Automatic Generation of Challenging Distractors Using Context-Sensitive Inference Rules

Torsten Zesch
Language Technology Lab
University of Duisburg-Essen
torsten.zesch@uni-due.de

Oren Melamud
Computer Science Department
Bar-Ilan University
melamuo@cs.biu.ac.il

Abstract

Automatically generating challenging distractors for multiple-choice gap-fill items is still an unsolved problem. We propose to employ context-sensitive lexical inference rules in order to generate distractors that are semantically similar to the gap target word in some sense, but not in the particular sense induced by the gap-fill context. We hypothesize that such distractors should be particularly hard to distinguish from the correct answer. We focus on verbs as they are especially difficult to master for language learners and find that our approach is quite effective. In our test set of 20 items, our proposed method decreases the number of invalid distractors in 90% of the cases, and fully eliminates all of them in 65%. Further analysis on that dataset does not support our hypothesis regarding item difficulty as measured by average error rate of language learners. We conjecture that this may be due to limitations in our evaluation setting, which we plan to address in future work.

1 Introduction

Multiple-choice gap-fill items as illustrated in Figure 1 are frequently used for both testing language proficiency and as a learning device. Each item consists of a *carrier sentence* that provides the context to a *target word*. The target word is blanked and presented as one possible gap-fill answer together with a certain number (usually 3) of *distractors*. Given a desired target word, carrier sentences containing it can be automatically selected from a corpus. Some methods even select only sentences where the target word is used in a certain sense (Liu et al., 2005). Then, the main problem is to pick challenging distractors that are

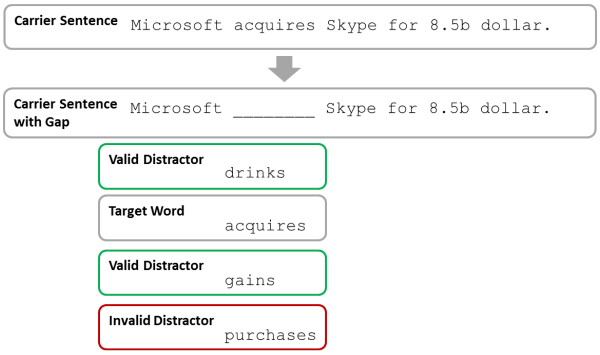

Figure 1: Multiple-choice gap-fill item.

reasonably hard to distinguish from the correct answer (i.e. the target word) on one hand, yet cannot be considered as correct answers on the other.

In this paper we propose to generate distractors that are semantically similar to the gap target word in some sense, but not in the particular sense induced by the gap-fill context, thereby making them difficult to distinguish from the target word. For example, the distractor *gain* in Figure 1 is semantically similar to *acquire*, but is not appropriate in the particular context of purchasing companies, and therefore has high distractive potential. On the other hand, the distractor *purchase* is a correct answer in this context and is therefore an invalid distractor. To generate challenging distractors, we utilize context-sensitive lexical inference rules that can discriminate between appropriate substitutes of a target word given its context and other inappropriate substitutes.

In the next section, we give an overview of previous work in order to place our contribution into context.

2 Previous Work

The process of finding good distractors involves two steps: *Candidate Selection* controls the difficulty of the items, while *Reliability Checking* ensures that the items remain solvable, i.e. it ensures

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 143–148,
Baltimore, Maryland USA, June 26, 2014.

that there is only one correct answer. We note that this work is focused on single-word distractors rather than phrases (Gates et al., 2011), and only on target isolated carrier sentences rather than longer texts as in (Mostow and Jang, 2012).

2.1 Candidates Selection

In some settings the set of possible distractors is known in advance, e.g. the set of English prepositions in preposition exercises (Lee and Seneff, 2007) or a confusion set with previously known errors like {two, too, to}. Sakaguchi et al. (2013) use data from the Lang-8 platform (a corpus of manually annotated errors[1]) in order to determine typical learner errors and use them as distractors. However, in the common setting only the target word is known and the set of distractors needs to be automatically generated.

