# Using Neural Machine Translation for Generating Diverse Challenging Exercises for Language Learners

**Frank Palma Gomez**[1] **Subhadarshi Panda**[*] **Michael Flor**[2] **Alla Rozovskaya**[1,3]
[1]Queens College, CUNY [2]Educational Testing Service [3]CUNY Graduate Center
frankpalma12@gmail.com, subhadarshipanda08@gmail.com
mflor@ets.org, arozovskaya@qc.cuny.edu

## Abstract

We propose a novel approach to automatically generate distractors for cloze exercises for English language learners, using round-trip neural machine translation. A carrier sentence is translated from English into another (pivot) language and back, and distractors are produced by aligning the original sentence with its round-trip translation. We make use of 16 linguistically-diverse pivots and generate hundreds of translation hypotheses in each direction. We show that using hundreds of translations allows us to generate a rich set of challenging distractors. Moreover, we find that typologically unrelated language pivots contribute more diverse candidate distractors, compared to language pivots that are closely related. We further evaluate the use of machine translation systems of varying quality and find that better quality MT systems produce more challenging distractors. Finally, we conduct a study with language learners, demonstrating that the automatically generated distractors are of the same difficulty as the gold distractors produced by human experts.[1]

## 1 Introduction

A common challenge for language learners involves understanding how to appropriately use words that may have similar meanings but are used in different contexts. For instance, "main" and "vital" are semantically related but "main importance" is not an acceptable expression while "vital importance" is. This subtle language knowledge is not explicitly available to learners. For this reason, word usage (collocation) errors are some of the most common types of errors even for advanced non-native speakers (Leacock et al., 2010).

| **Carrier sentence** |
|---|
| *Are these old plates of _____ importance or can I put them into storage?* |
| **Target word:** *vital* |
| **Valid distractors:** *main, urgent, lively* |
| **Invalid distractors:** *great, utmost* |

Table 1: A sentence for a fill-in-the-blank exercise with the target word "vital" removed. Multiple-choice list will include the target and 3 distractors. Examples of valid and invalid distractors are shown.

In this work, we develop exercises for mastering vocabulary use for second (foreign) language learners. We focus on cloze (fill-in-the-blank) exercises. A cloze exercise is a common method of teaching vocabulary, as well as assessing non-native speaker performance in a foreign language: a sentence is presented to the learner with one word (*target*) hidden. The target word is presented along with a list of *distractors* (usually 3), and the learner should identify the target word from that list. Table 1 shows a sample cloze item with the target word "vital". The *carrier sentence* along with a multiple-choice list is referred to as *cloze item*. A cloze (exercise) item is valid if exactly one word (the target) fits the context. Therefore, a valid distractor should be a word that does not fit the context. Thus, "great" and "utmost" in Table 1 are invalid distractors, since they both fit the context.

Given a carrier sentence and the target word, the problem is to generate distractors. Distractors are typically created manually by educational testing experts, a time-consuming procedure. The problem becomes more challenging once the exercises are aimed at high-proficiency learners, since distractors that are not semantically close to the target word or are grammatically unfit will be too easy for them (Zesch and Melamud, 2014).

[*]Work was done while the author was at the CUNY Graduate Center.

[1]The code is available at https://github.com/subhadarship/round-trip-distractors

We propose to generate distractors using round-trip neural machine translation (MT). Robust machine translation systems exist today for many language pairs. While translations produced with modern automated systems are reasonably good, these are not perfect, and, while a round-trip translation may preserve the sentence meaning, it will often not result in the exact same sentence. We use this observation to develop an approach to automatically generate distractors for cloze exercises.

We focus on exercises aimed at *advanced* English as a Second Language (ESL) learners. A carrier sentence is translated from English into another *pivot* language, where the top $n$ translation hypotheses are generated. For each hypothesis, the top $m$ back-translations into English are generated. Each back-translation is aligned with the original sentence, and the back-translated word aligned to the target is treated as a potential distractor.

The intuition behind the approach is that word choice errors are commonly affected by the learner's first language. In particular, the different meanings (or contextual uses) of an ambiguous word in the learner's native language may lead to different word choices in English. The assumption thus is that lexical challenges that are common with non-native speakers will also manifest themselves in the round-trip machine translation as back-translated words that are semantically close to the target. Such words should therefore serve as challenging distractors for advanced learners. Unlike previous work, this method also opens up a possibility of *customizing* the cloze task for speakers of different languages.

This work builds on a pilot study (Panda et al., 2022) that made use of five round-trip MT systems. However, the pivots used in the study were closely related languages spoken in Europe. In addition, the study did not evaluate *the difficulty of the automatic distractors* and did not test these with language learners.

In this paper, we use 16 language pivots from a diverse set of linguistic families and conduct a thorough evaluation of the proposed method, using a dataset of real cloze exercises for advanced learners. Our contributions are as follows: (1) We use MT systems of varying levels of quality. We show that, while poor MT systems generate a larger pool of candidate distractors, high quality systems tend to produce more challenging distractors that are semantically close to the target word; (2) We evaluate the approach as a function of using pivots from different language families and show that pivot languages that are typologically distant contribute more diverse distractors; (3) We conduct a human study with 32 advanced language learners and show that the generated distractors are of the same difficulty as distractors created by experts.

The rest of the paper is organized as follows. The next section presents related work. Section 3 describes the dataset of cloze exercises. Section 4 describes our approach. Section 5 presents the evaluation of the approach along several dimensions. Section 6 describes the human study. Section 7 concludes, by outlining avenues for future work and discussing the limitations of the study.

