Carrier Sentence Selection for Fill-in-the-blank Items

Shu Jiang
Department of
Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
jshmjs45@gmail.com

John Lee
Department of
Linguistics and Translation
City University of Hong Kong
Hong Kong SAR, China
jsylee@cityu.edu.hk

Abstract

Fill-in-the-blank items are a common form of exercise in computer-assisted language learning systems. To automatically generate an effective item, the system must be able to select a high-quality carrier sentence that illustrates the usage of the target word. Previous approaches for carrier sentence selection have considered sentence length, vocabulary difficulty, the position of the target word and the presence of finite verbs. This paper investigates the utility of word co-occurrence statistics and lexical similarity as selection criteria. In an evaluation on generating fill-in-the-blank items for learning Chinese as a foreign language, we show that these two criteria can improve carrier sentence quality.

1 Introduction

Fill-in-the-blank items are a common form of exercise in computer-assisted language learning (CALL) systems. Also known as cloze or gap-fill items, they are constructed on the basis of *carrier sentences*. One word in the carrier sentence — called the *target word*, or the "key" — is blanked out. The learner attempts to fill in this blank, sometimes with the help of several choices, which include the target word itself and several distractors. Consider the following item, in Chinese, for the target word *xiande* 顯得 'appear, seem, look':

> 與其他生物相比，人類的生產過程 ___ 複雜許多。
>
> Compared with other organisms, the human reproductive process ___ a lot more complex.
>
> (A) *xiande* 顯得 'appear, seem, look'
> (B) ... (C) ... (D) ...

This carrier sentence not only makes a point about the human reproductive process, but also illustrates a typical usage of *xiande* by providing a comparison. In contrast, the following carrier sentence is an inferior choice for illustrating the same target word. Though perfectly fluent and grammatical, it does not offer any reason or context for the appearance of the pilot:

> 他回到駕駛艙，臉色 ___ 蒼白。
>
> He returned to the cockpit, and his face ___ pale.

Authoring fill-in-the-blank items, and composing carrier sentences in particular, is labor-intensive. There has thus been much interest in automatic generation of these items for self-directed language learners. Previous research on fill-in-the-blank item generation has mostly focused on finding plausible distractors. In the only published studies on carrier sentence selection (Volodina et al., 2012; Pilán et al., 2013), the system makes the selection based on sentence length, vocabulary difficulty, the position of the target word, and the presence of finite verbs. We show that two additional criteria — word co-occurrence and lexical similarity, which take into account the relation between the target word and other words in the carrier sentence — can improve the quality of the selected carrier sentence.

2 Previous work

This section summarizes previous work on carrier sentence selection (Section 2.1), supplemented by two related research areas: example sentence selection for dictionaries (Section 2.2), and text readability prediction (Section 2.3).

Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications, pages 17–22,
Taipei, Taiwan, December 1, 2017

2.1 Carrier sentence selection

Automatic generation of fill-in-the-blank items requires selection of distractors and carrier sentences. Previous research mostly focused on the former (Liu et al., 2005; Sumita et al., 2005; Chen et al., 2006; Smith et al., 2010; Sakaguchi et al., 2013; Zesch and Melamud, 2014). As for carrier sentences, some systems use example sentences from dictionaries (Pino and Eskenazi, 2009), while others impose relatively simple requirements (Meurers et al., 2010; Knoop and Wilske, 2013).

We are not aware of any reported work on carrier sentence selection for Chinese. To the best of our knowledge, apart from general guidelines (Haladyna et al., 2002; Xu, 2012), the only work on sentence selection for language-learning exercises focused on Swedish. Volodina et al. (2012) proposed an algorithm that uses four weighted heuristics — sentence length, position of the target word, absence of rare words and presence of finite verbs — to score each candidate sentence. In manual evaluation, 56.6% of the sentences were considered "acceptable". In a subsequent evaluation on the *Lärka* system (Pilán et al., 2013), human judges rated 73% of the automatically selected sentences to be "understandable"; they rated about 60% as suitable as exercise items or as examples for vocabulary illustration.

2.2 Example sentence selection

A dictionary entry of a word often includes an example sentence. Various criteria for selecting an example sentence for dictionaries have been proposed (Kilgarriff et al., 2008; Didakowski et al., 2012). As far as the sentence is concerned, it must be authentic, complete, well-formed, self-contained, and not too complex. As for the target word, it should not be used as a proper noun, or in a metaphoric or abstract sense; further, it should co-occur often with, and be semantically related to, one or more words in the rest of the sentence. A number of systems have implemented some of these criteria as heuristic rules (Smith et al., 2009; Baisa and Suchomel, 2014). In one study, these criteria yielded 95% success rate in example sentence selection (Didakowski et al., 2012). Since a carrier sentence should likewise illustrate the usage of its target word, we will borrow some of these criteria — specifically, word co-occurrence and lexical similarity — in this work.

