Difficulty-aware Distractor Generation for Gap-Fill Items

Chak Yan Yeung, John Lee, Benjamin Tsou
Department of Linguistics and Translation
City University of Hong Kong
cyyeung91@gmail.com, {jsylee,rlbtsou}@cityu.edu.hk

Abstract

Many computer-assisted language learning (CALL) systems offer gap-fill items, often with multiple choices in order to facilitate immediate feedback. Automatic distractor generation can therefore be helpful in providing the multiple choices. While existing algorithms focus on proposing the most plausible distractors, many realistic scenarios make use of distractors at a variety of difficulty levels. This paper evaluates the use of a neural language model to rank distractors in terms of difficulty. Experiments show that BERT outperforms semantic similarity measures, in terms of both correlation to human judgment and classification accuracy of distractor plausibility.

1 Introduction

Many computer-assisted language learning (CALL) systems offer gap-fill items, also known as cloze or fill-in-the-blank items. A gap-fill item consists of a carrier sentence in which one word — called the key, or *target word* — is blanked out. Table 1 shows an example carrier sentence whose target word is 'served'.

To enable automatic feedback, multiple choices are sometimes provided for filling the gap. As shown in Table 1, these choices include the target word itself and several distractors. Judicious selection of distractors is important for generating effective items. A distractor should produce a sentence that seems plausible, yet unacceptable. Language pedagogy literature generally recommends that the target word and distractors belong to the same word class (Heaton, 1989), and be semantically related, ideally "false synonyms" (Goodrich, 1977). An empirical study confirmed that distractors indeed tend to be syntactically and semantically homogenous (Pho et al., 2014).

To reduce the manual effort and time needed for selecting distractors, there has been much interest

He ____ as class representative for two years.	
(a) served	[target word]
(b) acted	[distractor]
(c) brought	[distractor]

Table 1: A multiple-choice gap-fill item consists of a carrier sentence with a blank, and choices for filling the blank. In this example, the choices include two distractors and the target word (correct answer).

in developing algorithms for automatic distractor generation. Existing algorithms typically take a two-step approach (Jiang and Lee, 2017; Susanti et al., 2018). The first step generates distractor candidates, typically in a list ranked according to measures of semantic similarity and collocation strength. The second step removes candidates that are acceptable answers, using n-gram and collocation frequency or other criteria.

Evaluations on distractor generation tend to be limited to the highest-ranked distractors, for example the top-ranked or top three candidates only (Jiang and Lee, 2017; Susanti et al., 2018). Many practical scenarios, however, require not only the most challenging distractors, but distractors across the spectrum of plausibility. When authoring test items, for example, it can be useful to generate items at various difficulty levels for comprehensive assessment. In a CALL system, it can be effective to personalize distractor difficulty according to the user's language proficiency.

It is informative, then, to evaluate distractor algorithms on their ability to generate distractors at different levels of plausibility. We therefore propose to investigate the correlation between the estimated ranking of distractors and human judgment on plausibility. This research direction has indeed been taken up in item generation in the nat-

ural sciences (Liang et al., 2018; Gao et al., 2019). However, to the best of our knowledge, it has not yet been attempted for gap-fill items for language learning.

Language models provide a natural framework for this task by predicting how likely a word appears in a gap within the sentence. This paper is the first attempt to apply BERT (Devlin et al., 2019), a state-of-the-art model trained on the masked language modeling objective, on the task of distractor ranking. Experimental results show that BERT outperforms semantic similarity measures, in terms of both correlation to human judgment and classification accuracy of distractor plausibility.

The rest of the paper is organized as follows. In Section 2, we review related research areas. In Section 3, we outline our approach for distractor generation. In Section 4, we report experimental results on ranking distractors for gap-fill items for learning Chinese as a foreign language. Finally, we conclude in Section 5.

