Abstract

This paper reports the first study on automatic generation of distractors for fill-in-the-blank items for learning Chinese vocabulary. We investigate the quality of distractors generated by a number of criteria, including part-of-speech, difficulty level, spelling, word co-occurrence and semantic similarity. Evaluations show that a semantic similarity measure, based on the word2vec model, yields distractors that are significantly more plausible than those generated by baseline methods.

1 Introduction

The fill-in-the-blank item is a common form of exercise in computer-assisted language learning (CALL) systems. Also known as a cloze or gap-fill item, a fill-in-the-blank item is constructed on the basis of a carrier sentence. One word in the sentence — called the target word, or key — is blanked out, and the learner attempts to fill it. The top of Table 1 shows an example carrier sentence whose target word is *tiaojian* ‘condition’.

To enable automatic feedback, a fill-in-the-blank item often specifies choices, including the target word itself and several distractors, as shown at the bottom of Table 1. Distractors need to be carefully chosen: they must be sufficiently plausible, but must not be acceptable answers. Literature in language pedagogy generally recommends the following criteria to authors of fill-in-the-blank items: a distractor should belong to the same word class and same difficult level, and have approximately the same length, as the target word (Heaton, 1989); it should collocate strongly with a word in the sentence (Hoshino, 2013); and it should be semantically related with the target word, ideally a “false synonym” (Goodrich, 1977). An empirical study confirmed that distractors indeed tend to be syntactically and semantically homogenous (Pho et al., 2014).

To automate the time-consuming process of selecting distractors, there has been much interest in developing algorithms that, given a carrier sentence and a target word, can find appropriate distractors. To-date, most research effort on distractor generation for language learning has focused on English.

This paper presents the first attempt to automatically generate distractors in fill-in-the-blank items for learners of Chinese as a foreign language. In Section 2, we review related research areas. In Section 3, we present our datasets. In Section 4, we outline our criteria for distractor generation. In Section 5, we describe the evaluation procedure. In Section 6, we report evaluation results, show-
ing that a semantic similarity measure based on the word2vec model yields distractors that are significantly more plausible than those generated by baseline methods.

2 Previous work

An algorithm for generating distractors must attempt a trade-off between two objectives. One objective is plausibility. Most approaches require the distractor and the target word to have the same part-of-speech (POS) and similar level of difficulty, often approximated by word frequency (Coniam, 1997; Shei, 2001; Brown et al., 2005). They must also be semantically close, which can be quantified with semantic distance in WordNet (Lin et al., 2007; Pino et al., 2008; Chen et al., 2015; Susanti et al., 2015), thesauri (Sumita et al., 2005; Smith et al., 2010), ontologies (Karaninis et al., 2006; Ding and Gu, 2010), or hand-crafted rules (Chen et al., 2006). Another approach generates distractors that are semantically similar to the target word in some sense, but not in the particular sense in the carrier sentence (Zesch and Melamud, 2014). Others directly extract frequent mistakes in learner corpora to serve as distractors (Sakaguchi et al., 2013; Lee et al., 2016). Error-annotated Chinese learner corpora are still not large enough, however, to support broad-coverage distractor generation.

A second, often competing objective is to ensure that the distractor, however plausible, is not an acceptable answer. Most approaches require that the distractor never, or only rarely, collocate with other words in the carrier sentence. Some define collocation as n-grams in a context window centered on the distractor (Liu et al., 2005). Others also consider words elsewhere in the carrier sentence, for example those present in the Word Sketch of the distractor (Smith et al., 2010) or those that are grammatically related to the distractor in dependencies (Sakaguchi et al., 2013). Still others restrict potential distractors to antonyms of the target word, words with the same hypernym, and synonym of synonyms in WordNet (Knoop and Wilske, 2013).

To the best of our knowledge, there is not yet any reported attempt to generate distractors for learning Chinese vocabulary. The only previous work on Chinese distractor generation was designed for testing knowledge in the aviation domain, and leveraged a domain-specific ontology (Ding and Gu, 2010).