2.3 Text readability prediction

Text readability prediction classifies a document into a difficulty level, typically a school grade. State-of-the-art systems combine lexical, syntactic and discourse features, as well as n-gram language model scores, to perform the classification (Schwarm and Ostendorf, 2005; Pitler and Nenkova, 2008; Kate et al., 2010; Collins-Thompson, 2008).

For Chinese, we are aware of only two reported studies on this task. Using similar features as above, Sung et al. (2015) achieved 72.92% accuracy in classifying textbook material into the six grades at primary school. Chen et al. (2013) showed that tf-idf and lexical chains can further improve accuracy, ranging from 80% to 96% for various grade levels on a set of textbooks.

Although carrier sentences must also be highly readable, features used in text readability prediction are not directly transferrable to our task. Most of these features are intended for documents, and may not work well when applied on single sentences. In a recent study on sentence-level readability prediction for Swedish, Pilán et al. (2014) found that a heuristic approach based on example sentence selection (Kilgarriff et al., 2008) outperformed a statistical classifier that adapts features from document-level readability prediction.

3 Features for carrier sentence selection

Because of the lack of large-scale, annotated dataset for carrier sentence selection, it is impossible to use standard machine learning methods for scoring candidate carrier sentences. Instead, we developed a number of features, to be used by the system as heuristics. We first describe baseline features (Section 3.1) inspired by Volodina et al. (2012), and then investigate word co-occurrence and lexical similarity statistics (Section 3.2).

To tune the heuristics, we compiled two datasets. The "Textbook Set" consists of 299 carrier sentences, drawn from fill-in-the-blank questions in three Chinese textbooks (Liu, 2004, 2010; Wang, 2007)[1]. The "Wiki Set" contains 9.2 million sentences, harvested from Chinese Wikipedia. We performed word segmentation, part-of-speech tagging and syntactic parsing with the Stanford Chinese parser (Levy and Manning, 2003).

[1] We excluded carrier sentences whose target words are not nouns, verbs, or adjectives.

3.1 Baseline features

Sentence complexity. In fill-in-the-blank items intended for self-directed learning, as opposed to assessment, simple sentences are preferred. This is to minimize the learner's difficulties in sentence comprehension and optimize his/her acquisition of the target word. An indicator of sentence complexity is sentence length; for example, Volodina et al. (2012) favor sentences between 10 and 15 words. The average length of carrier sentences in the Textbook Set is 16.8 words, substantially shorter than that of sentences in the Wiki Set (24.7 words). Besides sentence length, the number of clauses can also serve as an approximate measure of complexity. On average, the parse tree of a carrier sentence in the textbooks contains 3.1 IP nodes[2]. *We require a carrier sentence to have between 10 to 20 words, and to have no more than 3 IP nodes.*

Vocabulary difficulty. For similar reasons as described above, carrier sentences tend to avoid difficult words. Word frequency is often used as a proxy for its difficulty level. While a straightforward strategy is to set a minimum frequency threshold (Volodina et al., 2012), no fixed threshold can suit all individual learners, and a conservative threshold would unnecessarily reject good candidate sentences. We instead take the target word as the point of reference — a carrier sentence designed for teaching that word should not assume the learner to know words that are more advanced. We ranked all words by frequency in the Wiki Set, and divide them into buckets of 1,000 words. *We require all words in the carrier sentence to belong to the same bucket as, or a higher word-frequency bucket than, the target word.*

Target word position. While Volodina et al. (2012) prefer target words to be located within the first 10 words of the sentence, the target words in the Textbook Set tend to occur in the second half of the sentence, and fewer than 1% are within the first tenth of the sentence.[3] *We require that the target word cannot be situated within the first tenth of the carrier sentence.*

Complete sentence. To favor complete sentences, the heuristic used by Volodina et al. (2012) rewards the presence of a finite verb. Since Chinese verbs do not mark finiteness, we instead require a carrier sentence to have a subject. The subject of a sentence is often dropped in Chinese, a pro-drop language. Although such a sentence is perfectly grammatical, it is undesirable as a carrier sentence since it cannot be interpretable in isolation. *We require the root of the carrier sentence to have a noun subject.*[4]

3.2 Target word features

Word Co-occurrence. A good sentence "should present words with which [the target word] typically co-occurs" (Didakowski et al., 2012). We measure co-occurrence with pointwise mutual information (PMI), estimated on the Wiki Set. We compute the "PMI Score" of a sentence by finding the word in the sentence that has the highest PMI with the target word. Table 1 shows the carrier sentence with the maximum PMI score for the target word *xiande* 'appear, seem, look'. The word that yields the highest PMI with the target word is *xiangbi* 'compare', reflecting a typical use of *xiangbi* to introduce a second element ("other organisms") to contrast with the subject ("human"). *We select the carrier sentence with the highest PMI Score with respect to the target word.*

