Automatic Generation of Distractors for Fill-in-the-Blank Exercises with Round-Trip Neural Machine Translation

Subhadarshi Panda
CUNY Graduate Center
spanda@gradcenter.cuny.edu

Frank Palma Gomez
Queens College, CUNY
frankpalma12@gmail.com

Michael Flor
Educational Testing Service
mflor@ets.org

Alla Rozovskaya
Queens College, CUNY
arozovskaya@qc.cuny.edu

Abstract

In a fill-in-the-blank exercise, a student is presented with a carrier sentence with one word hidden, and a multiple-choice list that includes the correct answer and several inappropriate options, called distractors. We propose to automatically generate distractors using round-trip neural machine translation: the carrier sentence is translated from English into another (pivot) language and back, and distractors are produced by aligning the original sentence and its round-trip translation. We show that using hundreds of translations for a given sentence allows us to generate a rich set of challenging distractors. Further, using multiple pivot languages produces a diverse set of candidates. The distractors are evaluated against a real corpus of cloze exercises and checked manually for validity. We demonstrate that the proposed method significantly outperforms two strong baselines.[1]

1 Introduction

A cloze (fill-in-the-blank) exercise is a common method of teaching vocabulary, as well as assessing non-native speaker performance in a foreign language: a passage (sentence) is presented to the learner with one word (*target*) being removed. The target word is presented along with a list of *distractors* (usually 3), and the task is to correctly 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 item is valid if and only if one word on the list (the target) fits the context. We also show valid and invalid distractors.

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

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.

A valid distractor is a word that does not fit the context. For example, "great" and "utmost" are invalid distractors, since they both fit the context.

Given a carrier sentence and the target word, the problem is to generate challenging distractors. In typical high-stakes tests, such as Test of English as a Foreign Language (TOEFL), distractors are generated manually by educational testing experts, a time-consuming procedure. An automated method to generate distractors would be extremely valuable. 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 grammatically unfit will be too easy for advanced speakers (Zesch and Melamud, 2014). To address this, previous work used context-sensitive inference rules (Zesch and Melamud, 2014), common collocation errors from large-scale learner corpora (Sakaguchi et al., 2013), co-occurrence likelihoods (Hill and Simha, 2016), and word embeddings (Jiang and Lee, 2017).

In this work, we propose to generate distractors using round-trip neural machine translation (MT). Word choice errors are commonly affected by the speaker's first language, and even advanced learn-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 391 - 401

ers struggle with word usage nuances and may inappropriately use semantically related words (Leacock et al., 2010). Our assumption is that lexical challenges 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.

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 top n translation hypotheses are generated. For each hypothesis, top m back-translations into English are generated. The back-translated words aligned to the target are treated as potential distractors. We use five round-trip MT systems and show that *using multiple pivot languages encourages diversity in the distractor generation*, as the distractors produced with different pivot language systems are often unique.

Using a corpus of cloze exercises for advanced ESL learners, we demonstrate that the proposed method retrieves over 31% of the gold distractors used in the exercises and over 70% percent of cloze items have at least one gold distractor retrieved with our approach. Evaluation shows that the proposed method outperforms two strong baselines – the word embeddings approach (Word2vec) and BERT. Manual evaluation of the distractor validity indicates that over 72.3% of all distractors are valid with our approach compared to 56.1% and 38.0% using Word2Vec and BERT, respectively.

Our contributions are as follows: (1) we propose to use round-trip machine translation to generate challenging distractors for cloze exercises and tests. We use hundreds of round-trip translations and multiple pivot languages, and generate challenging diverse distractors; (2) we validate our approach using a dataset of real cloze exercises for advanced ESL learners and show that it significantly outperforms the Word2vec and BERT baselines both in automatic and manual evaluation; (3) unlike previous work, we find that different pivot languages provide rather unique distractors for the same item, thereby allowing for customizing the exercises on the basis of the native language of the student.

The next section presents related work. Section 3 describes the dataset of cloze exercises. Section 4 describes the baseline methods, and Section 5 presents our approach. Section 6 presents the results of the automatic and manual evaluation of the generated distractors. Section 7 further discusses the results, while Section 8 concludes.

2 Related work

The general approach to automatic distractor generation can be broken down into candidate *generation* (identification), and candidate *ranking*.