2 Previous work

For the target word in a carrier sentence, a distractor generation algorithm aims to optimize two objectives: the distractor should look plausible for filling in the gap; but it should also *not* produce an acceptable sentence. Reflecting the twin goals, most existing algorithms perform two tasks (Jiang and Lee, 2017; Susanti et al., 2018). The first, Candidate Generation, identifies all possible distractor candidates. The second, Candidate Filtering, seeks to remove those candidates that are also acceptable answers, leaving only the distractors that are "reliable", i.e., those that yield an incorrect sentence.

Section 2.1 reviews existing approaches for the Candidate Generation task, which is the research focus of this paper. Section 2.2 then surveys related tasks in computer-assisted learning that have adopted evaluation on candidate ranking.

2.1 Candidate Generation

In most approaches, a distractor needs to have the same part-of-speech (POS) as the target word (Coniam, 1997). In addition, a number of features have been explored for ranking the candidates:

Word co-occurrence. Since a distractor should collocate strongly with a word in the carrier sentence (Hoshino, 2013), candidates are evaluated according to their co-occurrence frequencies with other words in the sentence. Various definitions of co-occurrence have been used, including bigram counts (Susanti et al., 2018) and, more generally, n-grams in a context window centered on the distractor (Liu et al., 2005); dependency relations (Sakaguchi et al., 2013); grammatical relations in a Word Sketch (Smith et al., 2010); as well as pointwise mutual information (PMI) (Jiang and Lee, 2017).

Learner error. Typical or frequent learner mistakes can be effective as distractors. When generating gap-fill items for prepositions, distractors based on learner errors were indeed found to be more challenging than those selected according to word co-occurrence (Lee et al., 2016). Distractor candidates have been mined from learner corpora and further selected with a discriminative model (Sakaguchi et al., 2013).

Semantic similarity. The distractor should be semantically close to the target word. Similarity can be quantified by semantic distance in WordNet (Pino et al., 2008; Chen et al., 2015), thesauri (Sumita et al., 2005; Smith et al., 2010), ontologies (Karamanis et al., 2006), hand-crafted rules (Chen et al., 2006), and word embeddings (Jiang and Lee, 2017; Susanti et al., 2018). Other approaches have also explored synonym of synonyms (Knoop and Wilske, 2013); and words 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).

A study on gap-fill items for learning Chinese as a foreign language compared the quality of distractors generated by a number of criteria, including spelling, word co-occurrence and semantic similarity (Jiang and Lee, 2017). Experimental results show that a semantic similarity measure, based on the `word2vec` model (Mikolov et al., 2013), yields distractors that are significantly more plausible than those generated by spelling similarity, and by word co-occurrence strength as estimated by PMI.

2.2 Evaluation on candidate ranking

Although the output of most distractor generation algorithms is a ranked list, most previous studies on distractor quality in gap-fill items have limited their attention to the top-ranked distractors. To the best of our knowledge, quantitative evaluations on *ranking* quality have been reported for item gen-

eration in the natural sciences, but not yet for language learning; furthermore, the focus has been on question-answering items, rather than gap-fill items.

Given a dataset of distractors and non-distractors, Liang et al. (2018) trained ranking models to rank the distractors higher. Experimental results suggested that random forest and Lambda MART performed best. However, the evaluation was restricted to pairwise prediction of distractor difficulty.

Gao et al. (2019) addressed the task of generating questions from a paragraph. Using bidirectional LSTMs, their system classified questions as either "easy" or "difficult". While their evaluation methodology was similar to the one advocated by this paper, it is applied to question generation rather than distractor generation.

3 Approach

Following most previous approaches, we adopt a two-step process for distractor generation: the first step generates distractor candidates, and the second step filters out candidates that constitute acceptable answers. Our research focus is on the first step, to which we introduce a re-ranking process with a neural language model.

3.1 Baseline

Our baseline uses semantic similarity measures, which have been reported in previous research to yield the best performance in identifying plausible distractor candidates (Section 2.1).