## 2 Related work

Previous work on distractor generation made use of word frequency, phonetic and morphological similarity, and grammatical fit (Hoshino and Nakagawa, 2005; Pino and Eskénazi, 2009; Goto et al., 2010). For advanced speakers, distractors should be selected more carefully, so that they are reasonably hard to distinguish from the target. Consider, for example, the target word "error" in the carrier sentence: "It is often only through long experiments of trial and *error* that scientific progress is made." The word "mistake" is semantically close to it but is not appropriate in the sentence, and thus could serve as a valid distractor. However, note that "mistake" can be substituted for "error" in the context of "He made a lot of mistakes in his test." and would therefore not be a valid distractor in that context. Thus, challenging distractors should be *semantically close* to the target word, yet, a valid distractor *should not produce an acceptable sentence*.

Most of the approaches to generating challenging distractors rely on semantic relatedness, computed through n-grams and collocations (Liu et al., 2005; Hill and Simha, 2016), thesauri (Sumita et al., 2005), or WordNet (Brown et al., 2005). Zesch and Melamud (2014) use semantic context-sensitive inference rules. Sakaguchi et al. (2013) propose generating distractors using errors mined from a learner corpus. The approach, however, assumes an annotated learner corpus, and both the choice of the target word and of the distractors are constrained by the errors in the corpus. Several studies showed that word embeddings are effective in distractor generation (Jiang and Lee, 2017; Susanti et al., 2018; Mikolov et al., 2013).

Our work builds on a study that employed five pivot languages (Panda et al., 2022), showing that the round-trip MT approach outperforms two strong baselines – word2vec and BERT (Section 5.4 and Appendix B provide more detail on the comparison of the MT approach with these methods). The present study focuses on an *in-depth evaluation of the MT approach to distractor generation* along several dimensions.

## 3 Data

We obtain cloze exercises from a reputable test preparation website, ESL Lounge.[2] The website contains study materials and preparatory exercises for ESL tests, such as FCE First Certificate, TOEFL, and International English Language Testing System (IELTS). There was significant effort put into the development of the exercises, which were manually curated for ESL students, and the exercises are of high quality. This is the first dataset that can be used by researchers working on the task.[3] Previous studies thus evaluate either on artificially created items or on proprietary data.

We use the advanced level multiple choice cloze exercises, which includes 142 cloze items.[4] Each *item* consists of a carrier sentence with the target word removed and is accompanied by four word choices that include the target word and three distractors provided by human experts. We refer to these distractors as *gold* distractors.

## 4 Generating Distractors with Neural MT

**Round-trip machine translation** Given a carrier sentence $X$ with the target word, a forward machine translation system from English to a pivot language $trg$ and a backward MT system from $trg$ to English, we can generate a round-trip translation for $X$. Importantly, we generate multiple hypotheses in each direction.

We first translate the sentence $X$ from English using a forward MT system $S_{en-trg}$ to obtain a set of top $N_f$ translation hypotheses $Y = \{Y_1, Y_2, \ldots, Y_{N_f}\}$ in the target language $trg$. We then translate the sentences in $Y$ using a backward MT system $S_{trg-en}$ and obtain a set of top $N_b$ translation hypotheses for $Y_i \in Y$. Finally, we obtain the set of round-trip translations $X_{RT} = \{X_{RT_1}, X_{RT_2}, \ldots, X_{RT_{N_f \times N_b}}\}$.

| Pivot group | Pivot language | BLEU Fwd | BLEU Bwd | BLEU Avg |
|---|---|---|---|---|
| Group 1 | Italian (it) | 48.2 | 70.9 | 59.6 |
| | Dutch (nl) | 57.1 | 60.9 | 59.0 |
| | Spanish (es) | 54.9 | 59.6 | 57.3 |
| | Russian (ru) | 48.4 | 61.1 | 54.8 |
| | French (fr) | 50.5 | 57.5 | 54.0 |
| | Czech (cs) | 46.1 | 58.0 | 52.1 |
| | German (de) | 47.3 | 55.4 | 51.4 |
| Group 2 | Indonesian (id) | 38.3 | 47.7 | 43.0 |
| | Vietnamese (vi) | 37.2 | 42.8 | 40.0 |
| Group 3 | Bislama (bi) | 37.1 | 31.3 | 34.2 |
| | Chinese (zh) | 31.4 | 36.1 | 33.8 |
| | Arabic (ar) | 14.0 | 49.4 | 31.7 |
| | Malayalam (ml) | 19.1 | 42.7 | 30.9 |
| Group 4 | Chuukese (chk) | 26.1 | 31.2 | 28.7 |
| | Hindi (hi) | 16.1 | 40.4 | 28.3 |
| | Urdu (ur) | 12.1 | 23.2 | 17.7 |

Table 2: Pivot languages used in the study sorted by their averaged BLEU scores. *Fwd* stands for forward MT system (from English); *Bwk* stands for backward MT system (from the pivot language into English).

Our earlier study included five Indo-European languages: German, Russian, Italian, French, and Czech. Presently, we include 16 languages from a diverse set of language families. For all language pairs, we use competitive neural MT systems of Tiedemann and Thottingal (2020). Table 2 lists the 16 languages, and includes BLEU scores in both directions and the averaged BLEU scores on the Tatoeba Machine Translation dataset from the Tatoeba Translation Challenge (Tiedemann, 2020). Tatoeba is a crowd-sourced collection of user-provided translations in a large number of languages. We split the languages into four groups, organized by the averaged BLEU scores. We assume higher BLEU scores correspond to back-translations of higher quality. Appendix A provides detail on the pivot grouping.

**Alignment computation** Given a round-trip translation $X_{RT_i}$ for carrier sentence $X$, we compute the alignment between the two sentences. The word in $X_{RT_i}$ that is aligned to the target word in $X$ is considered to be the back-translation of the target and can be a potential distractor. We use Simalign[5] (Sabet et al., 2020) that employs contextual

[2] https://www.esl-lounge.com

[3] A csv copy of the dataset for research purposes can be obtained from the authors.

