et al., 2012; Yang et al., 2008; Hsieh et al., 2007). Firstly, we extract title-categories mapping list, i.e., what categories are contained in each title page. Next, since there are many duplicated categories in different titles, we collect all categories as a set to reduce duplicated categories. This step can reduce the number of categories from around 2,691,151 to 378,540 and save around seven times of complexity. Then, we parse these categories one by one. After parsing the categories, we traverse through the parsed categories and extract the corresponding headwords of parsed categories only if the categories are noun-phrases. Finally, we create title-categories headword mapping list, that is, what headwords are contained in each title page. The procedure of getting noun-phrase headwords is shown as Figure 1. We denote the headwords obtained by this method as *NP-heads* during the evaluation section.

3.4. Title Embedding

Wikipedia title embedding, short for title embedding, is the improved word embedding with combination of context embedding and categories embedding. We acquire title embedding e_{title} by linear combining context embedding and categories embedding. The process of computing e_{title} is shown as following:

$$e_{title} = \alpha * e_{context} + (1 - \alpha) * e_{category}, \qquad (5)$$

where $0 < \alpha < 1$, $e_{context}$ is obtained from Skip-gram and $e_{category}$ is obtained from Section 3.2. The title embeddings are available at the link[5].

4. Evaluations

4.1. Datasets Size

In this section, we briefly introduce the size of datasets in the experiments. For word similarity task, there are six datasets: **SimLex-999**, **MEN-3k**, **MTurk-287**, **WordSim-353**, **WS353-Relatedness** and **WS353-Similarity** in Chinese version. For analogical reasoning task, there is one dataset: **Google dataset** in Chinese version. To get development set and testing set, we split every dataset in halves, each dataset's size is shown as Table 4.

[5]http://ckipsvr.iis.sinica.edu.tw/cwemb/reg.php

Dataset		dev	test
Word Similarity	SimLex-999	500	499
	MEN-3k	1,500	1,500
	MTurk-287	143	144
	WordSim-353	177	176
	WS353-R	126	126
	WS353-S	102	101
Analogical Reasoning	Google	5,563	5,563

Table 4: Size of each translated dataset, separated into development set and testing set.

4.2. Datasets Evaluation

To evaluate the translation quality, we ask some experts to manually rate similarity scores on our translated benchmarks. In Table 5, MEN-3k can get as high Spearman correlation with the original benchmark as 85, indicating that our translation process is able to preserve the original word semantics especially considering its large size. However MTurk-287 can only get Spearman correlation of 50. The reason could be either due to our translation process or because of the characteristics of the original dataset in English. To investigate the reason, we also obtain human rating from Amazon Mechanical Turk workers for MTurk-287. On Amazon Mechanical Turk, we assign each notation to three different workers and then average the similarity scores they provided. We collect human rating on both English and Chinese. In Table 6, we find that original dataset in English itself is difficult to reproduce the same correlation for Turkers, so it is more likely that MTurk-287 itself is hard to decide the similarity in nature instead of the translation problem.

	Score
MTruk-287 testing	50
MEN-3k testing	85

Table 5: Spearman correlation on original dataset and translated dataset.

	MTruk-287 testing
Original/ OurMTurk English	23
Original/ OurMTurk Chinese	21
mTruk English/ OurMTruk Chinese	49

Table 6: Spearman correlation on original dataset and translated dataset.

4.3. Word Similarity Tasks

Word similarity task datasets contain relatedness scores for word pairs; the cosine similarity of the two word embeddings should have high correlation. We have six

datasets: **SimLex-999**, **MEN-3k**, **MTurk-287**, **WordSim-353**, **WS353-Relatedness** and **WS353-Similarity** in Chinese version. We split datasets in halves to get development set and testing set, each dataset's size is shown as Table 4.
Table 2 shows the result of title embedding obtained in Section 3.4.. We tune the weight α via development set and then apply the best α, which $\alpha = 0.9$, on testing set. Comparing to the baseline, our proposed methods get significant improvement.
Furthermore, since we find out that about 30 percents of word pairs in the datasets are Wikipedia titles, we want to focus on these word pairs and figure out whether categories have positive effects on expressing the titles. Therefore, we evaluate the title embedding on only these word pairs which are Wikipedia titles in the datasets. Table 3 shows the result of title embedding obtained in Section 3.4., but only on the word pairs that are Wikipedia titles.
Based on both Table 2 and Table 3, we can conclude that categories indeed have positive effects on incorporating titles' information.

4.4. Analogical Semantics Tasks

Analogical reasoning dataset is compromised of analogous word pairs, i.e., pairs of tuples of word relations that follow a common syntactic relation. We use translated **Google dataset** and split it in halves as development and testing sets. Each set contains 5,563 questions.
Table 7 shows result of Linear Combination in 3.4.. We tune the weight α via development set and then apply the best α, which is $\alpha = 0.9$, on testing set. Comparing to the baseline, our proposed methods get significant improvement.

Method	Google	
	dev	test
Skip-gram	53.12	34.71
Avg(words)	**55.71**	**36.33**
Avg(headwords)	53.66	35.65
WSC (d=1)	**54.43**	**35.12**
WSC (d=2)	54.14	35.00
WSC (headwords)	53.17	34.96

Table 7: Accuracy on analogical reasoning task. All embeddings are 300 dimensions.