Randomly selecting distractors is a valid strategy (Mostow and Jang, 2012), but it is only suitable for the most beginner learners. More advanced learners can easily rule out distractors that do not fit grammatically or are too unrelated semantically. Thus, more advanced approaches usually employ basic strategies, such as choosing distractors with the same part-of-speech tag as the target word, or distractors with a corpus frequency comparable to the target word (Hoshino and Nakagawa, 2007) (based on the assumption that corpus frequency roughly correlates with word difficulty). Pino and Eskenazi (2009) use distractors that are morphologically, orthographically, or phonetically similar (e.g. *bread – beard*).

Another approach used in previous works to make distractors more challenging is utilizing thesauri (Sumita et al., 2005; Smith and Avinesh, 2010) or taxonomies (Hoshino and Nakagawa, 2007; Mitkov et al., 2009) to select words that are semantically similar to the target word. In addition to the target word, some approaches also consider the semantic relatedness of distractors with the whole carrier sentence or paragraph (Pino et al., 2008; Agarwal and Mannem, 2011; Mostow and Jang, 2012), i.e. they pick distractors that are from the same domain as the target word.

Generally, selecting more challenging distractors usually means making them more similar to the target word. As this increases the probability that a distractor might actually be another correct answer, we need a more sophisticated approach for checking the reliability of the distractor set.

[1] http://cl.naist.jp/nldata/lang-8/

2.2 Reliability Checking

In order to make sure that there is only one correct answer to a gap-fill item, there needs to be a way to decide for each distractor whether it fits into the context of the carrier sentence or not. In those cases, where we have a limited list of potential target words and distractors, e.g. in preposition exercises (Lee and Seneff, 2007), a supervised classifier can be trained to do this job. Given enough training data, this approach yields very high precision, but it cannot be easily applied to open word classes like nouns or verbs, which are much larger and dynamic in nature.

When we do not have a closed list of potential distractors at hand, one way to perform reliability checking is by considering collocations involving the target word (Pino et al., 2008; Smith and Avinesh, 2010). For example, if the target word is *strong*, we can find the collocation *strong tea*. Then we can use *powerful* as a distractor because it is semantically similar to *strong*, yet **powerful tea* is not a valid collocation. This approach is effective, but requires strong collocations to discriminate between valid and invalid distractors. Therefore it cannot be used with carrier sentences that do not contain strong collocations, such as the sentence in Figure 1.

Sumita et al. (2005) apply a simple web search approach to judge the reliability of an item. They check whether the carrier sentence with the target word replaced by the distractor can be found on the web. If such a sentence is found, the distractor is discarded. We note that the applicability of this approach is limited, as finding exact matches for such artificial sentences can be unlikely due to sparseness of natural languages. Therefore not finding an exact match does not necessarily rule out the possibility of an invalid distractor.

3 Automatic Generation of Challenging Distractors

Our goal is to automatically generate distractors that are as 'close' to the target word as possible, yet do not fit the carrier sentence context. To accomplish this, our strategy is to first generate a set of distractor candidates, which are semantically similar to the target word. Then we use context-sensitive lexical inference rules to filter candidates that fit the context, and thus cannot be used as dis-

tractors. In the remainder of this section we describe this procedure in more detail.

3.1 Context-Sensitive Inference Rules

A lexical inference rule '*LHS* → *RHS*', such as '*acquire* → *purchase*', specifies a directional inference relation between two words (or terms). A rule can be *applied* when its LHS matches a word in a text T, and then that word is substituted for RHS, yielding the modified text H. For example, applying the rule above to "*Microsoft acquired Skype*", yields "*Microsoft purchased Skype*". If the rule is true then the meaning of H is inferred from the meaning of T. A popular way to learn lexical inference rules in an unsupervised setting is by using distributional similarity models (Lin and Pantel, 2001; Kotlerman et al., 2010). Under this approach, target words are represented as vectors of context features, and the score of a rule between two target words is based on vector arithmetics.

One of the main shortcomings of such rules is that they are *context-insensitive*, i.e. they have a single score, which is not assessed with respect to the concrete context T under which they are applied. However, the appropriateness of an inference rule may in fact depend on this context. For example, '*Microsoft acquire Skype* → *Microsoft purchase Skype*', is an appropriate application of the rule '*acquire* → *purchase*', while '*Children acquire skills* → *Children purchase skills*' is not. To address this issue, additional models were introduced that compute a different *context-sensitive* score per each context T, under which it is applied (Dinu and Lapata, 2010; Melamud et al., 2013).