Lexical similarity. A good sentence should "contain words that are lexically-semantically related to [the target word]" (Didakowski et al., 2012). We trained a continuous bag of words (CBOW) model of 400 dimensions and window size 5 with `word2vec` (Mikolov et al., 2013) on the Wiki Set. We computed the "Similarity Score" of a sentence by finding the word in the sentence that has the highest `word2vec` similarity score with the target word[5]. Table 1 shows the carrier sentence with the maximum similarity score for the target word *xiande* 'appear, seem, look'. The word that yields the highest similarity score with the target word is *gengwei* 'even more', a verb that is often used in similar context. *We select the carrier sentence with the highest Similarity Score with respect to the target word.*

4 Experiment set-up

Among the target words in fill-in-the-blank items in the Textbook Set, we selected 100 words — 56 verbs, 31 nouns and 13 adjectives — such

[2] An IP node corresponds to an S or SBAR node in the Penn Treebank.

[3] More precisely, the average target word position is 7.1 out of a ten-word sentence. The optimal word position might differ by language, and deserves further investigation.

[4] That is, the root, typically a verb, a noun or an adjective, must have a child word in the `nsubj` or `nsubjpass` relation.

[5] The target word cannot be repeated in the rest of the sentence.

Method:	`Co-occur`
Related word:	*xiangbi* 相比 'compare' (Highest PMI with target word)
與其他生物 [相比 (*xiangbi*)]，人類的生產過程顯得 (*xiande*) 複雜許多。 When [compared] with other organisms, the human reproductive process appears a lot more complex.	
Method:	`Similar`
Related word:	*gengwei* 更為 'even more' (Highest similarity score with target word)
在政治和宗教的問題上，牛津比劍橋顯得 (*xiande*)[更為 (*gengwei*)] 保守。 On political and religious issues, Oxford appears [even more] conservative than Cambridge.	

Table 1: Carrier sentences selected for the target word *xiande* 'appear, seem, look' by the `Co-occur` method and `Similar` method.

that they are roughly equally spaced in the list of 20,000 most frequent words in the Wiki Set. For each of these 100 words, we retrieve candidate sentences from the Wiki Set that satisfy the constraints imposed by the **Baseline** features. From this pool of candidates, we used three methods to select a carrier sentence. The `Baseline` method randomly selects a sentence from the pool. The `+Similar` method selects the candidate that optimizes the **Lexical Similarity** feature. The `+Co-occur` method selects the candidate that optimizes the **Word co-occurrence** feature. Finally, the `Human` method uses the carrier sentence from the Textbook Set.

We asked two human judges, both native Chinese speakers, to evaluate the four carrier sentences for each of the 100 target words. They assigned two scores to each sentence: the *Sentence Score* (3="Good", 2="Fair", 1="Unacceptable") assesses the extent to which the sentence is grammatical, fluent and fit for pedagogical purpose; the *Word Score*, on the same 3-point scale, indicates how well the sentence succeeds in illustrating a typical usage of the target word. The kappa for these two scores are 0.342 and 0.227, respectively; both are considered a "fair" level of agreement (Landis and Koch, 1977).

Method	**Sentence Score**		**Word Score**	
	Avg	Good	Avg	Good
`Baseline`	2.15	50%	2.60	77%
`+Similar`	2.51	71%	2.67	80%
`+Co-occur`	**2.68**	**79%**	**2.70**	**82%**
`Human`	**2.70**	**81%**	**2.91**	**94%**

Table 2: Percentage of sentences rated "Good" (score 3 out of 3) and scores of carrier sentences generated by the various methods, averaged between the two judges

5 Experimental results

As shown in Table 2, human-authored carrier sentences attracted the highest scores, and have the highest percentage rated "Good" (81.0% and 93.5% for the sentence and word scores). In terms of the Sentence Score, both the `Co-occur` method and `Similar` method[6] outperformed the baseline. For the `Co-occur` method, 79.0% of the sentences were rated "Good". In most cases, the presence of a word of frequent co-occurrence seemed to be a reliable indicator of a high-quality sentence. The Word Score tends to be high across all methods; the baseline features already led to 77.0% of the selected sentences rated "Good". Both the `Co-occur` and `Similar` methods only slightly outperformed the baseline[7], suggesting a relatively high quality of word usage in general in Chinese Wikipedia articles.

6 Conclusions

We have presented a study on automatic selection of carrier sentences for fill-in-the-blank items for Chinese as a foreign language. Our evaluation results show that word co-occurrence and lexical similarity measures can improve the quality of the carrier sentences, over a baseline that considers only sentence complexity, vocabulary difficulty, sentence completeness, and target word position.

