Chengyu Cloze Test

Zhiying Jiang, Boliang Zhang, Lifu Huang and Heng Ji
Computer Science Department
Rensselaer Polytechnic Institute
{jiangz6, zhangb8, huangl7, jih}@rpi.edu

Abstract

We present a neural recommendation model for *Chengyu*, which is a special type of Chinese idiom. Given a query, which is a sentence with an empty slot where the Chengyu is taken out, our model will recommend the Chengyu candidate that best fits the slot context. The main challenge lies in that the literal meaning of a Chengyu is usually very different from its figurative meaning. We propose a neural approach to incorporate the definition of each Chengyu as background knowledge. Experiments on both Chengyu cloze test and coherence checking in college entrance exams show that our system achieves 89.5% accuracy on cloze test and outperforms human experts who attended competitive universities in China. We will make all of our data sets and resources publicly available as a new benchmark for research purposes[1].

1 Introduction

Chengyu ("成语", literal translation: "form phrases") is a special type of Chinese idiom, and represents one of the most beautiful, fascinating and unique aspects of the Chinese language. 96% Chengyus consist of four characters each. Chengyus were mainly created from ancient stories, literature and sayings which can be traced back to thousands of years ago. Some examples are shown in Table 1. More than 7,000 Chengyus are still widely used in the modern Chinese, Japanese, Korean and Vietnamese languages. Like idioms in other languages, using Chengyu appropriately makes communication more compelling and engaging because they introduce powerful imagery and figurative meanings that differ from their literal meanings.

When learning Chinese phrases, Chengyus are always the most difficult to understand and mem-orize. Second-language learners generally have a love-hate relation with Chengyu and tend to avoid it. A typical way to measure a Chinese learner's Chengyu knowledge is "*Cloze Test*", in which the learner is asked to supply the best Chengyu that has been removed from a sentence. It's considered as one of the most difficult problems in Chinese college entrance language and literature exams, and has been the focus of several TV talent shows in China such as the Chinese Idiom Congress by CCTV. This motivated us to develop the first Chengyu recommendation system to assist Chinese learners. Given a context sentence ("*query*") with a Chengyu removed, the system will automatically recommend the best Chengyu to fill in the blank.

The four characters in each Chengyu are often unintelligible without understanding the background story. For example, "沉鱼落雁 *(literal translation: sink fish fall swallow)*" and "闭月羞花 *(literal translation: hide moon shame flower)*" were used to summarize four stories of the top four beauties in ancient China: Xi Shi, Wang Zhaojun, Diao Chan and Yang Yuhuan. They were being so beautiful that fish sank, birds fell from the sky, the moon hid, and flowers were shamed. As a result, we cannot compose the meaning of a Chengyu only based on its four characters. Moreover, each Chengyu is highly succinct, compact and synthetic. For example, "一日三秋 *(literal translation: one day three autums)*" means greatly missing someone so that one day feels as long as three years. However, its key meaning "*missing*" is not in this Chengyu.

To address these challenges, we create a new Chengyu Cloze Test benchmark, which consists of 108,987 query sentences and 7,395 target Chengyus. Each Chengyu is associated with a definition, which describes its general meanings and scenarios where it occurs. Then we develop an