In this work, we use the resource provided by Melamud et al. (2013), which includes both context-sensitive and context-insensitive rules for over 2,000 frequent verbs.[2] We use these rules to generate challenging distractors as we show next.

3.2 Distractor Selection & Reliability

We start with the following illustrative example to motivate our approach. While the words *purchase* and *acquire* are considered to be almost perfect synonyms in sentences like *Microsoft acquires Skype* and *Microsoft purchases Skype*, this is not true for all contexts. For example, in *Children acquire skills* vs. *Children purchase skills*, the meaning is clearly not equivalent. These context-dependent senses, which are particularly typical to verbs, make it difficult for learners to understand how to properly use these words.

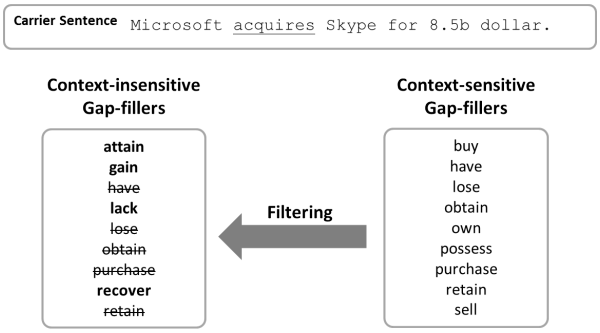

Figure 2: Filtering context-insensitive substitutions with context-sensitive ones in order to get challenging distractors.

Acquiring such fine-grained sense distinction skills is a prerequisite for really competent language usage. These skills can be trained and tested with distractors, such as *purchase* in the example above. Therefore, such items are good indicators in language proficiency testing, and should be specifically trained when learning a language.

To generate such challenging distractors, we first use the context-insensitive rules, whose LHS matches the carrier sentence target word, to create a *distractor candidate set* as illustrated on the left-hand side of Figure 2. We include in this set the top-n inferred words that correspond to the highest rule scores. These candidate words are inferred by the target word, but not necessarily in the particular context of the carrier sentence. Therefore, we expect this set to include both correct answers, which would render the item unreliable, as well as good distractors that are semantically similar to the target word in some sense, but not in the particular sense induced by the carrier sentence.

Next, we use context-sensitive rules to generate a *distractor black-list* including the top-m words that are inferred by the target word, but this time taking the context of the carrier sentence into consideration. In this case, we expect the words in the list to comprise only the gap-fillers that fit the given context as illustrated on the right-hand side of Figure 2. Such gap-fillers are correct answers and therefore cannot be used as distractors. Finally, we subtract the black-list distractors from the initial distractor candidate set and expect the remaining candidates to comprise only good distractors. We consider the candidates in this final set as our generated distractors.

[2] http://www.cs.biu.ac.il/nlp/downloads/wt-rules.html

3.3 Distractor Ranking

In case our approach returns a large number of good distractors, we should use ranking to select the most challenging ones. A simple strategy is to rely on the corpus frequency of the distractor, where less frequent means more challenging as it will not be known to the learner. However, this tends to put a focus on the more obscure words of the vocabulary while actually the more frequent words should be trained more often. Therefore, in this work we use the scores that were assigned to the distractors by the context-insensitive inference rules. Accordingly, the more similar a distractor is to the target word, the higher rank it will get (provided that it was not in the distractor black-list).

4 Experiments & Results

In our experiments we wanted to test two hypotheses: (i) whether context-sensitive inference rules are able to reliably distinguish between valid and invalid distractors, and (ii) whether the generated distractors are more challenging for language learners than randomly chosen ones.

We used the Brown corpus (Nelson Francis and Kuçera, 1964) as a source for carrier sentences and selected medium-sized (5-12 tokens long) sentences that contain a main verb. We then manually inspected this set, keeping only well-formed sentences that are understandable by a general audience without requiring too much context knowledge. In a production system, this manual process would be replaced by a sophisticated method for obtaining good carrier sentences, but this is beyond the scope of this paper. Finally, for this exploratory study, we only used the first 20 selected sentences from a much larger set of possible carrier sentences.

