Automatic Generation of Cloze Items for Prepositions

John Lee, Stephanie Seneff

Spoken Language Systems
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139 U.S.A.
{jsylee,seneff}@csail.mit.edu

Abstract

Fill-in-the-blank questions, or cloze items, are commonly used in language learning applications. The benefits of personalized items, tailored to the user's interest and proficiency, have motivated research on automatic generation of cloze items. This paper is concerned with generating cloze items for prepositions, whose usage often poses problems for non-native speakers of English.

The quality of a cloze item depends on the choice of distractors. We propose two methods, based on collocations and on non-native English corpora, to generate distractors for prepositions. Both methods are found to be more successful in attracting users than a baseline that relies only on word frequency, a common criterion in past research.

Index Terms: computer-assisted language learning, natural language generation

1. Introduction

Due to their ability to provide automatic and objective feedback, multiple choice questions are commonly used in education applications. One type that is especially popular in language learning and assessment is fill-in-the-blank questions, or *cloze items*, where one or more words is removed from a sentence, and a number of candidate words are offered to the user for filling in the gap. An example is shown in Figure 1.

As a language learning tool, cloze tests can be enhanced by using up-to-date, authentic text on topics in which the student takes an interest. Such personalization can "provide motivation, generate enthusiasm in learning, encourage learner autonomy, foster learner strategy and help develop students' reading skills as well as enhance their cultural understanding" [2].

It is clearly not practical to manually design tailor-made cloze tests for every student. This bottleneck has motivated research on automatic generation of cloze items.

2. Problem Definition

Broadly speaking, it takes the following steps to produce a cloze item from a *source corpus*:

1. Determine the *key*, i.e., the word to be removed from a sentence.
2. Select a *seed sentence* from the source corpus.
3. Generate *distractors*, i.e., incorrect choices, for the key.

Past research has focused on cloze items whose keys are of an *open-class* part-of-speech (POS), e.g., nouns, verbs, or adjectives. Words that occur relatively infrequently are selected as keys, with the intention of improving the vocabulary level of the user. The cloze item in Figure 1 is such an example.

While vocabulary build-up is essential, mastering the usage of function words is also important in language learning. Misuse of prepositions, for example, turns out to be a frequent type of error for Japanese speakers, according to the Japanese Learners of English (JLE) corpus, which consists of transcripts of spoken English [3]. Cloze items on prepositions, such as the one shown in Figure 2, can provide training that specifically targets this type of error. This paper is concerned with the automatic generation of such items.

Prepositions, as a closed-class POS, present some new challenges in cloze item generation. First, insertion and deletion of prepositions are common errors, whereas errors in open-class POS are predominantly substitutions. Secondly, the set of prepositions is much smaller than the set of their open-class counterparts. As a result, most prepositions are already familiar to the user, making it more difficult to select good distractors. To address these challenges, we propose two novel techniques for distractor generation.

Figure 1: *An example cloze item taken from [1].*

The child's misery would move even the most ____ heart.
(a) torpid (b) invidious (c) stolid (d) obdurate

Figure 2: *An example cloze item on prepositions, generated from the seed sentence "If you don't have anything planned for this evening, let's go to a movie". The key is "to". Distractor (b) is produced by the baseline method in §4.2, distractor (c) by the collocation method in §4.3, and distractor (d) by the non-native method in §4.4.*

If you don't have anything planned for this evening,
let's go ____ a movie.
(a) to (b) of (c) on (d) *null*

3. Related Work

Past research has addressed both key and distractor selection for open-class POS. The key is often chosen according to word frequency [2, 4], so as to match the user's vocabulary level. Machine learning methods are applied in [5] to determine the best key, using cloze items in a standard language test as training material.

3.1. Distractor Generation

The focus of this paper is on distractor generation. As is widely observed, a good distractor must satisfy two requirements. First and foremost, it must result in an incorrect sentence. Secondly, it must be similar enough to the key to be a viable alternative.

To secure the first requirement, the distractor must yield a sentence with zero hits on the web in [6]; in [7], it must produce a rare collocation with other important words in the sentence.

As for the second, various criteria have been proposed: matching patterns hand-crafted by experts [8]; similarity in meaning to the key, with respect to a thesaurus [6] or to an ontology in a narrow domain [9]. However, the most widely used criterion, again, is similarity in word frequency to the key [1, 2].

3.2. Evaluation

Mirroring the two requirements for distractors, our two main evaluation metrics are *usability* and *difficulty* of the cloze item.

3.2.1. Usability

A "usable" item has been defined in different ways, ranging from the simple requirement that only one choice is correct [7], to expert judgments [8]. Others take into account the time needed for manual post-editing [9], in relation to designing the item from scratch. We adopt the simple requirement as in [7].