Candidate generation Most of the work on automatic distractors focuses on generating distractor candidates. These include 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 picked 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 context, 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. Thus, on the one hand, challenging distractors should be *semantically close* to the target word, yet, on the other hand, a valid distractor *should not produce an acceptable sentence*.

Most of the approaches to generating challenging distractors rely on methods of semantic relatedness, such as 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 is quite limited, as both the choice of the target word and of the distractors are constrained by the errors in the corpus. Several recent studies showed that word embeddings are effective in distractor generation: Jiang and Lee (2017) and Susanti et al. (2018) generated distractors using semantically similar words obtained from Word2vec (Mikolov et al., 2013).

We propose to use round-trip neural machine translation to generate distractors. The only previ-

ous mention of using MT is that of Dahlmeier and Ng (2011) who aim at correcting ESL collocation errors using a statistical machine translation technique. To the best of our knowledge, ours is the first dedicated study that uses state-of-the-art NMT systems with 5 pivot languages and large sets of back-translations for generating distractors.

Several studies, while they do not generate distractors, address the complexity of the cloze task for language learners. Felice and Buttery (2019) focus on the contextual complexity of the generated gap itself. Marrese-Taylor et al. (2018) use LSTM models for gap generation. Gao et al. (2020) show that BERT is helpful in measuring the fit of the distractor in the context, and thus can be used for estimating distractor difficulty. Finally, we also note that there is a significant body of work on a task of generating reading comprehension (RC) items, that test a different set of examinee abilities, such as inference. That work (Chung et al., 2020) deals with generating phrases and complete sentences for distractors. RC item generation is a distinct problem from vocabulary item generation that is addressed in this work.

Candidate ranking can be used as an additional step to (re-)rank the candidates produced during candidate generation. One reason for this is that context is typically not taken into account when generating candidates. Yeung et al. (2019) used BERT (Devlin et al., 2018) to re-rank the candidate distractors generated with Word2vec for Chinese. We show that BERT is not effective at generating or re-ranking candidate distractors.

3 Data

It is important to note that there is no benchmark dataset for the task. Previous studies evaluate either on artificially created items with random words as targets or proprietary data. In contrast, 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]

Since we wish to generate distractors for advanced learners, we use the C1 advanced level multiple choice cloze exercises.[4] C1 level is part of CEFR scale.[5] It is used to prove high-level achievement in English and is designed for learners preparing for university or professional life.

We extract a total of 142 cloze items.[6] 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. We show two sample items in Table 2. 44.4% of the target words are verbs, 38.7% are nouns, 14.1% are adjectives, and 2.8% are some other part of speech.

4 The Baselines

We compare the round-trip MT method against Word2vec and BERT. Both Word2vec embeddings and BERT can be used *to generate* candidates, and *to rank* candidates generated with MT. Here, we describe how we generate candidates with Word2vec and BERT. In Section 5.3, we describe how we use the two methods for candidate ranking. Using Word2vec, we generate words that have the highest similarity to the target word and use these as potential distractors. We use the 300-dimensional Word2vec embeddings trained on Google News. For a given target word, we find k nearest neighboring words using cosine similarity in the word embedding space. With BERT, we produce a set of candidates by passing the carrier sentence 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 associated with probability; we select the top k candidates with the highest scores. The candidates are filtered out using the same filtering algorithm applied in round-trip MT (see Section 5.2). In addition, we filter out misspellings by using a wordlist of about 130,000 English word-forms.

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

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

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

[5] https://www.coe.int/en/web/common-european-framework-reference-/languages/level-descriptions

[6] Our data collection is in conformity with the website's terms as described at https://www.esl-lounge.com/student/copyright.php.

Sentence: *Much of the neighbourhood was demolished in the 1940s when living _____ had deteriorated.* **Choices**: *situations, conditions*, circumstances, states*
Sentence: *Scientists are yet to understand the full nutritional _____ of the humble olive.* **Choices**: *favours, helps, goods, benefits**

Table 2: Examples of multiple choice cloze exercises from the ESL Lounge website. Each item has exactly one correct choice, marked with a star (*).