Acknowledgments

This work is funded by the Language Fund under Research and Development Projects 2015-2016 of the Standing Committee on Language Education and Research (SCOLAR), Hong Kong SAR.

[6] $p < 0.001$ by McNemar's test for both methods

[7] Not statistically significant, at $p = 0.34$ and $p = 0.54$ by McNemar's test

References

Vít Baisa and Vít Suchomel. 2014. SkELL: Web Interface for English Language Learning. In *Proc. Recent Advances in Slavonic Natural Language Processing*.

Chia-Yin Chen, Hsien-Chin Liou, and Jason S. Chang. 2006. FAST: An Automatic Generation System for Grammar Tests. In *Proc. COLING/ACL Interactive Presentation Sessions*.

Yu-Ta Chen, Yaw-Huei Chen, and Yu-Chih Cheng. 2013. Assessing chinese readability using term frequency and lexical chain. *Computational Linguistics and Chinese Language Processing* 18(2):1–18.

Kevyn Collins-Thompson. 2008. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics* 165(2):97–135.

Jörg Didakowski, Lothar Lemnitzer, and Alexander Geyken. 2012. Automatic Example Sentence Extraction for a Contemporary German Dictionary. In *Proc. EURALEX*.

T. M. Haladyna, S. M. Downing, and M. C. Rodriguez. 2002. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education* 15(3):309–333.

Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, and Chris Welty. 2010. Learning to Predict Readability using Diverse Linguistic Features. In *Proc. 23rd International Conference on Computational Linguistics (COLING)*. pages 546–554.

Adam Kilgarriff, Mils Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proc. EURALEX*.

Susanne Knoop and Sabrina Wilske. 2013. WordGap: Automatic Generation of Gap-Filling Vocabulary Exercises for Mobile Learning. In *Proc. Second Workshop on NLP for Computer-assisted Language Learning, NODALIDA*.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33:159–174.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proc. ACL*.

Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. 2005. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In *Proc. 2nd Workshop on Building Educational Applications Using NLP*. pages 1–8.

Jennifer Lichia Liu. 2004. *Connections I: a Cognitive Approach to Intermediate Chinese*. Indiana University Press, Bloomington, IN.

Jennifer Lichia Liu. 2010. *Encounters I/II: a Cognitive Approach to Advanced Chinese*. Indiana University Press, Bloomington, IN.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing Authentic Web Pages for Language Learners. In *Proc. Fifth Workshop on Innovative Use of Nlp for Building Educational Applications*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. ICLR*.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013. Automatic Selection of Suitable Sentences for Language Learning Exercises. In *Proc. EUROCALL*.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and Machine Learning Approaches for Second Language Sentence-level Readability. In *Proc. 9th Workshop on Innovative Use of NLP for Building Educational Applications*.

Juan Pino and Maxine Eskenazi. 2009. Semi-automatic Generation of Cloze Question Distractors: Effect of Students' L1. In *Proc. SLaTE*.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: a unified framework for predicting text quality. In *Proc. EMNLP*.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners. In *Proc. ACL*.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proc. ACL*.

Simon Smith, P. V. S. Avinesh, and Adam Kilgarriff. 2010. Gap-fill Tests for Language Learners: Corpus-Driven Item Generation. In *Proc. 8th International Conference on Natural Language Processing (ICON)*.

Simon Smith, Adam Kilgarriff, W-L Gong, S. Sommers, and G-Z Wu. 2009. Automatic Cloze Generation for English Proficiency Testing. In *Proc. LTTC Conference*.

Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. 2005. Measuring Non-native Speakers'Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In *Proc. 2nd Workshop on Building Educational Applications using NLP*.

Yao-Ting Sung, Ju-Ling Chen, Ji-Her Cha, Hou-Chiang Tseng, Tao-Hsing Chang, and Kuo-En Chang. 2015. Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods* 47:340–354.

Elena Volodina, Richard Johansson, and Sofie Johansson Kokkinakis. 2012. Semi-automatic Selection of Best Corpus Examples for Swedish: Initial Algorithm Evaluation. In *Proc. Workshop on NLP in Computer-Assisted Language Learning*.

Youmin Wang. 2007. 實用商務漢語課本（漢韓版）準高級篇／高級篇. Commercial Press, Beijing.

Wei Xu. 2012. A Research on Blanked Cloze Exercises in Intermediate TCSL Comprehensive Textbooks Taking Four Textbooks as Examples [in Chinese]. In 第五届北京地区对外汉语教学研究生论坛 *(Proc. 5th Forum of CFL Graduate Students)*. School of Chinese as a Second Language, Peking University, Beijing, China.

Torsten Zesch and Oren Melamud. 2014. Automatic Generation of Challenging Distractors Using Context-Sensitive Inference Rules. In *Proc. Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.