3 Data

To facilitate our study, we compiled two datasets:

**Textbook Corpus** We collected 299 fill-in-the-blank items, each with a target word and two to three distractors, from three Chinese textbooks (Liu, 2004, 2010; Wang, 2007). An analysis on this corpus confirms many of the criteria proposed in the literature: in 63% of the items, all distractors have the same POS as the target word; and in 45% of the items, at least one distractor shares a common character with the target word.

**Wiki Corpus** We extracted 14 million sentences from Chinese Wikipedia for calculating word frequency, similarity and co-occurrence statistics in the Candidate Generation step. We then performed word segmentation, POS tagging and dependency analysis on a subset of 5.5 million sentences with the Stanford Chinese parser (Levy and Manning, 2003) for use in the Candidate Filtering step.

4 Approach

We follow a two-step process where the first step, Candidate Generation, optimizes distractor plausibility; and the second step, Candidate Filtering, aims to filter out distractor candidates that are acceptable answers.

4.1 Candidate Generation

We implemented the following criteria for generating a ranked list of distractor candidates:

**Baseline** (Baseline) The baseline re-implements the criteria proposed by Coniam (1997): the distractor must have the same POS and the similar difficulty level as the target word. We extract all words in the Wiki corpus with the same POS, and then rank them by the proximity of their word frequency and that of the target word. In Table 1, for example, pin-dao ‘channel’ was chosen because, among all nouns, its word frequency is closest to that of the target word tiaojian.

**Spelling similarity** (+Spell) Many Chinese words contain multiple characters; two words that have one or more characters
in common may be easily confusable for learners. This method requires the candidate
to share at least one common character with
the target word. In our running example in
Table 1, tiaoyue ‘agreement’ was chosen because,
among all words that contain the character tiao or jian (which combine to
form the target word tiaojian), it has the most
similar word frequency.

**Word co-occurrence** (+Co-occur) A distractor
that often co-occurs with the target word may
be easily confusable for learners. We ranked
the candidate distractors according to their
pointwise mutual information (PMI) score
with the target word, as estimated on the Wiki
corpus. In our running example in Table 1,
hanshu ‘function’ was chosen because of its
frequent co-occurrence with tiaojian ‘condition’.

**Word similarity** (+Similar) Words that are semantically
close to the target word tend to be plausible candidates. We ranked
candidate distractors according to their similarity
score with the target word. We obtained these scores by training a word2vec model (Mikolov et al., 2013) on the Wiki
corpus.² We opted for word2vec over thesauri or Chinese lexical databases such as HowNet because of its broader coverage. In the example in Table 1, yinsu ‘factor’ was chosen in-
stead since it never served as the subject of tiaojian.

### 4.1 Candidate Generation

A distractor is called “reliable” if it yields an incor-
correct sentence. This step aims to remove those can-
didates that are also acceptable answers, leaving
only the reliable distractors. We do so by examin-
ing whether the distractor can collocate with words
in the rest of the carrier sentence. The system ex-
amines the candidates in the ranked list produced
by the Candidate Generation step (Section 4.1),
and removes candidates that are rejected by both
filters below:

- **Trigram** The word trigram, formed by the distrac-
tor, the previous word and the following word
in the carrier sentence, must not appear in the

²We trained a bag-of-words (CBOW) model of 400 di-
mensions and window size 5 with word2vec.

**Figure 1:** In the Candidate Filtering step (Section 4.2), candidate distractors whose dependency relations are attested in the corpus are rejected. To determine whether yinsu can serve as a distractor in the carrier sentence in Table 1, the system determines whether the dependency relations nmod(yinsu, nali) or nsubj(hao, yinsu) is attested in a large corpus of Chinese texts.

dependency relations are attested in the corpus are rejected. To determine whether yinsu can serve as a distractor in the carrier sentence in Table 1, the system determines whether the dependency relations nmod(yinsu, nali) or nsubj(hao, yinsu) is attested in a large corpus of Chinese texts.