Word embeddings have been shown to be effective in measuring word similarity and relatedness in a large range of NLP tasks, including distractor generation (Jiang and Lee, 2017). We used word embeddings trained by Skipgram with negative sampling on Baidu Encyclopedia (Li et al., 2018). Specifically, we calculated cosine similarity between the word embeddings of the distractor candidate and the target word, and obtained candidates with the highest scores.

3.2 Re-ranking

The appropriateness of a distractor may depend not only on the target word but also on the context of the carrier sentence. Consider the word *served* as the target word. In the context of food being served at a restaurant, the word *brought* may be a plausible distractor since it is semantically close to the target word. However, in the context of serving in a position, the word *acted* would be a more plausible distractor, for example for the carrier sentences in Table 1. Hence, we propose to re-rank the distractors in the baseline list (Section 3.1) with a language model.

BERT (Devlin et al., 2019) is a state-of-the-art neural language model based on the Transformer architecture (Vaswani et al., 2017). The model is bi-directional, i.e., trained to predict the identity of a masked word based on both the words that precede and follow it. It has been shown to be effective in a variety of natural language processing tasks. This paper is the first to apply it to distractor generation.

Using its PyTorch implementation[1], we masked the target word in each carrier sentence and harvested the words most highly ranked by BERT for the masked position. We then re-ranked the candidates in the baseline list according to their relative ranking in BERT.

An alternative to re-ranking would be to directly use the ranked list from BERT. We did not adopt this approach because the list can include distractor candidates that are not semantically similar to the target word.

4 Experiments

We first present our dataset (Section 4.1), and then analyze experimental results (Section 4.2).

4.1 Data

We derived our evaluation data from the dataset compiled by Jiang and Lee (2017), which consists of 37 carrier sentences taken from a number of textbooks for Chinese as a foreign language. The target words include 37 distinct nouns and verbs.

To construct a pool of distractor candidates for the target word in each sentence, we intersected these two sets: the 20 words that are most similar to the target word according to the baseline algorithm (Section 3.1); and the 50 most likely words for the masked position according to BERT (Section 3.2).

Out of this pool, we randomly selected a total of 172 distractor candidates for human annotation. Five human raters, all native speakers of Chinese, independently annotated each candidate according to the scheme used in Jiang and Lee (2017).

[1] https://pypi.org/project/pytorch-pretrained-bert/

Method	Correlation		Classification accuracy
	Pearson's r	Spearman's rho	
Baseline	-0.233	-0.263	42.52%
Re-ranking	**-0.455**	**-0.487**	**60.63%**

Table 2: Evaluation of the baseline (Section 3.1) and re-ranking (Section 3.2) methods on correlation to human judgment on plausibility (left); and on classification of distractor plausibility (right).

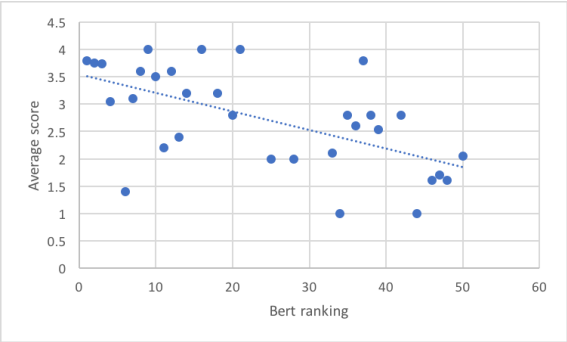

Figure 1: Correlation between human scores on the plausibility of the distractor candidates, and their ranking in BERT (Section 3.2).

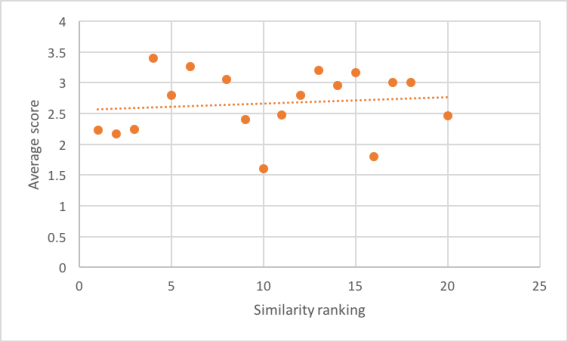

Figure 2: Correlation between human scores on the plausibility of the distractor candidates, and their semantic similarity ranking (Section 3.1).