---

[1] https://github.com/bazingagin/chengyu_data

Origin	Query	Recommended Chengyu and Its Definition
Historical Story	文学接受史上经常有这样的现象，某些作品在它的那个时代曾经风行一时，____。 Throughout history some literary works were extremely popular, so much so that ____.	**洛阳纸贵 (Luoyang's paper became expensive)** 原指西晋都城洛阳之纸，因大家争相传抄左思的作品《三都赋》，以至一时供不应求，货缺而贵。比喻著作有价值，流传广。 During the Western Jin Dynasty, people kept copying and propagating Zuosi's work "San Du Fu", which caused the supply of paper falling short of demand. Now a metaphor to describe that some great work is valuable and widely disseminated.
Ancient Chinese Literature	三年的高中生活如____一般转瞬而逝。 The three years of high school life was like ____, time takes wings.	**白驹过隙 (time passes quickly like a white pony's shadow across a crevice)** 《庄子》："人生天地之间，若白驹之过隙，忽然而已。" Chuang Tzu said ``Human life between heaven and earth is like the white pony seen through a crack in the wall, it's just a moment.
Fable Proverb Saying	____，各显神通，预示着省楹联艺术家协会成立后必将涌现出成千上万不计其数的佳联妙对。 ____, each with their own special powers. This indicates that thousands of great couplets will emerge and decorate Shanjin beautifully.	**八仙过海 (Eight immortals cross the sea)** 相传八仙过海时不用舟船，各有一套法术，后比喻各自拿出本领或办法，互相竞赛。 It's said that eight immortals crossed the sea without boats because each of them had special power. Now it's used to describe using one's unique skill to compete.
Foreign Literature	我们喜欢用经济去控制一个国家的命脉,用信仰去控制一个种族,用利益让别人为我们____。 We like to use economy to control a country's faith, use belief to control a race and use profit to control others so they can ____ for us.	**火中取栗 (pull chestnuts from the embers)** 出自十七世纪法国寓言诗人拉·封丹的寓言《猴子与猫》。比喻受人利用去冒险，吃了苦头却得不到一点好处。 From the 17 century French fabulist Jean de la Fontaine's "The Monkey and the Cat". Bertrand the monkey persuades Raton the cat to pull chestnuts from the embers amongst which they are roasting, promising him a share. As the cat scoops them from the fire one by one, burning his paw in the process, the monkey gobbles them up. It's used to describe a person used unwittingly or unwillingly by another to accomplish the other's own purpose with his own risk but gets nothing.
Metaphor	这篇小说情节完整生动，人物性格鲜明，但____，个别语句还欠推敲。 This novel includes a complete and vivid plot, and the characters have distinct personalities. But it's like ____ - some sentences need to be further polished.	**白璧微瑕 (white jade with a little blemish)** 洁白的玉上有些小斑点。比喻很好的人或物有些小缺点，美中不足。 A flaw in a white jade. It's a metaphor for a good person or a good thing with a little defect.

Table 1: Chengyu Examples

attentive neural network architecture to select the most appropriate Chengyu to fit in the slot context of each query. We first encode query sentence and Chengyu definitions using a bi-directional long short-term memory (Bi-LSTM) network (Hochreiter and Schmidhuber, 1997). To better capture the correlation between the query and the definition, we apply a soft attention to assign a weight to each word in the query sentence, and predict a matching score for each candidate Chengyu. Our system significantly outperforms human learners who attended top universities in China.

2 Related Work

Our Chengyu cloze test task is similar to reading comprehension (Hermann et al., 2015; Cui et al., 2016; Chen et al., 2016; Kadlec et al., 2016; Seo et al., 2016). However, it's more challenging because the context includes a sentence instead of a paragraph, the Chengyu phrase itself does not convey its figurative meaning, and there are many more candidate answers. Very few Natural Language Processing techniques have been applied to understand or recommend Chengyu.

Chung (2009) studied a subset of Chinese figurative language, focusing on Chinese five elements and body part terms. Limited efforts have used Chengyu dictionaries to expand Chinese emotion lexicon (Xu et al., 2010) and improve Chinese word segmentation (Chan and Chong, 2008; Sun and Xu, 2011; Wang and Xu, 2017). Chengyus differ from metaphors in other languages (Tsvetkov et al., 2014; Shutova, 2010) because they do not follow the grammatical structure and syntax of the modern Chinese.

3 Approach

Figure 1 shows the overall architecture of our approach. For a query and the definition of a candidate Chengyu, we first apply a word segmentation tool jieba[2] to segment query and definition into words, and apply a Bi-LSTM network to encode each word with a contextual embedding. In order to better capture the correlation between a query and a Chengyu, we further compare the representations of the Chengyu definition and the contextual embedding of each word in the query, and
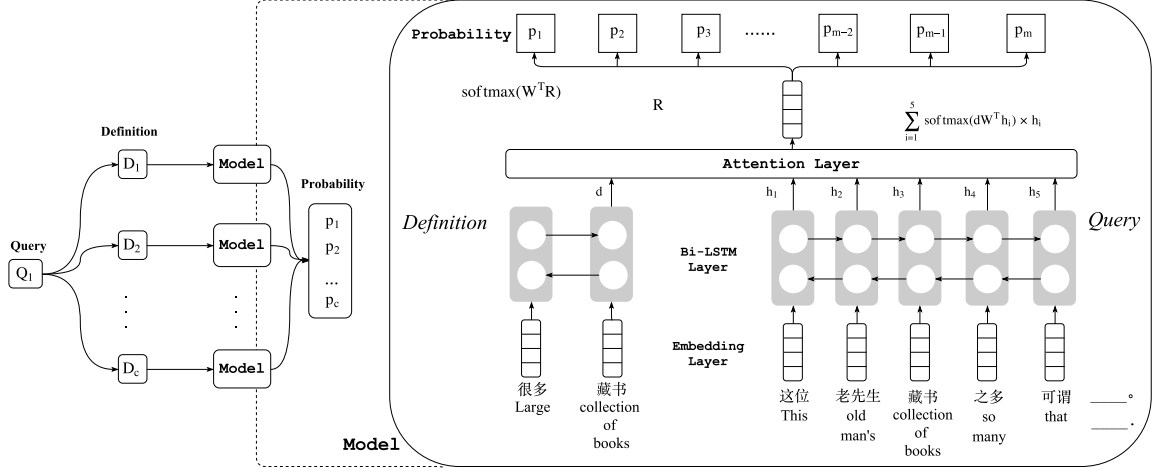
_____
[2]https://github.com/fxsjy/jieba

Figure 1: Architecture Overview

take the weighted sum of the query word contextual embeddings as input to a linear function to determine the probability score of the candidate Chengyu. Next we show the approach details.