[4] https://www.esl-lounge.com/student/advanced-multiple-choice-cloze.php

[5] https://github.com/cisnlp/simalign

word embeddings (Devlin et al., 2018) to produce an alignment model for a pair of sentences. Given the original sentence $X$ and a round-trip translation $X_{RT_i}$, the similarity between each token in $X$ with each token in $X_{RT_i}$ is computed, using contextual embeddings from multilingual BERT.

**Candidate filtering** In line with previous studies, we remove candidates that are of a different part-of-speech (POS) than the target word, and those that might fit the carrier sentence. While the first group of candidates would make the item too easy for advanced learners, the second group would make the exercise item invalid, as an item must have only one correct option. To rule out candidates that might fit the context, we use WordNet synonyms (Fellbaum, 1998). We use the NLTK POS tagger (Bird et al., 2009) to remove candidates that have a different tag than the target word in the carrier sentence. The tagger is applied to the carrier sentence with the target position filled by the appropriate word. Filtering removes about 50% of generated candidates. All results are shown with the filtering applied.

## 5 Evaluation

We evaluate the MT approach to distractor generation along 4 dimensions: (1) comparing the effect of using typologically diverse language pivots; (2) using MT systems of various quality; (3) using different number of translation hypotheses in the forward and backward direction; (4) evaluating the diversity of distractors produced with linguistically related versus linguistically unrelated pivots.

Evaluation for the distractor generation task is not straightforward, since the set of valid distractors for a given exercise item is not uniquely defined. For this reason, automatic evaluation against the set of distractors proposed by human experts does not provide a full picture of the quality of the generated distractors. Thus, we conduct several types of evaluation. First, we compare the generated distractors against the set of gold distractors for each item, making the assumption that a method that retrieves a higher percentage of gold distractors among its automatic candidates is better. Second, we conduct manual annotation with native English speakers to determine the percentage of valid distractors among the candidates proposed by MT: although filtering removes a majority of invalid candidates, there are still candidates that remain due to filtering errors. Third, we evaluate the *difficulty* of the generated distractors by annotating the distractors for their semantic similarity to the target. Our final test with language learners in Section 6 assesses the difficulty of the automatic distractors generated using the best settings for MT, as compared to the difficulty of gold distractors.

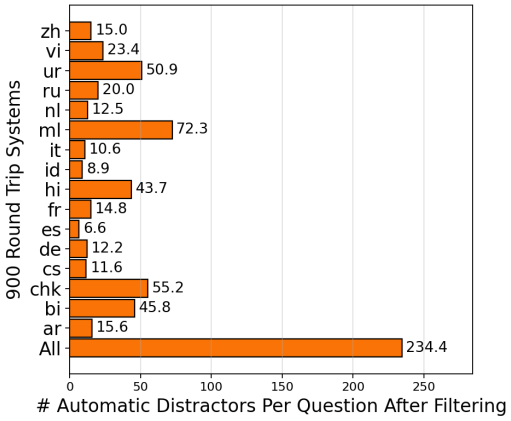


Figure 1: Average number of automatic distractors generated per exercise item with different pivot systems, using 30 translation hypotheses in each direction.

### 5.1 Diversity and quality of distractors by pivot language

With each of the 16 pivot language systems, we generate 900 back-translations for a single exercise item. We use 30 hypotheses in each direction. The carrier sentence is aligned with each of the back-translations, and the back-translated word that is aligned to the target in the original sentence is selected as a candidate distractor. Note that many of the hypotheses are similar and result in the same round-trip translation of the target word.

**How many distractors are generated?** In Figure 1, we show the average number of unique candidate distractors per exercise item, retrieved with each pivot language system and with the union of all the pivot systems. The average number of candidates generated per exercise item varies widely, from 6.6 (Spanish) to 72.3 (Malayalam). Notably, the union produces an average of 234 distractors per target word, suggesting that round-trip translations from different pivot languages contribute unique distractor candidates.

**Gold distractor retrieval** Our assumption is that a better method should generate, among its candidates, more gold distractors. Given a cloze item with its set of 3 gold distractors $D_{gold}$, and an automatic distractor $d$ generated for this cloze item, we compute the distractor retrieval score as follows:

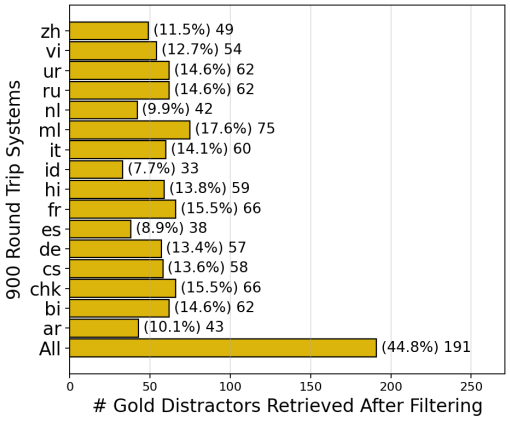


Figure 2: The total number and percentage of gold distractors retrieved for the 142 exercise items with different pivot systems, using 30 translation hypotheses in each direction.

$$r(d, D_{gold}) = \begin{cases} 1 \text{ if } d \in D_{gold} \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

We compute cumulative retrieval score $\sum r(d, D_{gold})$ across all cloze items (the total number of gold distractors is 426, since we have 142 cloze items, each containing 3 gold distractors). Figure 2 shows the cumulative retrieval score (and percentage of gold distractors retrieved) by pivot and for the union of all languages: 44.8% of gold distractors are retrieved with the automatic approach. Compared to the results over 5 language pivots in Panda et al. (2022), gold retrieval score is increased from 31.9% to 44.8% when using 16 pivot languages. The union of the pivot languages is able to retrieve 3 to 4 times as many gold distractors as the individual languages, indicating that *multiple pivots produce diverse candidate distractors*.

**Performance comparison by the quality of MT systems** Table 3 shows gold retrieval (column A) and the number of generated candidates (column B), averaged over the systems in each pivot group. Top MT systems (group 1) retrieve almost as many gold distractors as low-quality systems, but they generate substantially fewer candidates. Overall, better MT systems generate significantly fewer distractor candidates.