3.2.2. Difficulty

Cloze tests have been used both as a proficiency assessment tool [1, 6] and as a language learning tool [2]. For assessment purposes, the ability of the cloze test to discriminate between more advanced students and less advanced ones is important. This is expressed in two dimensions [4, 10]: First, *item difficulty* (or *facility index*), i.e., the distractor should be neither too obviously wrong nor too tricky. Second, *effectiveness* (or *discrimination index*), i.e., it should attract only the less proficient students.

For language learning applications, the discriminative power of a cloze test is not as important as its ability to cause users to make mistakes. An easy cloze test, on which the user scores perfectly, would not be very educational; arguably, the user learns most when his/her mistake is corrected. This paper will emphasize the generation of difficult cloze items.

4. Approach

Our input is a sentence from the source corpus and its key (a preposition). The output is a distractor, which, for our purposes, is ideally the one that is most likely to attract the user (cf. §3.2).

4.1. Context Representation

An important question is how to represent the *context* of the preposition in the sentence. The granularity of the representation reflects a trade-off similar to precision/recall.

Suppose one requires matching a rather large window of words centered on the preposition. With this fine-grained representation, new sentences are unlikely to match any sentences in the training set, and few cloze items can be generated. At another extreme, suppose one ignores the context, and determines the distractor solely on the basis of its frequency count. This coarse representation can produce a cloze item out of any sentence with a preposition, but it risks generating a less viable, and hence less difficult, distractor.

We now give a brief overview of the syntactic functions of prepositions [11] in order to motivate our context representation. A preposition can be a particle in a phrasal or prepositional verb; more frequently, however, it forms a prepositional phrase (PP) with a complement, typically a noun. The PP can serve as an adverbial, a post-modifier of a noun phrase, or the complementation of a verb or an adjective.

No attempt is made to distinguish these different functions. The context of a preposition is represented by the triplet $\langle A, p, B\rangle$, where A and B, possibly empty, are heads of the noun or verb phrases that are associated with the preposition p in one of its syntactic functions described above. From the sentence "*Let's go to a movie*", for example, the triplet $\langle$*go*,*to*,*movie*$\rangle$ is extracted.

Our task is to learn a mapping from such a triplet to $\overline{p}$, the distractor which the user is most likely to confuse with p:

$$\langle A, p, B\rangle \longmapsto \overline{p}$$

Either p or $\overline{p}$ can be an empty string, in which case it is written as *null*. If p is *null*, then A and B are the head nouns or verbs that are to be erroneously associated with $\overline{p}$. For example, the sentence "*So we decided to take the kitty* $\times$*to home*" is represented as $\langle$*take,null,home*$\rangle$, with "to" as $\overline{p}$.

Thus, this mapping is sufficient to represent substitution, insertion and deletion errors. We now describe three different ways to learn this mapping: first a baseline, then two novel methods that leverage the context of the preposition.

4.2. Baseline: Using frequencies

The baseline considers only word frequency, a criterion commonly used in cloze item generation for open-class POS. Given $\langle A, p, B\rangle$, it ignores A and B, and simply returns the $\overline{p}$ whose frequency count in a large English corpus is closest to that of p. According to Table 1, the frequency of "to" is closest to that of "of"; when the key is "to", as in the cloze item in Figure 2, the baseline distractor is "of". When p is *null*, the baseline method stochastically generates a random preposition according to the probability distribution observed in the English corpus.

Table 1: *Preposition frequencies in a corpus of 10-million sentences from the New York Times.*

Prep.	Count	Prep.	Count
to	5140589	on	1351260
of	5107531	with	1325244
in	3645151	at	991039
for	1865842	...	...

4.3. Using collocations

The context of the preposition may be helpful in choosing attractive distractors. In terms of our evaluation metrics, a preposition that collocates frequently with *either* A *or* B in a large English corpus might make a more *difficult* distractor for the user; on the other hand, one that has appeared in the corpus with *both* A *and* B is unlikely to be *usable*.

Following this intuition, this method returns the preposition that appears frequently with either A or B, but not both at the same time; formally, $\langle A, p, B\rangle \longmapsto \arg\max_{\overline{p}}\{c(\langle A, \overline{p}, *\rangle) + c(\langle *, \overline{p}, B\rangle)\}$ with the constraint that $c(\langle A, \overline{p}, B\rangle) = 0$, where $c(.)$ is the count. Consider the cloze item in Figure 2. On the strength of the popularity of the collocation "go on", and

Table 2: *Context representations extracted from a non-native English corpus. All errors in the original sentences not involving prepositions are suppressed before extraction. One example each of insertion, deletion and substitution errors are provided.*

Error	Version	Transcript	Context
Deletion	Corrected	I have movie tickets, so I'd like to **go to** the **movie** with you.	$\langle$*go,to,movie*$\rangle$
	Original	I have a movie tickets, so I'd like to **go movie** with you.	$\langle$*go,null,movie*$\rangle$
Insertion	Corrected	So we decided to **take** the kitty **home**.	$\langle$*take,null,home*$\rangle$
	Original	So we decided to **take** the kitty **to home**.	$\langle$*take,to,home*$\rangle$
Substitution	Corrected	He **studies at** the **university**.	$\langle$*study,at,university*$\rangle$
	Original	He **studies in** the **university**.	$\langle$*study,in,university*$\rangle$

the non-occurrence of $\langle$*go,on,movie*$\rangle$ in the English corpus, the preposition "on" is selected as the distractor.