Encoding Given a query q and a Chenyu definition d_j from the target Chengyu database $D = \{d_1, d_2, ..., d_m\}$, we apply two Bi-LSTM networks to encode them separately. Each Bi-LSTM network leverages long distance features from the whole sentence to capture the context information by using a memory cell (Hochreiter and Schmidhuber, 1997). Each word in q and d_j is assigned a contextual embedding.

Attention To better capture the correlation between a query and each Chengyu definition, we use an attention mechanism (Bahdanau et al., 2014; Sutskever et al., 2014) to compare the semantic relatedness of each word in the query sentence with the meaning of each Chengyu definition.

Given the hidden states $H = h_0, h_1, ..., h_n$ of the Bi-LSTM encoding the query sentence, where h_i denotes the concatenation of the hidden states of word w_i with forward and backward LSTMs, the attention layer sum over h_i with learnable weight α: $\mathbf{R} = \sum_{i=1}^{n} \alpha_i \cdot h_i$, where $\mathbf{R}$ is the weighted sum vector representation of the query. α_i is a learnable weight which is computed by $\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{n} \exp(e_i)}$ and $e_i = d^T \cdot \mathbf{W}_\alpha \cdot h_i$, where $\mathbf{W}_\alpha$ is a parameter to capture the relevance between a query and a definition flexibly (Chen et al., 2016). d^T is the last hidden hidden state of the Bi-LSTM encoding the definition.

Training With the weighted sum vector representation of the query $\mathbf{R}$, we apply a softmax function to compute the probability of each candidate Chengyu d_j to be filled into the slot.

$$\mathbf{o}_i = \mathbf{W}_\beta^T \mathbf{R}$$

$$p_i = \frac{\exp(\mathbf{o}_i)}{\sum_{j=1}^{m} \exp(\mathbf{o}_j)},$$

where $\mathbf{W}_\beta$ maps the final representation of the query into $\mathbb{R}^m$, and m is the number of classes. Then we optimize the log likelihood: $\mathbf{L} = \sum_{j=1}^{m} y_j log(p_j)$, where y_j is 0 or 1 depending on if the truth is Chengyu d_j or not.

Prediction For prediction, we take a query with each Chengyu definition (q, d_j), $1 \le j \le m$ as input, and predict a probability matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$, where m is the number of candidates. For example, a choose-one-from-four task will have $m = 4$. The final predicted Chengyu d_j is selected by $argmax(\mathbf{M}[:, j])$, $1 \le j \le m$.

4 Experiments

4.1 Data and Setting

We crawled 108,987 sentences including 7,395 unique idioms from http://zaojv.com, and the definitions of these idioms from http://cy.5156edu.com. Training and test set contain 108,432 and 555 sentences, and 7,071 and 508 Chengyus respectively. We use the whole Chengyu dataset to train word embeddings. We perform two tests: (1) **cloze test**: for each sentence in the test set, we take out the ground-truth Chengyu, and let the system select a Chengyu