Wiki corpus. In the example in Figure 1, the trigram “de yinsu bu” must not be attested.

**Dependency** The Trigram filter alone might be
too strict, since words that are grammatically
related to the distractor may be further away.
Among dependency relations in the parse tree
of the carrier sentence, we extract all those
with the distractor as head or child, and re-
quire that these relation must not be attested in the Wiki
corpus. This filter is similar to the approach by Smith et al. (2010), but instead of the grammatical relations in Word
Sketches, we consider all dependency rela-
tions. In our running example in Table 1, the
candidate 情况 qingkuang ‘situation’ was re-
jected because it is attested to serve as the
subject of hao ‘good’. The next distractor in the
ranked list, yinsu ‘factor’, was chosen in-
stead since it never served as the subject of
hao ‘good’, and was never modified by the
noun nali ‘there’.

### 5 Evaluation

#### 5.1 Test data

According to Da (2007), basic ability in Chi-
inese news reading require a vocabulary of around
20,000 words. Among the target words in the Text-
book Corpus, we selected 37 nouns and verbs such
that they were roughly equally spaced among the
20,000 most frequent words in the Wiki Corpus.

For each of these 37 words, we generated dis-
tractors using each of the four criteria in Sec-
2010), but instead of the grammatical relations in Word
Sketches, we consider all dependency rela-
tions. In our running example in Table 1, the
candidate 情况 qingkuang ‘situation’ was re-
jected because it is attested to serve as the
subject of hao ‘good’. The next distractor in the
ranked list, yinsu ‘factor’, was chosen in-
stead since it never served as the subject of
hao ‘good’, and was never modified by the
noun nali ‘there’.

### 5 Evaluation

#### 5.1 Test data

According to Da (2007), basic ability in Chi-
inese news reading require a vocabulary of around
20,000 words. Among the target words in the Text-
book Corpus, we selected 37 nouns and verbs such
that they were roughly equally spaced among the
20,000 most frequent words in the Wiki Corpus.

For each of these 37 words, we generated dis-
tractors using each of the four criteria in Sec-
2010), but instead of the grammatical relations in Word
Sketches, we consider all dependency rela-
tions. In our running example in Table 1, the
candidate 情况 qingkuang ‘situation’ was re-
jected because it is attested to serve as the
subject of hao ‘good’. The next distractor in the
ranked list, yinsu ‘factor’, was chosen in-
stead since it never served as the subject of
hao ‘good’, and was never modified by the
noun nali ‘there’.

### 5 Evaluation

#### 5.1 Test data

According to Da (2007), basic ability in Chi-
inese news reading require a vocabulary of around
20,000 words. Among the target words in the Text-
book Corpus, we selected 37 nouns and verbs such
that they were roughly equally spaced among the
20,000 most frequent words in the Wiki Corpus.

For each of these 37 words, we generated dis-
tractors using each of the four criteria in Sec-
2010), but instead of the grammatical relations in Word
Sketches, we consider all dependency rela-
tions. In our running example in Table 1, the
candidate 情况 qingkuang ‘situation’ was re-
jected because it is attested to serve as the
subject of hao ‘good’. The next distractor in the
ranked list, yinsu ‘factor’, was chosen in-
stead since it never served as the subject of
hao ‘good’, and was never modified by the
noun nali ‘there’.

### 5 Evaluation

#### 5.1 Test data

According to Da (2007), basic ability in Chi-
inese news reading require a vocabulary of around
20,000 words. Among the target words in the Text-
book Corpus, we selected 37 nouns and verbs such
that they were roughly equally spaced among the
20,000 most frequent words in the Wiki Corpus.