**Manual evaluation of distractors for validity** Although filtering removes a substantial number of invalid distractor candidates, there are still invalid candidates (contextual synonyms) that are not filtered out. To determine how many invalid candidates are generated, a set of 100 distractors produced with each pivot system, is evaluated for validity independently by 3 native English speakers. We then compute the percentage of candidates judged as valid (averaged over the 3 raters), shown in Table 3 (column C) by pivot group. Overall, languages in pivot group 1 with better MT systems produce the smallest percentage of valid candidates, while the languages with the poorest MT systems produce the highest percentage of valid candidates. We compute inter-annotator agreement for the 3 native speakers, as described in Appendix C.

| Pivot group | A: Gold distractors retrieved | B: Avg. number of cands. per exercise item | C: Valid cands. (%) |
|---|---|---|---|
| 1 | 55 (12.9%) | 13 | 70.8 |
| 2 | 44 (10.3%) | 16 | 72.4 |
| 3 | 57 (13.4%) | 37 | 75.5 |
| 4 | 62 (14.6%) | 138 | 83.1 |

Table 3: Gold retrieval results over 142 exercise items (column A), the average number of generated candidates per exercise item (column B), and percentage of valid candidates by pivot group (column C). Each value is an averaged result over pivot languages in each group. Using 30 translation hypotheses in each direction.

**Manual evaluation of the difficulty of the automatic distractors by pivot group** To evaluate the *difficulty* of distractors, a trained linguist is presented with an exercise item together with the target word and a proposed distractor and is asked to judge *whether the distractor has semantic similarity* to the context and to the target word (distractors that have semantic similarity are more difficult for a language learner to rule out and thus are more appropriate for advanced language learners). Only candidates judged as valid by all three raters are evaluated for semantic similarity. 10 pivot languages are selected: 4 from group 1, and 2 from each other group. Results averaged by pivot group are shown in Table 4. Better quality MT systems generate a higher percentage of challenging distractors among their candidates. Thus, although the pivots with better MT systems produce fewer candidates overall, there is a substantially higher proportion of difficult distractors among the candidates, compared to pivots with low-quality MT systems. Results by individual pivot are shown in Table D4. Table 5 presents examples of distractors that share semantic similarity with the carrier

| Pivot group | Number of valid cands. | Cands. that have seman. similarity |
|---|---|---|
| 1 | 227 | 125 (55.1%) |
| 2 | 123 | 47 (38.2%) |
| 3 | 135 | 39 (28.9%) |
| 4 | 136 | 50 (36.8%) |

Table 4: Number and percentage of candidates that are *semantically similar* to the target word and the carrier sentence context, among candidates judged as valid by all three raters. Using 30 translation hypotheses in each direction.

sentence and the target word, and those that do not.

### 5.2 Varying the number of generated hypotheses by translation direction

So far, we have evaluated our approach, using 30 translation hypotheses in each direction. We now compare three settings, generating 900 back-translations with 30.30, 900.1, and 1.900, where the first value is the number of hypotheses in the forward direction, while the second value is the number of hypotheses in the backward direction for each forward translation.[6] Table 6 summarizes gold retrieval results and the average number of candidates generated per exercise item, by pivot group. The highest retrieval score is obtained in the 900.1 setting (64.8% of gold distractors are retrieved), whereas the 30.30 setting produces the smallest number of gold distractors (44.8%). The 30.30 setting also produces the smallest number of candidates (234), while the other two settings generate a similar number of candidates (946 and 868). Results by pivot group show similar trends across the 3 settings and are shown in Appendix Table D5. Performance of select individual pivots for the 3 hypothesis settings can be viewed in Appendix Figures D3 and D4.

**Manual evaluation of distractors for validity, by hypothesis setting** We compute the percentage of valid candidates generated in each setting. We use six pivot languages: German and Russian (group 1), Indonesian (group 2), Malayalam (group 3), and Chuukese and Hindi (group 4). For each pivot, we generate 3 sets of distractors (1 set of 100 candidates for each of the 3 direction settings). Each candidate distractor is judged for validity by the three annotators. Results are shown in Table 7: the 900.1 setting generates the highest percentage of valid candidates (91.1%).

**Manual evaluation of the difficulty of the automatic distractors by hypothesis setting** As in previous section, we evaluate the difficulty of the generated distractors, as a function of the translation hypotheses used in each direction. For each of the 6 pivot systems annotated for validity, the same linguist judged, for each candidate considered as valid by all 3 raters, whether the candidate has semantic similarity to the target and to the carrier sentence context. Results are shown in Table 8. In groups 1 and 2, the 30.30 setting produces the highest percentage of candidates with semantic similarity. Overall, the 30.30 setting with languages in group 1 produces the highest percentage of difficult distractors. This is followed by the 30.30 setting group 2 (51.5%). This suggests that *using the 30.30 setting and good MT systems is preferred for generating challenging distractors*. Adding other language pivots might still be beneficial to obtain a more diverse set of distractors, however, more human feedback would be required to identify challenging candidates.

### 5.3 Distractor Diversity for Related vs. Unrelated Language Pivots

Section 5.1 has shown that the union of 16 pivot systems generates a diverse set of distractors. However, some of the pivots are more closely related than others. Here, we verify the claim that languages that are more closely related, tend to contribute similar distractors, whereas unrelated languages generate more diverse distractors. If this is true, this would also support the idea of customizing distractors to the native language of the learner.

We identify several pairs of most closely related languages among the 16 pivots used: French and Italian; Urdu and Hindi; Italian and Spanish; German and Dutch; Czech and Russian. For each language pair, we compute the gold retrieval score using the union of the candidates that the pivot pair generates. Let the first and second pivot in the pair be $r_1$ and $r_2$, respectively. We then identify for each pair another pivot $u_1$ that is unrelated to $r_1$, and compute gold retrieval score for the union of $r_1$ and $u_1$. We then compare the retrieval scores for the union of $r_1$ and $r_2$, and for the union of $r_1$ and $u_1$.

