Analogical Reasoning on Chinese Morphological and Semantic Relations

Shen Li[1,2,♠] **Zhe Zhao**[3,♣] **Renfen Hu**[1,2,♠,†] **Wensi Li**[1,2,♦] **Tao Liu**[3,♣] **Xiaoyong Du**[3,♣]

♠{shen, irishere}@mail.bnu.edu.cn
♣{helloworld, tliu, duyong}@ruc.edu.cn
♦zjklws@163.com
[1] Institute of Chinese Information Processing, Beijing Normal University
[2] UltraPower-BNU Joint Laboratory for Artificial Intelligence, Beijing Normal University
[3] School of Information, Renmin University of China

Abstract

Analogical reasoning is effective in capturing linguistic regularities. This paper proposes an analogical reasoning task on Chinese. After delving into Chinese lexical knowledge, we sketch 68 implicit morphological relations and 28 explicit semantic relations. A big and balanced dataset CA8 is then built for this task, including 17813 questions. Furthermore, we systematically explore the influences of vector representations, context features, and corpora on analogical reasoning. With the experiments, CA8 is proved to be a reliable benchmark for evaluating Chinese word embeddings.

1 Introduction

Recently, the boom of word embedding draws our attention to analogical reasoning on linguistic regularities. Given the word representations, analogy questions can be automatically solved via vector computation, e.g. *"apples - apple + car ≈ cars"* for morphological regularities and *"king - man + woman ≈ queen"* for semantic regularities (Mikolov et al., 2013). Analogical reasoning has become a reliable evaluation method for word embeddings. In addition, It can be used in inducing morphological transformations (Soricut and Och, 2015), detecting semantic relations (Herdagdelen and Baroni, 2009), and translating unknown words (Langlais and Patry, 2007).

It is well known that linguistic regularities vary a lot among different languages. For example, Chinese is a typical analytic language which lacks inflection. Figure 1 shows that function words and reduplication are used to denote grammatical and semantic information. In addition, many semantic

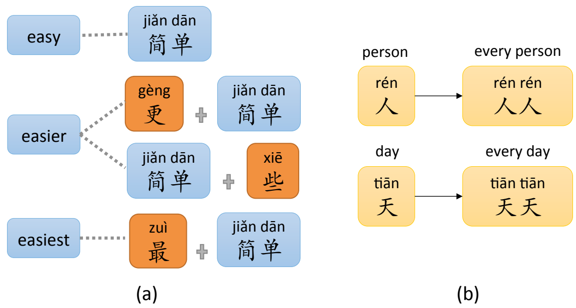

Figure 1: Examples of Chinese lexical knowledge: (a) function words (in orange boxes) are used to indicate the comparative and superlative degrees; (b) reduplication yields the meaning of *"every"*.

relations are closely related with social and cultural factors, e.g. in Chinese *"shī-xiān"* (god of poetry) refers to the poet *Li-bai* and *"shī-shèng"* (saint of poetry) refers to the poet *Du-fu*.

However, few attempts have been made in Chinese analogical reasoning. The only Chinese analogy dataset is translated from part of an English dataset (Chen et al., 2015) (denote as CA_translated). Although it has been widely used in evaluation of word embeddings (Yang and Sun, 2015; Yin et al., 2016; Su and Lee, 2017), it could not serve as a reliable benchmark since it includes only 134 unique Chinese words in three semantic relations (capital, state, and family), and morphological knowledge is not even considered.

Therefore, we would like to investigate linguistic regularities beneath Chinese. By modeling them as an analogical reasoning task, we could further examine the effects of vector offset methods in detecting Chinese morphological and semantic relations. As far as we know, this is the first study focusing on Chinese analogical reasoning. Moreover, we release a standard benchmark for evaluation of Chinese word embedding, together with 36 open-source pre-trained embeddings at

† Corresponding author.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 138–143
Melbourne, Australia, July 15 - 20, 2018.