4.1 Reliability

Our first goal was to study the effectiveness of our approach in generating reliable items, i.e. items where the target word is the only correct answer. In order to minimize impact of pre-processing and lemmatization, we provided the context-sensitive inference rules with correctly lemmatized carrier sentences and marked the target verbs. We found that we get better results when using a distractor black-list that is larger than the distractor candidate set, as this more aggressively filters invalid distractors. We used the top-20 distractor blacklist and top-10 distractor candidate set, which lead to generating on average 3.3 distractors per item.

Only valid distractors	13/20 (65%)
Mix of valid and invalid	5/20 (25%)
Only invalid distractors	2/20 (10%)

Table 1: Reliability of items after filtering

All our generated distractors were checked by two native English speakers. We count a distractor as "invalid" if it was ruled out by at least one annotator. Table 1 summarizes the results. We found that in 13 of the 20 items (65%) all distractors generated by our approach were valid, while only for 2 items all generated distractors were invalid. For the remaining 5 items, our approach returned a mix of valid and invalid distractors. We note that the unfiltered distractor candidate set always contained invalid distractors and in 90% of the items it contained a higher proportion of invalid distractors than the filtered one. This suggests that the context-sensitive inference rules are quite effective in differentiating between the different senses of the verbs.

A main source of error are sentences that do not provide enough context, e.g. because the subject is a pronoun. In *She [served] one four-year term on the national committee*, it would be acceptable to insert *sold* in the context of a report on political corruption, but a more precise subject like *Barack Obama* would render that reading much more unlikely. Therefore, more emphasis should be put on selecting better carrier sentences. Selecting longer sentences that provide a richer context would help to rule out more distractor candidates and may also lead to better results when using the context-sensitive inference rules. However, long sentences are also more difficult for language learners, so there will probably be some trade-off.

A qualitative analysis of the results shows that especially for verbs with clearly distinct senses, our approach yields good results. For example in *He [played] basketball there while working toward a law degree*, our method generates the distractors *compose* and *tune* which are both related to the "play a musical instrument" sense. Another example is *His petition [charged] mental cruelty*, where our method generates among others the distractors *pay* and *collect* that are both related to the "charge taxes" reading of the verb. *The ball [floated] downstream* is an example where our method did not work well. It generated the distractors *glide* and *travel* which also fit the context and

	Group 1	Group 2
Control Items	0.24 ± 0.12	0.20 ± 0.12
Test Items	0.18 ± 0.17	0.18 ± 0.15

Table 2: Average error rates on our dataset

should thus not be used as distractors. The verb *float* is different from the previous examples, as all its dominant senses involve some kind of "floating" even if only metaphorically used. This results in similar senses that are harder to differentiate.

4.2 Difficulty

Next, we wanted to examine whether our approach leads to more challenging distractors. For that purpose we removed the distractors that our annotators identified as invalid in the previous step. We then ranked the remaining distractors according to the scores assigned to them by the context-sensitive inference rules and selected the top-3 distractors. If our method generated less than 3 distractors, we randomly generated additional distractors from the same frequency range as the target word.

We compared our approach with randomly selected distractors that are in the same order of magnitude with respect to corpus frequency as the distractors generated by our method. This way we ensure that a possible change in distractor difficulty cannot simply be attributed to differences in the learners' familiarity with the distractor verbs due to their corpus frequency. We note that random selection repeatedly created invalid distractors that we needed to manually filter out. This shows that better methods for checking the reliability of items like in our approach are definitely required.

We randomly split 52 participants (all non-natives) into two groups, each assigned with a different test version. Table 2 summarizes the results. For both groups, the first 7 test items were identical and contained only randomly selected distractors. Average error rate for these items was 0.24 (SD 0.12) for the first group, and 0.20 (SD 0.12) for the second group, suggesting that the results of the two groups on the remaining items can be compared meaningfully. The first group was tested on the remaining 13 items with randomly selected distractors, while the second group got the same items but with distractors created by our method.