5 Generating Distractors with Neural MT

Formally, given a sentence $X = \{x_1, x_2, \ldots, x_n\}$ and a position $k \in [1, n]$ of the target word, the task is to generate a set of candidate distractors D such that $d \in D$ can be used as a challenging semantically-confusing distractor for the target word occupying position k in X. Since challenging distractors should be more similar to the target word (Zesch and Melamud, 2014), and because many word sense nuances are challenging for non-native speakers due to the differences between word usage in their native language and in English, we expect that candidates generated with round-trip MT that uses the target word together with the surrounding context will make good distractors for advanced ESL learners.

5.1 Candidate generation

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 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 in 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 $X_{RT} = \{X_{RT_1}, X_{RT_2}, \ldots, X_{RT_{N_f \times N_b}}\}$.

We use state-of-the-art NMT systems with German, Russian, Italian, French, and Czech as pivots. For German and Russian, we use the systems of Ng et al. (2019), and for the other languages we use the systems of Tiedemann and Thottingal (2020). We use $N_f = 1$, $N_b = 1,500$ for German, $N_f = 1, N_b = 1,000$ for Russian, and $N_f = N_b = 16$ for the other languages, and generate $1,500$ round-trip translations for German, $1,000$ for Russian, and 256 for Italian, French, and Czech. The number of hypotheses varies due to system specifications as well as the memory constraints in the machines we used. We do not attempt at comparing the machine translation models with various pivot languages and leave it for future work.

Alignment computation Given a round-trip translation X_{RT_i} for carrier sentence X, we need to compute the alignment between the two sentences. Then 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.We use Simalign[7] (Sabet et al., 2020) that employs contextual word embeddings (Devlin et al., 2018) to produce an alignment model for a pair of sentences in the same or different language, without parallel training data.

Given the original sentence and the round-trip translation, first the similarity between each source token is computed with each target token using contextual embeddings from multilingual BERT. This results in a matrix that stores similarity scores between all the source and target tokens. The alignment computation is framed as an alignment problem where we search for a maximum-weight maximal matching in the bipartite weighted graph induced by the similarity matrix (see details in Sabet et al. (2020)).

5.2 Candidate filtering

Not all the words obtained by alignment can serve as distractors because (a) the candidate might fit the context, which would make the item invalid, or (b) a word may make the sentence grammatically incorrect and thus too easy for advanced students. We use two filtering mechanisms.

Filtering distractors that are synonymous with the target We use the synonyms provided in WordNet (Fellbaum, 1998) to determine the candidate words that are synonymous with the target word. We note that this approach will not weed out distractors that are synonymous in specific contexts. For example, in the sentence *Though we always turn right here, I often _____ what's down*

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

the other road. with the target "wonder", the algorithm generates "think" as a candidate distractor. Although "think" and "wonder" are not synonyms, they are equivalent in the context of the sentence.

Filtering distractors based on POS tag An obvious approach to filter out grammatically inappropriate distractors is to ensure that the candidate word is of the same part-of-speech as the target word in the carrier sentence. We use NLTK (Bird et al., 2009) to compute the POS tag for the candidate words and only keep those which have the same part-of-speech as the target word. Both for the target word and the distractor candidates, the POS tag is obtained by applying the tagger to the entire carrier sentence with the target position filled by the appropriate word.

5.3 Candidate ranking with BERT and word2Vec

Typically, fewer than 5 distractors are used in a cloze exercise, however, as we show below, the MT method typically generates more than 5 candidates. One approach to selecting distractors from the available pool is uniformly at random. However, previous studies typically rank candidates based on their difficulty, assumed to be related to the degree of semantic similarity to the target. We thus wish to determine whether we can use Word2vec and BERT to rank the distractors instead of simply selecting candidates uniformly at random.

Using Word2vec, we define the difficulty of a candidate distractor d for sentence X with target t as the cosine similarity of their word embeddings as in Equation 1:

$$\text{difficulty}(d,t) = \frac{Emb(d) \cdot Emb(t)}{|Emb(d)||Emb(t)|} \quad (1)$$

The $Emb(w)$ is a pre-trained embedding for word w. We use the 300 dimensional Word2vec embeddings trained on Google news (Mikolov et al., 2013). We pick candidates with the highest similarity values. Similarly, we rank the candidates using the scores obtained with BERT.

6 Evaluation

We evaluate the generated distractors using both automatic and manual evaluation.