GitHub[1], which could serve as a solid basis for Chinese NLP tasks.

2 Morphological Relations

Morphology concerns the internal structure of words. There is a common belief that Chinese is a morphologically impoverished language since a morpheme mostly corresponds to an orthographic character, and it lacks apparent distinctions between roots and affixes. However, Packard (2000) suggests that Chinese has a different morphological system because it selects different "settings" on parameters shared by all languages. We will clarify this special system by mapping its morphological analogies into two processes: reduplication and semi-affixation.

2.1 Reduplication

Reduplication means a morpheme is repeated to form a new word, which is semantically and/or syntactically distinct from the original morpheme, e.g. the word *"tiān-tiān"(day day)* in Figure 1(b) means *"everyday"*. By analyzing all the word categories in Chinese, we find that nouns, verbs, adjectives, adverbs, and measure words have reduplication abilities. Given distinct morphemes A and B, we summarize 6 repetition patterns in Figure 2.

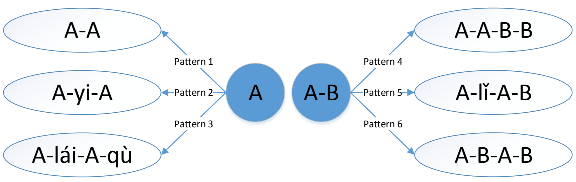

Figure 2: Reduplication patterns of A and A-B.

Each pattern may have one or more morphological functions. Taking *Pattern 1 (A→AA)* as an example, noun morphemes could form kinship terms or yield every/each meaning. For verbs, it signals doing something a little bit or things happen briefly. AA reduplication could also intensify an adjective or transform it to an adverb.

- *bà(dad) → bà-bà(dad)*
- *tiān(day) → tiān-tiān(everyday)*
- *shuō(say) → shuō-shuo(say a little)*
- *kàn(look) → kàn-kàn(have a brief look)*
- *dà(big) → dà-dà(very big; greatly)*
- *shēn(deep) → shēn-shēn(deeply)*

2.2 Semi-affixation

Affixation is a morphological process whereby a bound morpheme (an affix) is attached to roots or stems to form new language units. Chinese is a typical isolating language that has few affixes. Liu et al. (2001) points out that although affixes are rare in Chinese, there are some components behaving like affixes and can also be used as independent lexemes. They are called semi-affixes.

To model the semi-affixation process, we uncover 21 semi-prefixes and 41 semi-suffixes. These semi-suffixes can be used to denote changes of meaning or part of speech. For example, the semi-prefix *"dì-"* could be added to numerals to form ordinal numbers, and the semi-suffix *"-zi"* is able to nominalize an adjective:

- *yī(one) → dì-yī(first)*
  èr(two) → dì-èr(second)
- *pàng(fat) → pàng-zi(a fat man)*
  shòu(thin) → shòu-zi(a thin man)

3 Semantic Relations

To investigate semantic knowledge reasoning, we present 28 semantic relations in four aspects: geography, history, nature, and people. Among them we inherit a few relations from English datasets, e.g. country-capital and family members, while the rest of them are proposed originally on the basis of our observation of Chinese lexical knowledge. For example, a Chinese province may have its own abbreviation, capital city, and representative drama, which could form rich semantic analogies:

- *ān-huī* vs *zhè-jiāng* (province)
- *wǎn* vs *zhè* (abbreviation)
- *hé-féi* vs *háng-zhōu* (capital)
- *huáng-méi-xì* vs *yuè-jù* (drama)

We also address novel relations that could be used for other languages, e.g. scientists and their findings, companies and their founders.