We compute the gold distractor retrieval for each group using the 30.30 setting. Since each language

[6] For the 30.30 setting, we use a beam size of 30. For 1.900 and 900.1 directions, sampling with a beam size of 40 is used.

| |
|---|
| **Sentence**: *We paid the lawyer to _____ up a totally new will.*<br>**Target word**: *draw*; **candidate**: *realize*; **semantic similarity**: yes |
| **Sentence**: *Due to the fact you weren't listening , you understood _____ nothing of what I said.*<br>**Target word**: *virtually*; **candidate**: *barely*; **semantic similarity**: yes |
| **Sentence**: *Despite past good performances , the actor was fired when the studio decided he had become a _____ .*<br>**Target word**: *liability*; **candidate**: *decision*; **semantic similarity**: no |
| **Sentence**: *It was the child's history teacher that first realised she was being _____ at home.*<br>**Target word**: *neglected*; **candidate**: *aware*; **semantic similarity**: no |

Table 5: Examples of distractors with and without semantic similarity to the sentence context and the target word.

| Setting | Gold retrieval | Avg. number of cands |
|---|---|---|
| 30.30 | 191 (44.8%) | 234 |
| 1.900 | 250 (58.7%) | 946 |
| 900.1 | 276 (64.8%) | 868 |

Table 6: Gold retrieval results and the average number of candidates per question, when using a different number of hypotheses in each direction, for a total of 900 back-translations in all settings.

| Setting | Valid candidates (%) |
|---|---|
| 30.30 | 85.5 |
| 1.900 | 88.6 |
| 900.1 | 91.1 |

Table 7: Percentage of valid distractors by direction setting. Averaged over 6 languages and 3 annotators.

produces a different number of gold distractors, for a fair comparison, we select a $u_1$, such that the gold retrieval score of $u_1$ on its own is the same as or close to the score of $r_2$. Our hypothesis is that since $r_1$ and $u_1$ are unrelated, their candidates should have less of an overlap than the candidates of $r_1$ and $r_2$. Therefore, the gold retrieval score of the union of $r_1$ with an unrelated language should be higher than for the union of $r_1$ and $r_2$. Indeed, we confirm our hypothesis in Table 9.

We further analyze the distractors proposed by various pivots and find that 52/191 gold distractors in the 30.30 setting (27%) are proposed by a single pivot and not proposed by the other 15 pivots.

### 5.4 Comparison with baseline methods

Our earlier study (Panda et al., 2022) compared the round-trip MT against word2vec and BERT, two approaches that showed competitive results for distractor generation (Mikolov et al., 2013; Gao et al., 2020). Table 10 shows gold distractor retrieval for the three methods when generating the same number of candidates (51) with each method. Table 11 shows the percentage of valid distractors among the proposed candidates for each method, demonstrating the superiority of the MT approach over word2vec and BERT. Further, neither word2vec nor BERT are effective at ranking the candidates, because word2vec and BERT tend to prefer words that are synonymous with the target and thus fit the context. Appendix B provides more detail on the two baseline methods and how comparisons are performed.

| Pivot group | Perc. (%) of cands. with semantic similarity to the target/context | | |
|---|---|---|---|
| | **1.900** | **900.1** | **30.30** |
| Group 1 | 32.4 | 43.8 | **62.0** |
| Group 2 | 32.1 | 40.6 | **42.5** |
| Group 3 | 25.6 | **45.5** | 26.7 |
| Group 4 | 37.9 | **39.8** | 41.3 |

Table 8: Percentage of candidate distractors judged as semantically similar to the target word and the carrier sentence context. Results are shown by the hypothesis setting. Best result for each pivot group is in bold.

## 6 Study with Language Learners

To evaluate the difficulty of automatically generated distractors, we conduct a cloze exercise test with English learners. We use a pool of manually validated items from the 30.30 setting and the pivots in group 1 to create a cloze test for participants. Manual validation ensured that all of the automatically generated candidates are valid. We sample 32 exercise items uniformly at random from the pool.

**Participants** Our participants are adult non-native English speakers of diverse language backgrounds. To ensure that the participants are advanced learners, we asked them to provide their

| Related | Pivots | Gold retrieval |
|---|---|---|
| Yes | Italian (60), French (66) | 85 |
| No | Italian (60), Chuukese (66) | **94** |
| Yes | Urdu (62), Hindi (59) | 90 |
| No | Urdu (62), Czech (58) | **98** |
| Yes | Spanish (38), Italian (60) | 68 |
| No | Spanish (38), Hindi (59) | **78** |
| No | Spanish (38), Russian (62) | 77 |
| Yes | Dutch (42), German (57) | 74 |
| Yes | Dutch (42), Czech (58) | 76 |
| No | Dutch (42), Urdu (57) | **84** |
| Yes | Czech (58), Russian (62) | 88 |
| No | Czech (58) Urdu (62) | **98** |

Table 9: Gold distractor retrieval for related and unrelated pivots. Best result for each comparison is in bold.

| Gold distractors retrieved | | |
|---|---|---|
| **Word2vec** | **BERT** | **MT** |
| 39 (9.2%) | 97 (22.8%) | **136** (31.9%) |

Table 10: **Word2vec** vs. **BERT** vs. **round-trip MT**: Number of gold distractors retrieved.

TOEFL or IELTS scores. We also gave them a sample test to complete to exclude those whose English was too good or not good enough. Participants were informed that the results of their tests would be used to collect statistics for research, without disclosing personal information. Participants were provided with $25 gift cards.

**Cloze exercise setup** We create two versions of a cloze test with the same set of 32 carrier sentences. Each version contains 16 sentences with gold distractors and 16 sentences with automatic distractors. The sentences that come with gold distractors in the first version, come with automatic distractors in the second version of the test, and vice versa. The order of the cloze items in each version is randomized. Additionally, we ensure that for each item the target always appears in the same position with both gold and automatic distractors on the multiple-choice list.