6.1 Automatic evaluation

Number of distractors generated We first show the average number of unique candidate distractors retrieved with each pivot language system and with the union of all the pivot systems, with and without filtering (Figure 1). The number of unique distractors is smaller than the total number of back-translated sentences since many of the hypotheses result in the same round-trip translation of the target word. The smallest average number of distractors is 18.1 for Italian, and the largest average number is 51.8 for German, when no filtering is used. Notably, the union produces an average of 104.6 distractors per target word, suggesting that round-trip translations from different pivot languages contribute unique distractor candidates. Filtering removes a significant number of generated candidates by reducing the average number of candidates from 104.6 to 51.1 for the union.

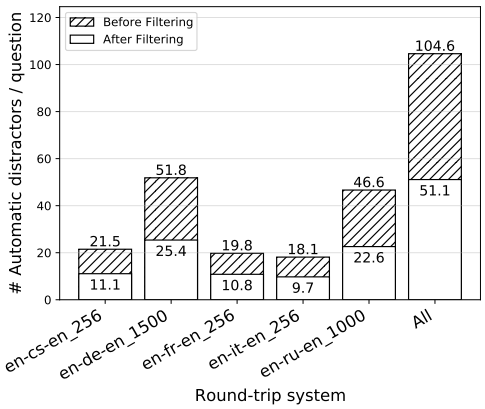

Figure 1: Average number of automatic distractors generated per cloze item using different pivots before and after filtering. The average is computed over 142 cloze items.

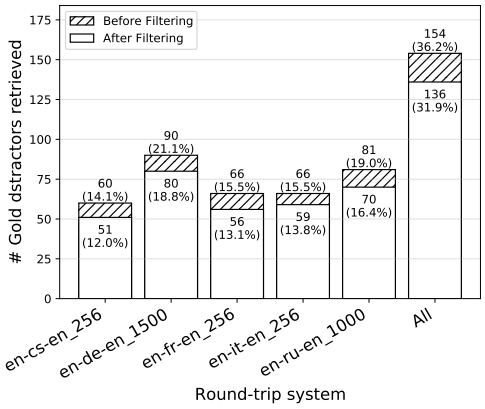

Figure 2: The number and percentage of gold distractors retrieved as a function of round-trip translations used, before and after filtering.

Gold distractor retrieval While there may be many valid challenging distractors for a given ex-

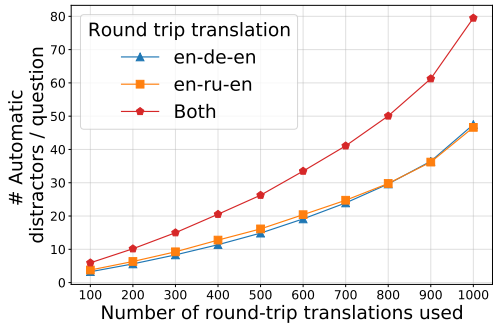

Figure 3: Average number of automatic distractors per item as a function of the number of round-trip translations used. The average is computed over 142 cloze items.

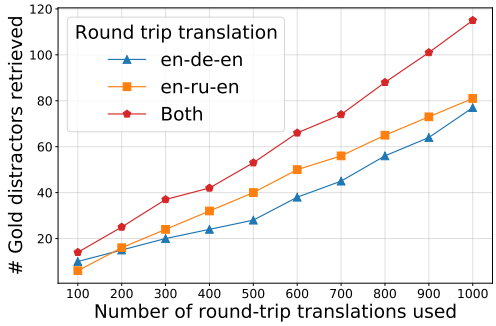

Figure 4: Number of gold distractors retrieved as a function of round-trip translations used.

ercise item, we nevertheless wish to evaluate the distractors generated automatically against the set of gold distractors (distractors used in the cloze items in the dataset). 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 following Equation 2.

$$r(d, D_{gold}) = \begin{cases} 1 \text{ if } d \in D_{gold} \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

We compute cumulative retrieval score[8] $\sum r(d, D_{gold})$ across all the generated distractors and 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 language and for the union of all pivot languages before and after filtering is applied: 36.2% of gold distractors are retrieved with the automatic approach (without filtering). Filtering reduces this number to 31.9%, however, as we showed above, filtering removes about 50% of the generated candidates. We also note that by-pivot performance is surprisingly consistent: for German and Russian, we retrieve 21.1% and 19.0% of gold distractors, and for the other pivots – between 14.1% and 15.5%. We attribute the differences between the first and second group to the number of round-trip translations we generate (1,000 and 1,500 for Russian and German, respectively, and 256 for the other pivots). Importantly, the union of the pivot languages is able to retrieve almost twice as many gold distractors as the individual languages, indicating that *multiple pivots produce diverse candidate distractors*.