5. Discussion

5.1. Datasets

According to the evaluation result, we find out there is a high correlation between the quality of dataset and the evaluation scores. To be specific, according to Table 5, since Spearman correlation of MEN-3k is much higher than MTurk-287, we infer that MEN-3k has better quality compared with MTurk-287. The evaluation results on table 2 and table 3 indeed show that the title embeddings have better performance on MEN-3k compared with MTurk-287.
We observe that our approaches consistently get worse performances on some benchmarks than others. The reasons could be due to that for certain datasets, the similarity scores between words in word pairs are determined based on more various aspects. In general, the performance on SimLex-999 is the worst, so we look into the results of this dataset and indeed observe that there are more different aspects to express relations between words. Take a word pair in SimLex-999, which consists of word pairs with similarity score in a 10 scale, for instance, (**晚上 *(night)*, 白天 *(day)*) 1.88** has a low similarity score in the original dataset; nevertheless, words from title embedding that applies different categories embedding have much higher similarity as 5.45. Actually, we find out that *"*晚上 *(night)"* and *"*白天 *(day)"* have the same category: "一天里的时刻 (parts of a day)". This could explain why these two words in title embedding have higher similarity. On the orther hand, in original English dataset, *"*晚上 *(night)"* and *"*白天 *(day)"* are emphasized as entirely different time, which causes low similarity score.

5.2. Wikipedia Title Embedding

According to the experiment result, we can confirm that categories can provide valuable information for improving embedding sole based on context using skipgram.
We also find that for around half benchmarks, linear combination using category embedding which obtains from averaging all category words has the better performance. This circumstance reflects that in some cases, other words except headwords in the category could play more critical roles in the expression of the meaning of the category.
Another interesting finding is that, in some cases, the categories' representation degree fails to detect by the frequency of co-occurrence in the context since they are more related to other factors, such as the position of category appears in context. It is plausible that the category which denotes the first sentence in the context deserves most attention.
From error analysis, we find that the restriction of NP-head for categories based on parsing is especially able to filter out some noisy information in some benchmarks, such as SimLex999. For example, a title 神经 *(nerves)* has categories including 神经 *(nerves)*, 周围神经系统 *(peripheral nervous system)*, 神经解剖学 *(neuroanatomy)*, 软组织 *(soft tissue)* and 日语借词 *(Chinese words of Japanese origin)*. In this case, only 日语借词 *(Chinese words of Japanese origin)* is incorrectly parsed as verb-phrase and is discarded. Since it indeed has nothing to define the meaning of nerve, thus, without considering it, the title will be expressed more precisely. Although we can see the effect of the restriction on some benchmarks, we also observe that in some other benchmarks, the effect of this restriction is below our expectation. Based on our error analysis, one reason is that some categories are mistakenly parsed into verb-phrases due to no context was provided during parsing process and thus some important information is missed. For instance, a title 琴酒 *(gin)* has categories including 蒸馏酒 *(distilled wine)*; nevertheless, 蒸馏酒 *(distilled wine)* is incorrectly parsed as verb-phrase: 蒸馏　酒 *(to distill wine)*. In this case, the information of 蒸馏酒 will be discarded while we are acquiring categories' heads to be NP's head; hence the title embedding of 琴酒 *(gin)* will have no information from its categories. On the other hand, categories embedding which use the last word as headwords could still contain the information of 蒸馏酒 *(distilled drinks)*.

Another possible reason is that for some titles, their verb-phrased categories could be actually pretty critical to define or express the title. This possibility is worth further investigation in our future work.

6. Conclusion

In this paper, we create several benchmarks for Chinese word embedding on both word similarity task and analogical task through translating some existing popular evaluation sets from English to Chinese. To assess the quality of translated datasets, we obtain human rating from both experts and Amazon Mechanical Turk workers. And we also confirm that a Wikipedia title's categories can help define or complement the meaning of the title besides the title's Wikipedia context. Experimental results and the comprehensive error analysis demonstrate that the benchmarks can precisely reflect the approaches' quality. Furthermore, we compare with two different approaches to extract headwords of categories - rule-base method and parsing-based method, and find out both approaches have their pros and cons and are thus worth further investigating how to remains the pros and eliminate the cons in the future work.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Chen, C.-Y. and Ma, W.-Y. (2017). Embedding wikipedia title based on its wikipedia text and categories. In *Asian Language Processing (IALP), 2017 International Conference on*, pages 146–149. IEEE.

Chen, K.-J., Huang, S.-L., Shih, Y.-Y., and Chen, Y.-J. (2005). Extended-hownet: A representational framework for concepts. *Proceedings of OntoLex 2005-Ontologies and Lexical Resources*.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Group, C. et al. (2009). Lexical semantic representation and semantic composition-an introduction to e-hownet. Technical report, Technical Report), Institute of Information Science, Academia Sinica. 16 Yu-Ta Chen et al.

Hill, F., Reichart, R., and Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Hsieh, Y.-M., Yang, D.-C., and Chen, K.-J. (2007). Improve parsing performance by self-learning. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 2, June 2007*, 12(2):195–216.

Hsieh, Y.-M., Bai, M.-H., Chang, J. S., and Chen, K.-J. (2012). Improving pcfg chinese parsing with context-dependent probability re-estimation. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 216–221.

Huang, S.-L., Chung, Y.-S., and Chen, K.-J. (2008). E-hownet: the expansion of hownet. In *Proceedings of the First National HowNet Workshop*, pages 10–22. Citeseer.

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

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.

Yang, D.-C., Hsieh, Y.-M., and Chen, K.-J. (2008). Resolving ambiguities of chinese conjunctive structures by divide-and-conquer approaches. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Zesch, T., Müller, C., and Gurevych, I. (2008). Using wiktionary for computing semantic relatedness. In *AAAI*, volume 8, pages 861–866.