Each version of the test was completed by exactly 16 participants, so each cloze item was completed by 16 learners who were given gold distractors, and by another group of 16 learners who received automatic distractors. We use the first 2 cloze items as training items, to help the test takers familiarize themselves with the task. The statistics are computed using the remaining 30 cloze items. These remaining 30 cloze items contain an equal number (15) of items with gold distractors and automatic distractors.

| Method | % of valid distractors | | | | Gold distr. retrieved |
|---|---|---|---|---|---|
| | R1 | R2 | R3 | Avg. | |
| MT | **67.9** | **73.5** | **75.4** | **72.3** | 16 (3.8%) |
| Word2vec | 57.2 | 48.7 | 62.4 | 56.1 | 23 (5.4%) |
| BERT | 22.7 | 46.3 | 45.1 | 38.0 | 24 (5.6%) |
| MT (word2vec rank.) | 50.4 | 47.1 | 52.1 | 49.9 | 47 (11.0%) |
| MT (BERT rank.) | 27.7 | 41.8 | 55.4 | 41.6 | 36 (8.5%) |

Table 11: Percentage of valid distractors in the top-5 list by rater and distractor generation method. The last column shows the number and percentage of the gold distractors in the top-5 list.

We set up the test in a user interface setting, where a participant can see the carrier sentence and the four choices on the screen and has to pick one choice. As part of the test instructions, the participants were asked not to leave the response blank. We asked the participants not to get help from external resources to solve the exercise. The participants took between 20 to 30 minutes to complete the test.

**Paired t-test** A paired t-test was used to compare the human performance on cloze items with gold and automatic distractors. For computing the paired t-test statistics, we use the 30 cloze items that were not used as training items, and compare scores of gold vs. automatic distractors used, where the *score* is defined as proportion of participants that correctly solved the item. There was no significant difference in the scores of gold distractors (with mean 9.57, standard deviation 3.83) and automatic distractors (with mean 10.23, standard deviation 3.47). The two-tailed P value is 0.2884. These results suggest that the scores on cloze items using gold distractors and automatic distractors are not significantly different. Specifically, our results show that *when automatic distractors are used in the cloze items instead of gold distractors, the difficulty of the cloze items remains the same*.

## 7 Conclusion

We present a novel approach to generate challenging distractors for cloze exercises with round-trip neural MT. We show that using multiple pivot systems and a large set of round-trip translations produces diverse candidates, and each pivot contributes unique distractors. The latter opens up a possibility of customizing the cloze task for speakers of different languages, by tying the pivot

choice to the learner's native language, an interesting promise that BERT-based and other models cannot do. We conduct a thorough evaluation of the distractors, using a set of real cloze exercises for advanced ESL learners. Finally, we conduct a study with language learners that demonstrates that the automatic distractors produced with our approach result in cloze items of the same difficulty as those that use gold distractors. For future work, we will focus on customizing distractors based on the learner's native language, by prioritizing that language as pivot for MT.

## Limitations

A qualitative analysis of distractors generated via MT shows that this method can produce some inadequate candidates (and so do word2vec and BERT-based methods). Thus, a *human-in-the-loop* is needed to ensure the validity of the generated distractors. However, human-in-the-loop is standard practice, when producing language exercises and tests (Attali et al., 2022). We therefore believe that the proposed approach does not need to be fully automatic to be useful, as it can still help speed up distractor generation to create advanced vocabulary exercises. The MT method can thus be of huge help to human test developers.

The MT approach can be computationally more expensive than the methods proposed in prior work such as BERT and word2vec. Although we make use of pre-trained MT systems, the approach can be still costly, as it requires running two MT systems (forward and backward) with each pivot, and a BERT-based word alignment model to align the carrier sentence with each of its 900 back-translations. In terms of cost comparison, it takes 1-2 hours in a single Nvidia Tesla A100 GPU to generate 900 translations and produce candidate distractors for a single pivot, versus 0.5 hour with BERT and word2vec. However, the MT approach can potentially offer advantages that other methods cannot, such as producing a more diverse pool of distractors and, importantly, relating the native language of the learner to the pivot systems used to produce distractors. As our analyses show, each pivot system generates unique distractors. We stress that, while we show that using multiple pivots generates diverse distractors, we leave the question of whether using a pivot based on learner's first language is useful, to future work. We do hypothesize, however, that using pivots tied to the first language might be useful, however, but verifying this claim is left for future work. This is because verifying whether tying the pivot to learner's native language would be useful would require a human study with a relatively large group of learners of at least 20-30 students (all of advanced level) that all share the same first language. In fact, we would need to have several groups of learners, such that students in each group have the same first language background. This would be a large-scale study that is out of the scope of the paper. Note that the current work already presents a human study with 32 students that demonstrates that the automatically generated pivots are of the same difficulty as those created manually.

We also note that the method requires relatively good MT systems for generating more difficult distractors. Finally, our study is limited to cloze items that include single words as targets and does not consider fixed expressions, such as phrasal verbs and idioms. In the language testing community, such expressions are typically tested separately from the generic cloze items. The basic approach is to detect them before the carrier sentence is cleared to be used for cloze exercises. Our current work is not focused on carrier sentence selection. But it makes sense to include this consideration in a larger suite of tools for cloze item generation.

## Acknowledgments

The authors would like to thank the anonymous ARR reviewers for their insightful comments. This work was partly supported by the PSC-CUNY grant 64487-00 52.

## References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210. Association for Computational Linguistics.

Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary

assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Lingyu Gao, Kevin Gimpel, and Arnar Jensson. 2020. Distractor analysis and selection for multiple-choice cloze questions for second-language learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal*, 2(3):210–224.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30, San Diego, CA. Association for Computational Linguistics.