We stress that, while the distractors are not uniquely defined, it is encouraging that over 30% of gold distractors are retrieved with our approach.

Gold distractor retrieval as a function of the number of round-trip translations Next, we evaluate how increasing the size of the round-trip translations affects the number of distractors generated, and whether it improves gold distractor retrieval. We use 2 pivot languages, German and Russian, since we generate a large number of translations with these pivots. We limit the number of round-trip translations to 1,000 since this is the maximum number of translations we can generate with the Russian pivot. These NMT models also have similar implementations, which would allow for a fair cross-pivot comparison. We use $N_f = 1$ in all cases, and vary N_b between 100 and 1,000.

Figure 3 shows that *the average number of distractors generated per item* increases with the number of round-trip translations. With 100 hypotheses, fewer than 5 candidates are generated with each pivot, but this number increases to around 50 when 1,000 are used. Interestingly, the number of candidates for each pivot is almost the same, but the union of the pivots generates almost twice as many candidates indicating that the pivots generate non-overlapping candidates.

While the number of candidates increases with the number of round-trip translations used, it is not obvious if the lower-ranked hypotheses are useful or they simply generate noise. Figure 4 shows the gold retrieval scores as a function of the number of translations. Both systems behave similarly in terms of the number of gold distractors retrieved, and the retrieval score continues to increase as the

[8] We do not evaluate precision here, as the set of potential valid distractors is not unique, and candidates that are not in the gold set can also serve as valid distractors, so precision cannot be computed in automatic evaluation.

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

Method	**% of valid distractors**				**Gold distr.**
	R1	**R2**	**R3**	**Avg.**	**retrieved**
MT-no-ranking	**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 4: 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.

number of translations goes up. For example, with 200 round-trip translations, each language generates around 15 gold distractors among its candidates, and this number increases linearly, to almost 80 when $1,000$ translations are used. This suggests that lower-ranked hypotheses are still very useful. Furthermore, the information produced by each pivot system is complementary: *the union of the pivots retrieves almost twice as many gold distractors as the individual languages*. This motivates the use of multiple round-trip translation systems.

Finally, Figure 5 shows the percentage of cloze items for which at least $x \in \{1, 2, 3\}$ gold distractors were retrieved for the German and Russian round-trip translations. For both pivots, when using $1,000$ translations, less than 5% of cloze items have all 3 distractors retrieved. However, at least 1 gold distractor is retrieved in around 40 % of the cloze items. With the union of the two pivots, we retrieve at least 1 gold distractor for about 55% of the items, which, again, demonstrates that using multiple pivots introduces diversity and provides complementary information. We also find that some of the distractors might be more difficult to retrieve using the MT approach, as discussed further in Section 7.

Comparing generated distractors with BERT and Word2vec Using Word2vec and BERT, we generate a list of n nearest neighbors for each target word. 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, we use the average number of candidates produced with round-trip MT with the union of 5 pivot languages, and generate 104 neighbors without filtering and 51 neighbors with filtering applied. Table 3 shows the results. Round-trip MT retrieves significantly more gold distractors compared to Word2vec and BERT.

Method	**Annotators**			**Avg.**
	1,2	**1,3**	**2,3**	
MT-no-ranking	0.573	0.619	0.590	0.594
Word2vec	0.379	0.389	0.624	0.463
BERT	0.294	0.705	0.364	0.454
MT (Word2vec rank.)	0.496	0.476	0.696	0.556
MT (BERT rank.)	0.439	0.495	0.413	0.449

Table 5: Pairwise agreement for the 3 annotators.

6.2 Manual evaluation of item validity

Evaluation of the item validity needs to ensure that the distractors cannot be used in the carrier sentence (see Table 1). Many invalid examples involve contextual synonyms that have not been filtered out with WordNet, as well as other, non-synonymous candidates that simply fit the context.