4.4. Using a non-native English corpus

From a corpus of non-native sentences and their corrections, mappings from a triplet to a preposition mistake can be directly estimated. Table 2 illustrates the context extraction of prepositions in such a corpus. The most frequent mistake for each context would then make a reasonable distractor; formally, $\langle A, p, B\rangle \longmapsto \arg\max_{\overline{p}}\{c(\langle A, \overline{p}, B\rangle)\}$. For example, the cloze item in Figure 2 has *null* as the distractor because, for the triplet $\langle$*go,to,movie*$\rangle$, the deletion error is more common than substitution errors in the non-native corpus.

Indeed, more than half of the preposition mistakes in the JLE corpus are deletion errors. One advantage of using a non-native corpus is the ability to directly model contexts where deletion errors are common. It is difficult to do so with native English corpora only, as in the two methods above.

The main drawback is data sparseness[1]. Compared to normal English corpora, non-native corpora are much more expensive to collect; they tend to be much smaller, and restricted to speakers of only a few mother tongues, if not just one.

5. Experiments

This section describes experiments that compare the quality of the distractors generated by the three methods described in §4.2, §4.3 and §4.4. The distractors will be referred to as the *baseline distractor*, *collocation distractor* and *non-native distractor*, respectively. We begin by discussing our corpora.

5.1. Set-up

The 72 prepositions listed in [11] are considered to be the set of prepositions. The context representations are extracted from parse trees derived by a statistical parser [12].

English corpus The English corpus consists of about 10 million sentences from the New York Times.

Non-native corpus The non-native corpus is the Japanese Learners of English (JLE) corpus [3], which contains about 1,300 instances of preposition mistakes. As illustrated in Table 2, one $\langle A, p, B\rangle$ and one $\langle A, \overline{p}, B\rangle$ are harvested from each mistake.

Source corpus The source corpus is the BTEC corpus [13], used in the evaluation campaign of the International Workshop on Spoken Language Translation. It consists of about 24,000 transcripts from the travel domain. Only sentences at least five words long are utilized.

To ensure a fair comparison, a sentence qualifies as a seed sentence only when all three methods can generate a distractor. In practice, the non-native corpus is the constraining factor. To select the most reliable distractors, we require the seed sentence's triplet $\langle A, p, B\rangle$ to occur two times or more in the non-native corpus, or its $\langle A, p, *\rangle$ or $\langle *, p, B\rangle$ to occur four times or more. With these restrictions, 328 cloze items were generated.

Interestingly, the three methods rarely generate the same distractor. The non-native distractor agrees with the collocation distractor 9.5% of the time, and intersects with the baseline only 4.4% of the time. The collocation and baseline distractors are identical 12.7% of the time. Most cloze items thus offer four different choices.

5.2. Analyses

5.2.1. Usability

A cloze item is considered usable when all distractors result in an inappropriate sentence [7]. The second author of this paper, who is a native speaker of English and was not involved in the cloze item generation process, performed the usability study. She took the cloze test, identifying all choices which yield acceptable English sentences.

In 12 out of 328 cloze items, one of the distractors yielded a correct sentence; in other words, 96.3% of the automatically generated items were usable. To put this performance level in context, usability rates of 93.5% [6] and 91.5% [7] have been reported in the literature, although their tasks and corpora are different, and the results are hence not directly comparable.

Among the unusable distractors, more than half are collocation distractors. For example, from the seed sentence "*I have sore pain here*", the collocation method produces the distractor "around", yielding an acceptable sentence "*I have sore pain around here*".

5.2.2. Difficulty

After omitting the unusable cloze items identified in the previous step, we split the remaining 316 cloze items into two tests, with 158 questions each. Our subjects are four students whose mother tongue is Mandarin. They are all students in their second or third year of senior high school in Taiwan.

The overall performance of the subjects is listed in Table 3. The subjects made a total of 106 mistakes, of which 12% are insertions, 29% are deletions, 58% are substitutions. A breakdown of the distractors responsible for the mistakes is provided in Table 4 with respect to the subjects, and in Table 5 with respect to error types. Figure 3 shows a few cloze items for which both subjects made an error.

[1] A possible mitigation of this problem, which we have not yet explored, is to initially generate cloze items from collocations only, then harvest the mistakes made by users to grow a non-native corpus.