4 Task of Chinese Analogical Reasoning

Analogical reasoning task is to retrieve the answer of the question "a is to b as c is to ?". Based on the relations discussed above, we firstly collect word pairs for each relation. Since there are no explicit word boundaries in Chinese, we take dictionaries and word segmentation specifications as references to confirm the inclusion of each word

[1] https://github.com/Embedding/Chinese-Word-Vectors

Benchmark	Category	Type	#questions	#words	Relation
CA_translated	Semantic	Capital	506	46	capital-country
		State	175	54	city-province
		Family	272	34	family members
CA8	Morphological	Reduplication A	2554	344	A-A, A-yi-A, A-lái-A-qù
		Reduplication AB	2535	423	A-A-B-B, A-lǐ-A-B, A-B-A-B
		Semi-prefix	2553	656	21 semi-prefixes: 大, 小, 老, 第, 亚, etc.
		Semi-suffix	2535	727	41 semi-suffixes: 者, 式, 主义, 性, etc.
	Semantic	Geography	3192	305	country-capital, country-currency, province-abbreviation, province-capital, province-dramma, etc.
		History	1465	177	dynasty-emperor, dynasty-capital, title-emperor, celebrity-country
		Nature	1370	452	number, time, animal, plant, body, physics, weather, reverse, color, etc.
		People	1609	259	finding-scientist, work-writer, family members, etc.

Table 1: Comparisons of CA_translated and CA8 benchmarks. More details about the relations in CA8 can be seen in GitHub.

Window (dynamic)	Iteration	Dimension	Sub-sampling	Low-frequency threshold	Context distribution smoothing	Negative (SGNS/PPMI)	Vector offset
5	5	300	1e-5	50	0.75	5/1	3COSMUL

Table 2: Hyper-parameter details. Levy and Goldberg (2014b) unifies SGNS and PPMI in a framework, which share the same hyper-parameter settings. We exploit 3COSMUL to solve the analogical questions suggested by Levy and Goldberg (2014a).

pair. To avoid the imbalance problem addressed in English benchmarks (Gladkova et al., 2016), we set a limit of 50 word pairs at most for each relation. In this step, 1852 unique Chinese word pairs are retrieved. We then build CA8, a big, balanced dataset for Chinese analogical reasoning including 17813 questions. Compared with CA_translated (Chen et al., 2015), CA8 incorporates both morphological and semantic questions, and it brings in much more words, relation types and questions. Table 1 shows details of the two datasets. They are both used for evaluation in Experiments section.

5 Experiments

In Chinese analogical reasoning task, we aim at investigating to what extent word vectors capture the linguistic relations, and how it is affected by three important factors: vector representations (sparse and dense), context features (character, word, and ngram), and training corpora (size and domain). Table 2 shows the hyper-parameters used in this work. All the text data used in our experiments (as shown in Table 3) are preprocessed via the following steps:

- Remove the html and xml tags from the texts and set the encoding as utf-8. Digits and punctuations are remained.
- Convert traditional Chinese characters into simplified characters with Open Chinese Convert (OpenCC)[2].
- Conduct Chinese word segmentation with HanLP(v_1.5.3)[3].

5.1 Vector Representations

Existing vector representations fall into two types, dense vectors and sparse vectors. SGNS (skip-gram model with negative sampling) (Mikolov et al., 2013) and PPMI (Positive Pointwise Mutual Information) (Levy and Goldberg, 2014a) are respectively typical methods for learning dense and sparse word vectors. Table 4 lists the performance of them on CA_translated and CA8 datasets under different configurations.

We can observe that on CA8 dataset, SGNS representations perform better in analogical reasoning of morphological relations and PPMI representations show great advantages in semantic relations. This result is consistent with performance of English dense and sparse vectors on MSR (morphology-only), SemEval (semantic-only), and Google (mixed) analogy datasets (Levy and Goldberg, 2014b; Levy et al., 2015). It is