Contrary to our hypothesis, the average error rate for both groups was equal (0.18, SD_1=0.17, SD_2=0.15). One reason might be that the English language skills of the participants (mostly computer science students or faculty) were rather high, close to the native level, as shown by the low error rates. Furthermore, even if the participants were more challenged by our distractors, they might have been able to finally select the right answer with no measurable effect on error rate. Thus, in future work we want measure answer time instead of average error rate, in order to counter this effect. We also want to re-run the experiment with lower grade students, who might not have mastered the kind of sense distinctions that our approach is focused on.

5 Conclusions

In this paper we have tackled the task of generating challenging distractors for multiple-choice gap-fill items. We propose to employ context-sensitive lexical inference rules in order to generate distractors that are semantically similar to the gap target word in some sense, but not in the particular sense induced by the gap-fill context.

Our results suggest that our approach is quite effective, reducing the number of invalid distractors in 90% of the cases, and fully eliminating all of them in 65%. We did not find a difference in average error rate between distractors generated with our method and randomly chosen distractors from the same corpus frequency range. We conjecture that this may be due to limitations in the setup of our experiment.

Thus, in future work we want to re-run the experiment with less experienced participants. We also wish to measure answer time in addition to error rate, as the distractive powers of a gap-filler might be reflected in longer answer times more than in higher error rates.

Acknowledgements

We thank all participants of the gap-fill survey, and Emily Jamison and Tristan Miller for their help with the annotation study. This work was partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287923 (EXCITEMENT).

References

Manish Agarwal and Prashanth Mannem. 2011. Automatic Gap-fill Question Generation from Text Books. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–64.

Georgiana Dinu and Mirella Lapata. 2010. Measuring Distributional Similarity in Context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1162–1172.

Donna Gates, Margaret Mckeown, Juliet Bey, Forbes Ave, and Ross Hall. 2011. How to Generate Cloze Questions from Definitions : A Syntactic Approach. In *Proceedings of the AAAI Fall Symposium on Question Generation*, pages 19–22.

Ayako Hoshino and Hiroshi Nakagawa. 2007. Assisting Cloze Test Making with a Web Application. In *Proceedings of the Society for Information Technology and Teacher Education International Conference*, pages 2807– 2814.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

John Lee and Stephanie Seneff. 2007. Automatic Generation of Cloze Items for Prepositions. In *Proceedings of INTERSPEECH*, pages 2173–2176, Antwerp, Belgium.

Dekang Lin and Patrick Pantel. 2001. DIRT – Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*.

Chao-lin Liu, Chun-hung Wang, and Zhao-ming Gao. 2005. Using Lexical Constraints to Enhance the Quality of Computer-Generated Multiple-Choice Cloze Items. *Computational Linguistics and Chinese Language Processing*, 10(3):303–328.

Oren Melamud, Jonathan Berant, Ido Dagan, Jacob Goldberger, and Idan Szpektor. 2013. A Two Level Model for Context Sensitive Inference Rules. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1331–1340, Sofia, Bulgaria.

Ruslan Mitkov, Le An Ha, Andrea Varga, and Luz Rello. 2009. Semantic Similarity of Distractors in Multiple-choice Tests: Extrinsic Evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 49–56.

Jack Mostow and Hyeju Jang. 2012. Generating Diagnostic Multiple Choice Comprehension Cloze Questions. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 136–146, Stroudsburg, PA, USA.

W. Nelson Francis and Henry Kuçera. 1964. Manual of Information to Accompany a Standard Corpus of Present-day Edited American English, for use with Digital Computers.

Juan Pino and Maxine Eskenazi. 2009. Semi-Automatic Generation of Cloze Question Distractors Effect of Students L1. In *SLaTE Workshop on Speech and Language Technology in Education*.

Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A Selection Strategy to Improve Cloze Question Quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 9th Internationnal 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 *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–242, Sofia, Bulgaria.

Simon Smith and P V S Avinesh. 2010. Gap-fill Tests for Language Learners: Corpus-Driven Item Generation. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*.

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 *Proceedings of the second workshop on Building Educational Applications Using NLP*, EdAppsNLP 05, pages 61–68, Stroudsburg, PA, USA.