Figure 3: *Some cloze items for which both subjects made an error. The* **bolded** *items are the selected distractors.*

It's really different driving ____ the right side of the street	
(a) on [key]	**(b) *null* [non-native]**
(c) with [baseline]	**(d) *to* [collocation]**
Could I take the leftovers ____ home?	
(a) *in* [collocation]	(b) about [baseline]
(c) *to* [non-native]	(d) *null* [key]

Table 3: *Overall performance on the cloze tests.*

Test 1		Test 2	
Subject	Score	Subject	Score
Student 1	75.9%	Student 3	91.1%
Student 2	76.6%	Student 4	91.1%

Overall, distractors produced by the collocation and non-native methods were more successful in attracting the subjects. The subjects were two to three times more likely to choose a non-native distractor than a baseline one; with the exception of Student 1, the same difference is observed between the collocation and baseline distractors.

6. Conclusion & Future Work

Prepositions present some new challenges in cloze item generation. We have proposed two novel distractor generation methods, one based on collocations, the other on direct observations in a non-native corpus. We have compared them with a baseline method based on word frequency. The distractors generated by the two novel methods were more successful in attracting the subjects than the baseline method.

We believe there is still much room for improvement. In particular, we plan to further enrich the context representation, and also explore techniques that would work well even on a small non-native corpus.

In the future, we would like to generate cloze tests tailored for other error classes. One class of interest to us, which also occurs frequently in the JLE corpus, is the misuse of verb forms, e.g., the infinitive, participle, gerund and base forms.

7. Acknowledgements

This research was partially supported by MIT Lincoln Laboratory and by a post-graduate fellowship from the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank the four subjects and Hsu-Chih Wu for assisting in the user study, and Terry Koo for parsing the English corpus.

Table 4: *Number of distractors chosen by subjects.*

Subject	Non-native	Collocation	Baseline
Student 1	17	10	10
Student 2	20	12	6
Student 3	4	9	2
Student 4	6	8	2
Total	**47**	**39**	**20**

Table 5: *A breakdown of the the distractors into the error types. "Success" refers to the number of distractors that were selected as answer by the subjects, out of the "Total" number that appeared in the cloze tests.*

Error Type	Non-native Success (Total)	Collocation Success (Total)	Baseline Success (Total)
Del	31 (211)	0 (0)	0 (0)
Ins	5 (79)	5 (79)	3 (79)
Sub	11 (26)	34 (237)	17 (237)
Total	**47** (316)	**39** (316)	**20** (316)

8. References

[1] Brown, J. C., Frishkoff, G. A., and Eskenazi, M. Automatic Question Generation for Vocabulary Assessment. *Proc. HLT-EMNLP*, 2005.

[2] Shei, C.-C. FollowYou!: An Automatic Language Lesson Generation System. *Computer Assisted Language Learning*, 14(2):129-144, 2001.

[3] Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., and Isahara, H. Automatic Error Detection in the Japanese Learners' English Spoken Data. *Proc. ACL*, 2003.

[4] Coniam, D. From Text to Test, Automatically — An Evaluation of a Computer Cloze-Test Generator. *Hong Kong Journal of Applied Linguistics* 3(1):41-60, 1998.

[5] Hoshino, A. and Nakagawa, H. A Real-Time Multiple-Choice Question Generator for Language Testing: A Preliminary Study. *Proc. 2nd Workshop on Building Educational Applications using NLP*, Ann Arbor, MI, 2005.

[6] Sumita, E., Sugaya, F., and Yamamoto, S. Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. *Proc. 2nd Workshop on Building Educational Applications using NLP*, Ann Arbor, MI, 2005.

[7] Liu, C.-L., Wang, C.-H., Gao, Z.-M., and Huang, S.-M. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. *Proc. 2nd Workshop on Building Educational Applications using NLP*, Ann Arbor, MI, 2005.

[8] Chen, C.-Y., Liou, H.-C., and Chang, J. S. FAST — An Automatic Generation System for Grammar Tests. *Proc. COLING/ACL Interactive Presentation Sessions*, 2006.

[9] Karamanis, N., Ha, L. A., and Mitkov, R. Generating Multiple-Choice Test Items from Medical Text: A Pilot Study. *Proc. 4th International Natural Language Generation Conference*, Sydney, Australia, 2006.

[10] Mitkov, R., Ha, L. A. Computer-aided Generation of Multiple-choice Tests. *Proc. HLT-NAACL Workshop on Building Educational Applications using NLP*, 2003.

[11] Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. A Comprehensive Grammar of the English Language. Longman, New York. 1985.

[12] Collins, M. and Koo, T. Discriminative Reranking for Natural Language Parsing. *Computational Linguistics* 31(1):25-69. 2005.

[13] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World. *Proc. LREC*, Las Palmas, Spain, 2002.