For each carrier sentence, we compare 5 sets of automatically-generated distractors:[9] (1) round-trip MT (without ranking);[10] (2) round-trip MT with Word2vec ranking; (3) round-trip MT with BERT ranking; (4) using Word2vec for generation; (5) using BERT for generation.

The manual evaluation is performed by three annotators who are college students and native English speakers. The annotators were presented with a carrier sentence, the target, and manually evaluated 5 sets of distractors by marking each distractor as valid or invalid.

We obtain the "precision" of each method, i.e. the percentage of the distractors judged as valid (Table 4). MT without ranking produces the highest percentage of valid candidates with all three annotators. On average , 72.3% of candidates are valid for MT without ranking, vs. 56.1% with Word2vec and 38.0% with BERT. Using BERT and word2Vec for ranking reduces the percentage of valid candidates in the top-5 list. The last column shows the retrieval scores for the top-5 list. Interestingly, BERT and word2Vec retrieve more gold candidates than the MT method, however, the proportion of invalid candidates is much higher for these methods, pos-

[9]The number of candidates is set to 5 because in a typical setting one would need to use 3 distractors for creating the exercises, and some of the automatic distractors would turn out to be invalid.

[10]5 distractors are selected uniformly at random.

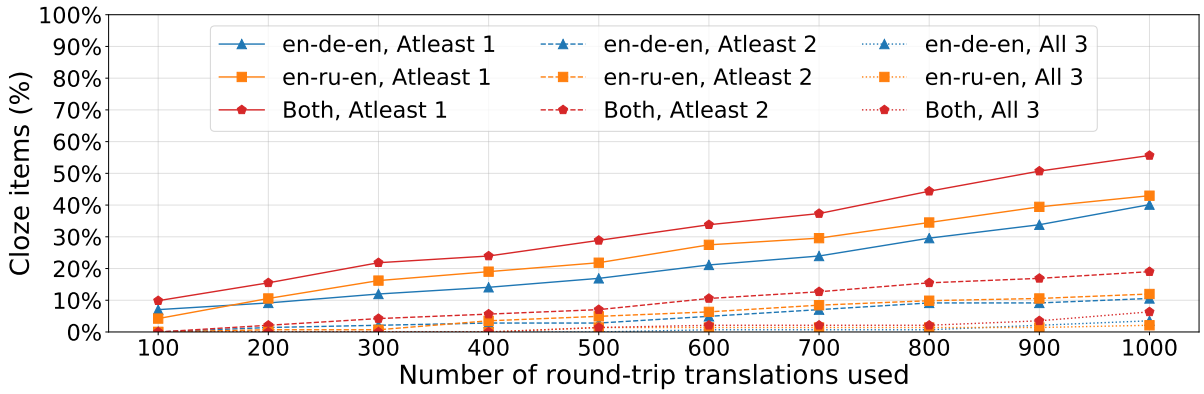

Figure 5: Percentage of cloze items with at least 1, 2, and 3 (all) gold distractors retrieved as a function of the number of round-trip translations used.

Sentence: *When choosing for this role, don't _____ the talents of Brian, one of the best actors in the academy.* **Choices**: *overlook*, overvalue, oversee, overrate*
Sentence: *You simply must invite Carol to the party. She's always the life and _____ of any evening.* **Choices**: *light, soul*, blood, flesh*

Table 6: Examples of multiple choice cloze exercises where none of the gold distractors were identified with the round-trip NMT approach. Each item has exactly one correct choice, marked with a star (*).

sibly, due to the higher proportion of synonyms of unrelated words that fit the sentence context.

Overall, manual evaluation demonstrates the superiority of the MT approach over Word2vec and BERT. We also find that neither Word2vec nor BERT are effective at ranking the candidates. With Word2vec, we conjecture this is due to the nature of the word embedding models that tend to prefer words that are not simply semantically similar but also synonymous with the target. Similarly, BERT is good at producing words that are most likely in the context of the carrier sentence.

Inter-annotator agreement We compute pairwise agreement using Cohen kappa's (Cohen, 1960) and present the results in Table 5. 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 moderate agreement in all cases.