TYPE	QUERY	SYSTEM	GROUND TRUTH	ANALYSIS
Incorporating Definition	这事已势不可遏，任何想阻挡他的人都如____，简直是不自量力。 This event is unstoppable, anyone who tries to stop it will be like ____, almost not recognizing his/her own limited power.	蚍蜉撼树 an ant shaking a tree, to describe one fails to recognize one's limited power	蚍蜉撼树 an ant shaking a tree, to describe one fails to recognize one's limited power	The definition significantly enriches the semantic meanings of Chengyu itself. 蚍蜉撼树(an ant shaking a tree) is a metaphor to describe 自不量力(fail to recognize one's own limited power).
Attention Mechanism	刘备思贤若渴，三请诸葛亮的故事在我国可是____，人人皆知的佳话。 It's ____ well known by everyone in our country that Liu Bei was eager to recruit talents and invited Zhu Geliang three times.	家喻户晓 well known by every family	家喻户晓 well known by every family	By incorporating the attention mechanism, our approach can better capture the correlations between query context and Chengyu definition. our approach successfully selects 家喻户晓 (well known by every family) to fill in the slot since it shares similar semantic meanings with query context word 知 (known).
World Knowledge	村上春树____，29岁才写他的第一部作品。 Haruki Murakami ____, he was already at age 29 when he wrote his first works.	画龙点睛 bring the painted dragon to life by putting in the pupils of its eyes	大器晚成 takes a long time to make a great instrument	We need to know "age 29" is relatively late to produce the first works for a writer.
Discourse Coherence	人们面临灾难，不得不____，离开他们自己的村落。 When facing disasters, people had to ____ and leave their own villages.	逍遥法外 at large	背井离乡 leave one's hometown	Our system focused on the shared meaning of escape/leave while ignored this Chengyu has a specific object "the arm of the law".
Sentiment Analysis	多少人认为一个作家不仅能妙笔生花，也是____的。 Many people think that a writer can not only write like an angel but also ____.	大言不惭 brag shamelessly	口若悬河 speak eloquently	大言不惭 (brag shamelessly) expresses very negative sentiment while 口若悬河 (speak eloquently) includes positive sentiment.
Negation Detection	你在他面前说那些话，实在是班门弄斧，不知____。 The words you said in front of him were really like showing off axe in front of Lu Ban, without knowing ____.	孤陋寡闻 with very limited knowledge and scanty information	天高地厚 high as heaven, deep as earth	Our system did not detect negation clues and thus failed to select the right Chengyu antonyms.
Grammatical Structure	写文章先要构思好，不要下笔千言，____。 We should think about the plot carefully before write an article, don't write down thousands of words,	词不达意 the words fail to express the meaning	离题万里 get away from the title ten thousands of miles	When multiple Chengyus appear in the same query sentence, they tend to follow the same grammatical structure.
Rhythm	爱是人性的美的力量，爱是爱你年少时的____，更爱你年老时白发苍苍。 Love is the beauty of humanity. To love is to love your youthful vigor like ____ as well as your gray hair.	意气风发 high-spirited and vigorous	桃之夭夭 the peach trees in full blossom	Multiple chengyus tend to appear in rhythmical form. In this example, "苍苍"(pronunciation: Cāng Cāng) and "夭夭"(Yāo Yāo) are both reduplication with similar vowel pronunciations.

🟩 Correct   🟦 Remaining Challenges

Table 2: Detailed Analysis on Correct Examples and Remaining Challenges

from four candidates consisting of the ground-truth and three other randomly selected ones to fill in the slot. (2) **coherence checking in college entrance exam**: we collected 14 problem sets from (1998, 2000) China college entrance exam, where each problem set consists of four sentences including Chengyus. We let the system select the sentence that contains the most appropriate Chengyu that fits into the context in a coherent way. For comparison with human, we asked two Chinese native speakers (not system developers) who attended top universities in China to perform the same tests.

4.2 Results and Analysis

	Cloze Test	Coherence Checking in College Entrance Exam
Human	70%	42.3%
System	89.5%	35.7%

Table 3: System and Human Accuracy Comparison

Table 3 shows our approach achieves comparable performance as human experts. For 18% of our system recommended Chengyus which don't exactly match the ground truth, they are also acceptable choices for the given query contexts.

For example, our system output "白驹过隙*(time passes quickly like a white pony's shadow across a crevice)*" and ground truth "光阴似箭*(time flies)*" are near synonyms. Table 2 shows some correct examples and the remaining challenges that require capabilities beyond lexical semantics.

5 Conclusions and Future Work

We created a new benchmark dataset for a new task of Chengyu cloze test. We also proposed a neural model which leverages the definitions of Chengyu as background knowledge and outperforms human experts. In the future we will explore collective inference to rank multiple Chengyus in the same discourse simultaneously, and incorporate richer linguistic clues based on structures and rhythms.

Acknowledgments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Samuel W.K. Chan and Mickey W.C. Chong. 2008. An agent-based approach to chinese word segmentation. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Siaw-Fong Chung. 2009. A corpus-based study on figurative language through the chinese five elements and body part terms. In *Proceedings of Computational Linguistics and Chinese Language Processing*.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension. *arXiv preprint arXiv:1607.02250*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Ekaterina Shutova. 2010. Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697. Association for Computational Linguistics.

Weiwei Sun and Jia Xu. 2011. Enhancing chinese word segmentation using unlabeled data. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 248–258.

Chunqi Wang and Bo Xu. 2017. Convolutional neural network with word embeddings for chinese word segmentation. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*.

Ge Xu, Xinfan Meng, and Houfeng Wang. 2010. Build chinese emotion lexicons using a graph-based algorithm and multiple resources. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*.