For each of these 37 words, we generated dis-
tractors using each of the four criteria in Sec-
2010), but instead of the grammatical relations in Word
Sketches, we consider all dependency rela-
tions. In our running example in Table 1, the
candidate 情况 qingkuang ‘situation’ was re-
jected because it is attested to serve as the
subject of hao ‘good’. The next distractor in the
ranked list, yinsu ‘factor’, was chosen in-
stead since it never served as the subject of
hao ‘good’, and was never modified by the
noun nali ‘there’.

### 5 Evaluation

#### 5.1 Test data

According to Da (2007), basic ability in Chi-
inese news reading require a vocabulary of around
20,000 words. Among the target words in the Text-
book Corpus, we selected 37 nouns and verbs such
that they were roughly equally spaced among the
20,000 most frequent words in the Wiki Corpus.

For each of these 37 words, we generated dis-
tractors using each of the four criteria in Sec-
2010), but instead of the grammatical relations in Word
Sketches, we consider all dependency rela-
tions. In our running example in Table 1, the
candidate 情况 qingkuang ‘situation’ was re-
jected because it is attested to serve as the
subject of hao ‘good’. The next distractor in the
ranked list, yinsu ‘factor’, was chosen in-
stead since it never served as the subject of
hao ‘good’, and was never modified by the
noun nali ‘there’.

### 5 Evaluation

#### 5.1 Test data

According to Da (2007), basic ability in Chi-
inese news reading require a vocabulary of around
20,000 words. Among the target words in the Text-
book Corpus, we selected 37 nouns and verbs such
that they were roughly equally spaced among the
20,000 most frequent words in the Wiki Corpus.

For each of these 37 words, we generated dis-
tractors using each of the four criteria in Sec-
2010), but instead of the grammatical relations in Word
Sketches, we consider all dependency rela-
tions. In our running example in Table 1, the
candidate 情况 qingkuang ‘situation’ was re-
jected because it is attested to serve as the
subject of hao ‘good’. The next distractor in the
ranked list, yinsu ‘factor’, was chosen in-
stead since it never served as the subject of
hao ‘good’, and was never modified by the
noun nali ‘there’.

### 5 Evaluation

#### 5.1 Test data

According to Da (2007), basic ability in Chi-
inese news reading require a vocabulary of around
20,000 words. Among the target words in the Text-
book Corpus, we selected 37 nouns and verbs such
that they were roughly equally spaced among the
20,000 most frequent words in the Wiki Corpus.

For each of these 37 words, we generated dis-
tractors using each of the four criteria in Sec-
2010), but instead of the grammatical relations in Word
Sketches, we consider all dependency rela-
tions. In our running example in Table 1, the
candidate 情况 qingkuang ‘situation’ was re-
jected because it is attested to serve as the
subject of hao ‘good’. The next distractor in the
ranked list, yinsu ‘factor’, was chosen in-
stead since it never served as the subject of
hao ‘good’, and was never modified by the
noun nali ‘there’.
Table 2: Reliability of the various distractor generation methods.

| Method       | Reliability |
|--------------|-------------|
| Baseline     | 100%        |
| +Co-occur    | 98.6%       |
| +Spell       | 93.2%       |
| +Similar     | 93.2%       |
| Human        | 100%        |

Table 3: Average scores, out of a 3-point scale (see Section 5.2), of distractors generated by the various methods in the human evaluation.

|h | Average score | Plausible or somewhat plausible |
|---|---------------|---------------------------------|
| Baseline | 1.06 | 5.2% |
| +Co-occur | 1.27 | 8.6% |
| +Spell | 1.66 | 39.7% |
| +Similar | 1.76 | 46.6% |
| Human | 1.68 | 53.4% |

6.2 Plausibility

Table 3 shows the results on plausibility. Both the +Similar method and the +Spell method outperformed the baseline, both in terms of the average score and the proportion of distractors considered at least somewhat plausible.

5.2 Human annotation

We asked two human judges, both native Chinese speakers, to annotate these choices, without revealing the target word. For each choice in the item, the judges decided whether it was correct or incorrect; they may identify zero, one or multiple correct answers. For an incorrect answer, they further assessed its plausibility as a distractor on a three-point scale: “Plausible” (3), “Somewhat plausible” (2), or “Obviously wrong” (1).