Ayako Hoshino and Hiroshi Nakagawa. 2005. A real-time multiple-choice question generation for language testing: A preliminary study. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 17–20, Ann Arbor, Michigan. Association for Computational Linguistics.

Shu Jiang and John Lee. 2017. Distractor generation for Chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, Copenhagen, Denmark. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers.

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 *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 1–8, Ann Arbor, Michigan. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Subhadarshi Panda, Frank Palma Gomez, Michael Flor, and Alla Rozovskaya. 2022. Automatic generation of distractors for fill-in-the-blank exercises with round-trip neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 391–401, Dubln, Ireland. Association for Computational Linguistics.

Juan Pino and Maxine Eskénazi. 2009. Semi-automatic generation of cloze question distractors effect of students' l1. In *SLaTE*.

Masoud Jalili Sabet, Philipp Dufter, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*.

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. Association for Computational Linguistics.

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*, pages 61–68, Ann Arbor, Michigan. Association for Computational Linguistics.

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(1):15.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the

World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Chak Yan Yeung, John Lee, and Benjamin Tsou. 2019. Difficulty-aware distractor generation for gap-fill items. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 159–164, Sydney, Australia. Australasian Language Technology Association.

Torsten Zesch and Oren Melamud. 2014. Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, Baltimore, Maryland. Association for Computational Linguistics.

## Appendix A: Grouping Pivot Languages by Machine Translation Quality

**Using BLEU scores on Tatoeba dataset** To evaluate the contribution of the quality of MT systems to the problem of distractor generation, we use BLEU scores of the MT systems on the Tatoeba dataset (since Bislama and Chuukese are not part of Tatoeba, for these languages we report BLEU score results on the JW300 corpus for low-resource languages (Agić and Vulić, 2019)).

We then split the pivot languages into four groups, organized by the averaged BLEU scores. We assume higher BLEU scores correspond to back-translations of higher quality. Generally speaking, higher BLEU scores correspond to language pairs with more training data (high-resource), whereas lower scores correspond to language pairs that are low-resource. Table A1 shows the averaged number of parallel sentences per pivot group, supporting this claim. Although the training size varies by language, languages in group 1 have substantially more training data than languages in other groups. The number of parallel sentences is between 141-905M in group 1, 66-105M in group 2, 1.9K-126M in group 3, and 9.2-28M in group 4. Another factor that might be contributing to the BLEU score levels is the typological distance of the pivot and English (all languages in group 1 are Indo-European languages more closely related to English, compared to languages in other groups.)

| **Pivot group** | **Parallel corpus size (sents.)** |
|---|---|
| Group 1 | 392M |
| Group 2 | 85M |
| Group 3 | 63M |
| Group 4 | 18M |

Table A1: Number of parallel sentences per language pivot group (each group shows the number of parallel sentences averaged over the languages in that group).

**Using BLEU scores of the carrier sentences** Since BLEU is dependent on the n-grams in the reference, we also perform the following experiment:

1. Calculate the BLEU score for every carrier sentence and its 900 round-trip translations. We use the carrier sentence as the reference and the round-trip translation as the hypothesis.
2. Average the resulting BLEU scores to get the overall BLEU score for each language pair.

We find that the resulting BLEU scores are drastically small, ranging between 1.5 and 2.30, making it hard to provide a ranking between the language pairs. This is because lower-ranked hypotheses tend to diverge from the original sentence. We thus perform the same experiment by including only top 10 hypotheses. BLEU scores are slightly higher but still low. We obtained the following BLEU scores, averaged by language group: 6.9 (group 1); 6.4 (group 2); 5.0 (group 3); 2.2 (group 4).

While the averaged BLEU scores are all very small, they do support the ranking based on the BLEU scores on the Tatoeba dataset.

## Appendix B: Comparison with Other Approaches

Below, we compare the MT approach with word2vec and BERT, two methods that showed competitive results on the task of distractor generation. This comparison was carried out in our earlier study (Panda et al., 2022), and is presented here for convenience.

Using word2vec, candidate distractors are generated by producing a list of words that have the highest similarity to the target word. 300-dimensional word2vec embeddings trained on Google News are used. For a given target word, $k$ nearest neighboring words based on cosine similarity in the word embedding space are considered as candidates. With BERT, the carrier sentence is passed to the model, with the target word replaced by a masked token. BERT returns a list of words that best fit the context of the carrier sentence at the position of the masked token. Each word is asso-

| | Gold distractors retrieved | | |
|---|---|---|---|
| | **Word2vec** | **BERT** | **MT** |
| Before filt. | 66 (15.5%) | 144 (33.8%) | **154** (36.2%) |
| After filt. | 39 (9.2%) | 97 (22.8%) | **136** (31.9%) |

Table B2: **Word2vec** vs. **BERT** vs. **round-trip MT**: Number and percentage of gold distractors retrieved.

ciated with probability; top $k$ candidates with the highest scores are selected. The candidates are filtered out using the same filtering algorithm applied in round-trip MT (see Section 4).

**Comparing generated distractors with BERT and word2vec on gold distractor retrieval** Using word2vec and BERT, a list of $n$ nearest neighbors for each target word is generated. Since the round-trip MT method produces a different number of candidate distractors per target, whereas word2vec and BERT generate a long list of candidates, the average number of candidates produced with round-trip MT with the union of 5 pivot languages is used, to generate 104 neighbors without filtering and 51 neighbors with filtering applied. Results are shown in Table B2 before and after filtering is applied. Round-trip MT retrieves significantly more gold distractors compared to word2vec and BERT. Word2vec performs the worst among the three methods.

**Manual evaluation of distractor validity for the three methods** For each carrier sentence, 5 sets of automatically-generated distractors are compared: (1) round-trip MT (without ranking);[7] (2) round-trip MT with word2vec ranking; (3) round-trip MT with BERT ranking; (4) using word2vec for generation; (5) using BERT for generation. BERT and word2vec can be used to rank candidates produced with MT by using the semantic similarity of the candidate to the target. The most similar candidates would rank as the highest.