[2]https://github.com/BYVoid/OpenCC
[3]https://github.com/hankcs/HanLP

Corpus	Size	#tokens	$\lvert V \rvert$	Description
Wikipedia	1.3G	223M	2129K	Wikipedia data obtained from https://dumps.wikimedia.org/
Baidubaike	4.1G	745M	5422K	Chinese wikipedia data from https://baike.baidu.com/
People's Daily News	3.9G	668M	1664K	News data from People's Daily (1946-2017) http://data.people.com.cn/
Sogou news	3.7G	649M	1226K	News data provided by Sogou Labs http://www.sogou.com/labs/
Zhihu QA	2.1G	384M	1117K	Chinese QA data from https://www.zhihu.com/, including 32137 questions and 3239114 answers
Combination	14.8G	2668M	8175K	We build this corpus by combining the above corpora

Table 3: Detailed information of the corpora. #tokens denotes the number of tokens in corpus. $|V|$ denotes the vocabulary size.

		CA_translated			CA8									
		Cap.	Sta.	Fam.	A	AB	Pre.	Suf.	**Mor.**	Geo.	His.	Nat.	Peo.	**Sem.**
SGNS	word	.706	.966	.603	.117	.162	.181	.389	.222	.414	.345	.236	.223	.327
	word+ngram	.715	**.977**	.640	.143	.184	.197	.429	.250	.449	.308	.276	.310	.368
	word+char	.676	.966	.548	**.358**	**.540**	**.326**	**.612**	**.455**	.468	.226	.296	.305	.368
PPMI	word	.925	.920	.548	.103	.139	.138	.464	.226	.627	.501	.300	.515	.522
	word+ngram	**.943**	.960	**.658**	.102	.129	.168	.456	.230	**.680**	**.535**	**.371**	**.626**	**.586**
	word+char	.913	.886	.614	.106	.190	.173	.505	.260	.638	.502	.288	.515	.524

Table 4: Performance of word representations learned under different configurations. Baidubaike is used as the training corpus. The top 1 results are in **bold**.

probably because the reasoning on morphological relations relies more on common words in context, and the training procedure of SGNS favors frequent word pairs. Meanwhile, PPMI model is more sensitive to infrequent and specific word pairs, which are beneficial to semantic relations.

The above observation shows that CA8 is a reliable benchmark for studying the effects of dense and sparse vectors. Compared with CA_translated and existing English analogy datasets, it offers both morphological and semantic questions which are also balanced across different types [4].

5.2 Context Features

To investigate the influence of context features on analogical reasoning, we consider not only word features, but also ngram features inspired by statistical language models, and character (Hanzi) features based on the close relationship between Chinese words and their composing characters [5]. Specifically, we use word bigrams for ngram features, character unigrams and bigrams for character features.

Ngrams and Chinese characters are effective features in training word representations (Zhao et al., 2017; Chen et al., 2015; Bojanowski et al., 2016). However, Table 4 shows that there is only a slight increase on CA_translated dataset with ngram features, and the accuracies in most cases decrease after integrating character features. In contrast, on CA8 dataset, the introduction of ngram and character features brings significant and consistent improvements on almost all the categories. Furthermore, character features are especially advantageous for reasoning of morphological relations. SGNS model integrating with character features even doubles the accuracy in morphological questions.

Besides, the representations achieve surprisingly high accuracies in some categories of CA_translated, which means that there is little room for further improvement. However it is much harder for representation methods to achieve high accuracies on CA8. The best configuration only achieves 68.0%.

5.3 Corpora

We compare word representations learned upon corpora of different sizes and domains. As shown in Table 3, six corpora are used in the experiments: Chinese Wikipedia, Baidubaike, People's Daily News, Sogou News, Zhihu QA, and "Com-

[4] CA_translated and SemEval datasets contain only semantic questions, MSR dataset contains only morphological questions, and in Google dataset the capital:country relation constitutes 56.72% of all semantic questions.

[5] The SGNS with word and character features are implemented by fasttext toolkit, the rest are implemented by ngram2vec toolkit.