The kappa for the human annotation is 0.529, which is considered a “moderate” level of agreement (Landis and Koch, 1977). As a annotation quality check, we found that overall, in 6.8% of the times, a judge labels the target word as a distractor.

7 Conclusions

We presented the first study on automatic generation of distractors for fill-in-the-blank items for learning Chinese. Evaluations showed that a semantic similarity measure, based on the word2vec model, offers a significant improvement over a baseline that considers only part-of-speech and word frequency, and achieves competitive plausibility in comparison to human-authored items.

---

3Except that in 5 items, the +Co-occur and +Similar methods generated the same distractor; in another item, Baseline and +Co-occur generated the same distractor.

4p < 0.001, by McNemar’s test.

5p < 0.021 by McNemar’s test.

6Since we randomly selected one distractor out of three in the Textbook Corpus, the Human score reflects the average plausibility of the human-authored distractors, rather than the best one.
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.

References

Jonathan C. Brown, Gwen A. Frishkoff, and Maxine Eskenazi. 2005. Automatic Question Generation for Vocabulary Assessment. In Proc. HLT-EMNLP.

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.

Tao Chen, Naijia Zheng, Yue Zhao, Muthu Kumar Chandrasekaran, and Min-Yen Kan. 2015. Interactive Second Language Learning from News Websites. In Proc. 2nd Workshop on Natural Language Processing Techniques for Educational Applications.

David Coniam. 1997. A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests. CALICO Journal 14(2-4):15–33.

Jun Da. 2007. Reading News for Information: How Much Vocabulary a CFL Learner Should Know. In Andreas Guder, Jiang Xin, and Yexin Wan, editors, The Cognition, Learning and Teaching of Chinese Characters. Beijing Language and Culture University Press, pages 251–277.

Xiangmin Ding and Hongbin Gu. 2010. Automatic Generation Technology of Chinese Multiple-choice Items based on Ontology [in Chinese]. Computer Engineering and Design 31(6):1397–1400.

Hubbard C. Goodrich. 1977. Distractor Efficiency in Foreign Language Testing. TESOL Quarterly 11(1):69–78.

J. B. Heaton. 1989. Writing English Language Tests. Longman.

Yuko Hoshino. 2013. Relationship between Types of Distractor and Difficulty of Multiple-Choice Vocabulary Tests in Sentential Context. Language Testing in Asia 3(16).

Nikiforos Karamanis, Le An Ha, and Ruslan Mitkov. 2006. Generating Multiple-Choice Test Items from Medical Text: A Pilot Study. In Proc. 4th International Natural Language Generation Conference.

S. Knoop and S. 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.

John Lee, Donald Sturgeon, and Mengqi Luo. 2016. A CALL System for Learning Preposition Usage. In Proc. ACL.

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

Y.-C. Lin, L.-C. Sung, and M.-C. Chen. 2007. An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding. In Proc. Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, 15th International Conference on Computers in Education.

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.

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

Van-Minh Pho, Thibault André, Anne-Laure Ligozat, B. Grau, G. Illouz, and Thomas François. 2014. Multiple Choice Question Corpus Analysis for Distractor Characterization. In Proc. LREC.

Juan Pino, M. Helman, and Maxine Eskenazi. 2008. A Selection Strategy to Improve Cloze Question Quality. In Proc. Workshop on Intelligent Tutoring Systems for Ill-Defined Domains, 9th International Conference on Intelligent Tutoring Systems.

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

Chi-Chiang Shei. 2001. FollowYou!: An Automatic Language Lesson Generation System. Computer Assisted Language Learning 14(2):129–144.

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

Yuni Susanti, Ryu Iida, and Takenobu Tokunaga. 2015. Automatic Generation of English Vocabulary Tests. In Proc. 7th International Conference on Computer Supported Education (CSEDU).

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

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).