7 Analysis and Discussion

We further analyze the distractors generated with round-trip MT. First, we examine the gold distractors that have not been identified with the MT approach. We find that some gold distractors are not semantically close to the target. Table 6 shows two such examples. In the first sentence, the gold distractors are based on morphology/phonology (common prefix), while in the second sentence, the distractors ("light", "blood", and "flesh"), arguably, are not semantically close to the target "soul".

Next, we focus on the differences between the distractors generated with Word2vec, BERT, and MT, and show an example that demonstrates the ability of round-trip MT to model sentential context. First example in Table 7 illustrates that Word2vec distractors are independent of the context of the sentence: the distractors are all latched on the "music" sense of the target word "band". However, round-trip MT models the context of the complete sentence and generates more appropriate distractors. The second example compares BERT-generated and MT-generated distractors: while not all of the MT distractors are valid, BERT is more likely to generate candidates that are synonymous with the target, and thus are invalid as distractors. In fact, Zhou et al. (2019) successfully use BERT for the task of lexical substitution, while Qiang et al. (2020) use BERT for lexical simplification. The idea of using BERT in such tasks is to provide good substitutes that are close synonyms in the given context. This is precisely the opposite of our goal: difficult distractors for a gap-filling task should not be substitutes of the target word.

Finally, the example below demonstrates that MT systems are capable of generating unique pivot-dependent distractors. Consider the carrier sentence "Despite being such a *frequent* visitor to Paris, Sam never bored of exploring it." with the

Sentence: *The _____ of thieves had been captured.* **Target word**: *band*; **gold distractors**: *bunch, crew, range* **Top-5 word2vec distractors**: *keyboardist, vocalist, drummer, quintet, guitarist* **Round-trip MT distractors**: *crew, group, orchestra, gang, squad*
Sentence: *The _____ of the report have yet to be analysed by the government so they can formulate new policies.* **Target word**: *findings*; **gold distractors**: *invenstions, discoveries, rulings* **Top-5 BERT distractors**: *recommendations, assertions, observations, results, conclusions* **Round-trip MT distractors**: *outcomes, familiarities, shows, results, achievements*

Table 7: Word2vec and BERT distractors vs. round-trip MT distractors.

target word "frequent" the French system generates "usual" as a distractor, while the Russian system does not. We believe this might be related to the fact that one of the translations of "frequent" into French is "habituel", which also has a meaning of "usual", and thus "usual" can be produced as a round-trip translation with the French pivot. This is not the case for Russian.

8 Conclusion

We present a novel approach to generating challenging distractors for cloze exercises using round-trip neural machine translation. We show that using multiple pivot systems and a large set of round-trip translations produces diverse candidates, and each pivot language contributes unique distractors. This opens up a possibility of customizing the cloze generation task for speakers of different languages (groups), an interesting promise that BERT-based and other models cannot do. We conducted a thorough evaluation of the distractors, using a set of real cloze exercises for advanced ESL learners. Comparison with Word2vec and BERT showed that the round-trip MT retrieves substantially more gold distractors given the same size of the candidate set.

For future work, we will focus on customizing distractors based on the learner's native language, by prioritizing that language as pivot for MT. We will also conduct a study with language learners to determine whether the automatic distractors produced with our approach result in cloze items of the same difficulty as those that use gold distractors.

For the current work for English, we used high-quality machine translation systems. However, for many language pairs that do not include English as one of the languages, high-quality MT systems are not available. Further, high-quality MT systems are also rarely available for low-resource languages paired with English. The future work will also focus in determining whether and how translation quality might affect the quality of generated distractors. We hypothesize that the proposed method might require special approaches when used to develop exercises for languages other than English and when generating English distractors using low-resource pivots. This is another exciting direction for future work.

Acknowledgments

We thank the anonymous reviewers for their insightful comments.

References

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.

Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In *EMNLP*. Association for Computational Linguistics.

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

Daniel Dahlmeier and Hwee Tou Ng. 2011. Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

Mariano Felice and Paula Buttery. 2019. Entropy as a proxy for gap complexity in open cloze tests. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. INCOMA Ltd.

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.

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.

Edison Marrese-Taylor, Ai Nakajima, Yutaka Matsuo, and Ono Yuichi. 2018. Learning to automatically generate fill-in-the-blank quizzes. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*.

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.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

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

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*. Association for the Advancement of Artificial Intelligence.

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 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.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.