Each distractor candidate can be rated as "Obviously wrong", "Somewhat plausible", "Plausible", or "Correct" (and hence unacceptable as a distractor). The pairwise Kappa for the five human raters was 0.420, which is considered a "Moderate" level of agreement (Landis and Koch, 1977).

4.2 Results

Correlation with human judgment. Table 2 shows the level of correlation between the automatically produced distractor ranking and the average human rating.[2] A larger negative coefficient indicates a higher degree of correlation, since distractors with higher plausibility scores should have a smaller rank number. BERT achieved a Pearson correlation coefficient of -0.455; visualized in Figure 1, the correlation is statistically significant ($p = 0.002$). In contrast, the coefficient for the semantic similarity baseline was only -0.233; visualized in Figure 2, the correlation is not statistically significant ($p = 0.123$). The trend was similar with Spearman's rho, for which BERT achieved a coefficient of -0.487, which is statistically significant ($p = 0.0007$). The baseline obtained a coefficient of -0.263, which is not statistically significant ($p = 0.081$).

Generation accuracy. We generated distractors by setting thresholds in the re-ranked list for different plausibility levels. We tuned the thresholds by leave-one-out cross-validation to optimize accuracy in classifying a candidate as "Correct", "Plausible"[3], or "Less Plausible". The gold label is the majority label out of the five raters. As shown in Table 2, BERT achieved 60.63% classification accuracy, outperforming the similarity baseline, which achieved 42.52% by always predicting the majority label "Less plausible". For our dataset, the optimized thresholds for BERT were to classify candidates ranked 1 to 10 as "Correct"; those ranked 11 to 39 as "Plausible"; and the rest as "Less Plausible".

5 Conclusions

To support automatic generation of gap-fill items with distractors at a variety of plausibility levels, we have introduced distractor ranking as a new evaluation framework for distractor generation. This study is the first to apply BERT to the task of distractor ranking. Experimental results show that it outperforms semantic similarity measures in terms of both correlation to human judgment on distractor plausibility, and classification accuracy of distractor plausibility.

[2]We assigned score 1 to the "Obviously wrong" candidates, score 2 to "Somewhat plausible", score 3 to "Plausible" and score 4 to "Correct" distractors.

[3]The "Plausible" and "Somewhat Plausible" labels in the human annotation were collapsed into the "Plausible" label.

Acknowledgment

This work was supported by the Innovation and Technology Fund (Ref: ITS/389/17) of the Innovation and Technology Commission, the Government of the Hong Kong Special Administrative Region. We thank Ka Po Chow for his assistance. The first author completed this research at City University of Hong Kong and now works at Google.

References

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*.

Yifan Gao, Lidong Bing, Wang Chen, Michael R. Lyu, and Irwin King. 2019. Difficulty Controllable Generation of Reading Comprehension Questions. In *Proc. 28th International Joint Conference on Artificial Intelligence (IJCAI)*.

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).

Shu Jiang and John Lee. 2017. Distractor Generation for Chinese Fill-in-the-blank Items. In *Proc. 12th Workshop on Innovative Use of NLP for Building Educational Applications*, page 143–148.

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*.

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.

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

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyang Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations. *arXiv preprint arXiv:1805.06504*.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. 2018. Distractor Generation for Multiple Choice Questions using Learning to Rank. In *Proc. 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 284–290.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. International Conference on Learning Representations (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. Heilman, 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*.

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)*.

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, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2018. Automatic Distractor Generation for Multiple-choice English Vocabulary Questions. *Research and Practice in Technology Enhanced Learning*, 13(15).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems*, pages 6000–6010.

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)*.