	CA_translated			CA8									
	Cap.	Sta.	Fam.	A	AB	Pre.	Suf.	**Mor.**	Geo.	His.	Nat.	Peo.	**Sem.**
Wikipedia 1.2G	.597	.771	.360	.029	.018	.152	.266	.180	.339	.125	.147	.079	.236
Baidubaike 4.3G	.706	.966	.603	.117	.162	.181	**.389**	.222	.414	**.345**	.236	**.223**	.327
People's Daily 4.2G	**.925**	**.989**	.547	.140	.158	**.213**	.355	**.226**	**.694**	.019	.206	.157	**.455**
Sogou News 4.0G	.619	.966	.496	.057	.075	.131	.176	.115	.432	.067	.150	.145	.302
Zhihu QA 2.2G	.277	.491	**.625**	**.175**	**.199**	.134	.251	.189	.146	.147	**.250**	.189	.181
Combination 15.9G	**.872**	**.994**	**.710**	**.223**	**.300**	**.234**	**.518**	**.321**	**.662**	**.293**	**.310**	**.307**	**.467**

Table 5: Performance of word representations learned upon different training corpora by SGNS with context feature of word. The top 2 results are in **bold**.

bination" which is built by combining the first five corpora together.

Table 5 shows that accuracies increase with the growth in corpus size, e.g. Baidubaike (an online Chinese encyclopedia) has a clear advantage over Wikipedia. Also, the domain of a corpus plays an important role in the experiments. We can observe that vectors trained on news data are beneficial to geography relations, especially on People's Daily which has a focus on political news. Another example is Zhihu QA, an online question-answering corpus which contains more informal data than others. It is helpful to reduplication relations since many reduplication words appear frequently in spoken language. With the largest size and varied domains, "Combination" corpus performs much better than others in both morphological and semantic relations.

Based on the above experiments, we find that vector representations, context features, and corpora all have important influences on Chinese analogical reasoning. Also, CA8 is proved to be a reliable benchmark for evaluation of Chinese word embeddings.

6 Conclusion

In this paper, we investigate the linguistic regularities beneath Chinese, and propose a Chinese analogical reasoning task based on 68 morphological relations and 28 semantic relations. In the experiments, we apply vector offset method to this task, and examine the effects of vector representations, context features, and corpora. This study offers an interesting perspective combining linguistic analysis and representation models. The benchmark and embedding sets we release could also serve as a solid basis for Chinese NLP tasks.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities, National Natural Science Foundation of China with Grant (No.61472428) and Chinese Testing International Project (No.CTI2017B12).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *IJCAI*. pages 1236–1242.

Etienne Denoual. 2007. Analogical translation of unknown words in a statistical machine translation framework. *Proceedings of Machine Translation Summit XI, Copenhagen* .

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*. pages 8–15.

Amac Herdagdelen and Marco Baroni. 2009. Bagpack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Association for Computational Linguistics, pages 33–40.

Philippe Langlais and Alexandre Patry. 2007. Translating unknown words by analogical learning. In *EMNLP-CoNLL*. pages 877–886.

Omer Levy and Yoav Goldberg. 2014a. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*. pages 171–180.

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

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

Yuehua Liu, Wenyu Pan, and Wei Gu. 2001. *Practical grammar of modern Chinese*. The Commercial Press.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *hlt-Naacl*. volume 13, pages 746–751.

Jerome L Packard. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

Radu Soricut and Franz Josef Och. 2015. Unsupervised morphology induction using word embeddings. In *HLT-NAACL*. pages 1627–1637.

Tzu-ray Su and Hung-yi Lee. 2017. Learning chinese word representations from glyphs of characters. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 264–273.

Peter D Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 905–912.

Liner Yang and Maosong Sun. 2015. Improved learning of chinese word embeddings with semantic knowledge. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Springer, pages 15–25.

Rongchao Yin, Quan Wang, Peng Li, Rui Li, and Bin Wang. 2016. Multi-granularity chinese word embedding. In *EMNLP*. pages 981–986.

Zhe Zhao, Tao Liu, Shen Li, Bofang Li, and Xiaoyong Du. 2017. Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 244–253.