The manual evaluation was performed by three annotators who are college students and native English speakers. The annotators were presented with a carrier sentence, the target word, and the manually evaluated five sets of distractors. The annotator's task was to mark each distractor as valid or invalid. Results are presented in Table 11 in the main text and demonstrate that MT without ranking produces the highest percentage of valid candidates with all three annotators.

[7] Five distractors are selected uniformly at random.

| **Method** | **Annotators** | | | **Avg.** |
|---|---|---|---|---|
| | **1,2** | **1,3** | **2,3** | |
| MT-all-pivots | 0.805 | 0.816 | 0.861 | 0.827 |

Table C3: Pairwise agreement for the 3 annotators.

| **Group** | **Pivot language** | **Number of valid cands.** | **Cands. with seman. fit** |
|---|---|---|---|
| Group 1 | Spanish | 49 | 34 (69.4%) |
| | German | 54 | 31 (57.4%) |
| | Russian | 65 | 31 (47.7%) |
| | French | 59 | 29 (49.2%) |
| | Total | 227 | **125 (55.1%)** |
| Group 2 | Indonesian | 49 | 20 (40.8%) |
| | Vietnamese | 74 | 27 (36.5%) |
| | Total | 123 | **47 (38.2%)** |
| Group 3 | Chinese | 63 | 23 (36.5%) |
| | Malayalam | 72 | 16 (24.6%) |
| | Total | 135 | **39 (28.9%)** |
| Group 4 | Chuukese | 67 | 26 (38.8%) |
| | Hindi | 69 | 24 (34.8%) |
| | Total | 136 | **50 (36.8%)** |

Table D4: Number and percentage of candidate distractors judged as *semantically similar* to the target word and the carrier sentence context, among candidates considered as valid by all three raters. Using 30 translation hypotheses in each direction.

## Appendix C: Inter-Annotator Agreement

The annotators made a binary decision on each distractor, determining whether the distractor is valid. We compute pairwise agreement using Cohen kappa's (Cohen, 1960) and present the results in Table C3. Our average pairwise agreement values are shown in the last column. These values are better than those obtained by Yeung et al. (2019), although their annotation task included 3 classes. Cohen's kappa results indicate strong agreement in all cases. The numbers in the table indicate excellent agreement (Landis and Koch, 1977).

## Appendix D: Additional Results

**Manual evaluation of the difficulty of the automatic distractors by pivot group** Table D4 shows the number and percentage of candidate distractors that are judged as semantically similar to the target word and the carrier sentence.

**Varying the number of generated hypotheses by translation direction** Table D5 shows gold retrieval results by pivot group and the hypothesis setting. Performance of individual select pivots for the 3 hypothesis settings can be viewed in Figures D3 and D4.

| Setting | Pivot group | Gold retrieval | Avg. number of cands |
|---|---|---|---|
| 30.30 | 1 | 54 (12.7%) | 13 |
| | 2 | 44 (10.2%) | 16 |
| | 3 | 58 (13.6%) | 38 |
| | 4 | 62 (14.6%) | 50 |
| | All | 191 (44.8%) | 234 |
| 1.900 | 1 | 106 (24.9%) | 122 |
| | 2 | 110 (25.8%) | 130 |
| | 3 | 73 (17.1%) | 82 |
| | 4 | 87 (20.4%) | 138 |
| | All | 250 (58.7%) | 946 |
| 900.1 | 1 | 149 (35%) | 112 |
| | 2 | 144 (33.8%) | 125 |
| | 3 | 116 (27.2%) | 121 |
| | 4 | 96 (22.5%) | 152 |
| | All | 276 (64.8%) | 868 |

Table D5: Gold retrieval results and average number of candidate per question, when using a different number of hypotheses in each direction.

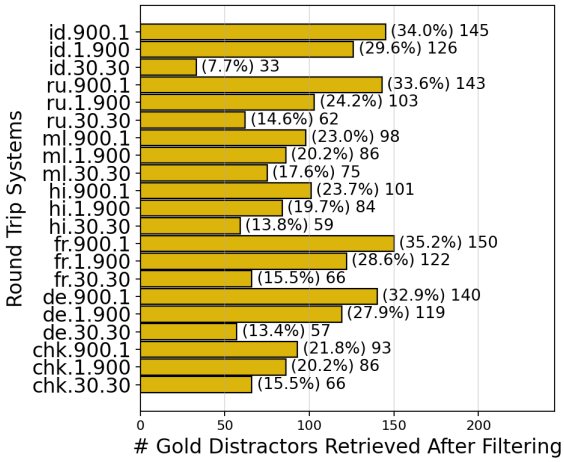


Figure D3: Gold retrieval results using 900 hypotheses with the different number of translations used in each direction.

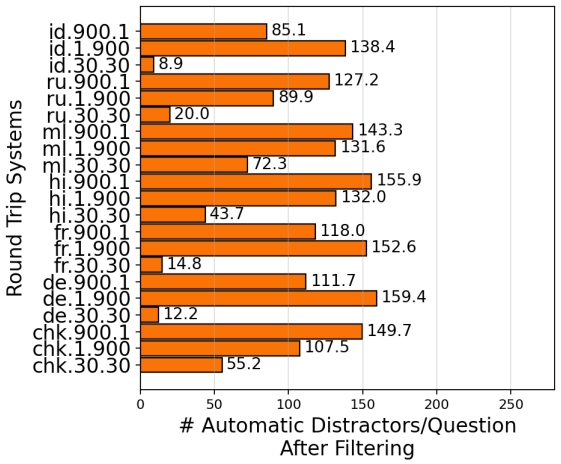


Figure D4: Average number of candidates generated per exercise item, using 900 hypotheses with the different number of hypotheses in each direction.