Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners

Keisuke Sakaguchi[1*] **Yuki Arase**[2] **Mamoru Komachi**[1†]
[1]Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
[2]Microsoft Research Asia
Bldg.2, No. 5 Danling St., Haidian Dist., Beijing, P. R. China
{keisuke-sa, komachi}@is.naist.jp, yukiar@microsoft.com

Abstract

We propose discriminative methods to generate semantic distractors of fill-in-the-blank quiz for language learners using a large-scale language learners' corpus. Unlike previous studies, the proposed methods aim at satisfying both *reliability* and *validity* of generated distractors; distractors should be exclusive against answers to avoid multiple answers in one quiz, and distractors should discriminate learners' proficiency. Detailed user evaluation with 3 native and 23 non-native speakers of English shows that our methods achieve better reliability and validity than previous methods.

1 Introduction

Fill-in-the-blank is a popular style used for evaluating proficiency of language learners, from homework to official tests, such as TOEIC[1] and TOEFL[2]. As shown in Figure 1, a quiz is composed of 4 parts; (1) sentence, (2) blank to fill in, (3) correct answer, and (4) *distractors* (incorrect options). However, it is not easy to come up with appropriate distractors without rich experience in language education. There are two major requirements that distractors should satisfy: *reliability* and *validity* (Alderson et al., 1995). First, distractors should be *reliable*; they are exclusive against the answer and none of distractors can replace the answer to avoid allowing multiple correct answers in one quiz. Second, distractors should be *valid*; they discriminate learners' proficiency adequately.

Each side, government and opposition, is _____
the other for the political crisis, and for the violence.

(a) blaming (b) accusing (c) BOTH

Figure 1: Example of a fill-in-the-blank quiz, where (a) *blaming* is the answer and (b) *accusing* is a distractor.

There are previous studies on distractor generation for automatic fill-in-the-blank quiz generation (Mitkov et al., 2006). Hoshino and Nakagawa (2005) randomly selected distractors from words in the same document. Sumita et al. (2005) used an English thesaurus to generate distractors. Liu et al. (2005) collected distractor candidates that are close to the answer in terms of word-frequency, and ranked them by an association/collocation measure between the candidate and surrounding words in a given context. Dahlmeier and Ng (2011) generated candidates for collocation error correction for English as a Second Language (ESL) writing using paraphrasing with native language (L1) pivoting technique. This method takes an sentence containing a collocation error as input and translates it into L1, and then translate it back to English to generate correction candidates. Although the purpose is different, the technique is also applicable for distractor generation. To our best knowledge, there have not been studies that fully employed actual errors made by ESL learners for distractor generation.

In this paper, we propose automated distractor generation methods using a large-scale ESL corpus with a discriminative model. We focus on *semantically confusing* distractors that measure learners' competence to distinguish word-sense and select an appropriate word. We especially target verbs, because verbs are difficult for language learners to use correctly (Leacock et al., 2010). Our proposed methods use discriminative models

[*] This work has been done when the author was visiting Microsoft Research Asia.

[†] Now at Tokyo Metropolitan University (Email: komachi@tmu.ac.jp).

[1] http://www.ets.org/toeic

[2] http://www.ets.org/toefl

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 238–242,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

Orig.	I	stop		company	on	today	.
Corr.	I	quit	a	company		today	.
Type	NA	#REP#	#DEL#	NA	#INS#	NA	NA

Figure 2: Example of a sentence correction pair and error tags (Replacement, Deletion and Insertion).

trained on error patterns extracted from an ESL corpus, and can generate exclusive distractors with taking context of a given sentence into consideration.

We conduct human evaluation using 3 native and 23 non-native speakers of English. The result shows that 98.3% of distractors generated by our methods are reliable. Furthermore, the non-native speakers' performance on quiz generated by our method has about 0.76 of correlation coefficient with their TOEIC scores, which shows that distractors generated by our methods satisfy validity.

Contributions of this paper are twofold; (1) we present methods for generating reliable and valid distractors, (2) we also demonstrate the effectiveness of ESL corpus and discriminative models on distractor generation.

2 Proposed Method

To generate distractors, we first need to decide which word to be blanked. We then generate candidates of distractors and rank them based on a certain criterion to select distractors to output.

In this section, we propose our methods for extracting target words from ESL corpus and selecting distractors by a discriminative model that considers long-distance context of a given sentence.

2.1 Error-Correction Pair Extraction

We use the Lang-8 Corpus of Learner English[3] as a large-scale ESL corpus, which consists of 1.2M sentence correction pairs. For generating semantic distractors, we regard a correction as a target and the misused word as one of the distractor candidates.

In the Lang-8 corpus, there is no clue to align the original and corrected words. In addition, words may be deleted and inserted in the corrected sentence, which makes the alignment difficult. Therefore, we detect word deletion, insertion, and replacement by dynamic programming[4]. We compare a corrected sentence against its original sentence, and when word insertion and deletion errors are identified, we put a spaceholder (Figure 2). We then extract error-correction (i.e. replacement) pairs by comparing trigrams around the replacement in the original and corrected sentences, for considering surrounding context of the target. These error-correction pairs are a mixture of grammatical mistakes, spelling errors, and semantic confusions. Therefore, we identify pairs due to semantic confusion; we exclude grammatical error corrections by eliminating pairs whose error and correction have different part-of-speech (POS)[5], and exclude spelling error corrections based on edit-distance. As a result, we extract 689 unique verbs (lemma) and 3,885 correction pairs in total.

Feature	Example
Word[i-2]	*,*
Word[i-1]	*is*
Word[i+1]	*the*
Word[i+2]	*other*
Dep[i]_child	*nsubj_side, aux_is, dobj_other, prep_for*
Class	*accuse*

Table 1: Example of features and class label extracted from a sentence: *Each side, government and opposition, is *accusing/blaming the other for the political crisis, and for the violence.*

Using the error-correction pairs, we calculate conditional probabilities $P(w_e|w_c)$, which represent how probable that ESL learners misuse the word w_c as w_e. Based on the probabilities, we compute a confusion matrix. The confusion matrix can generate distractors reflecting error patterns of ESL learners. Given a sentence, we identify verbs appearing in the confusion matrix and make them blank, then outputs distractor candidates that have high confusion probability. We rank the candidates by a generative model to consider the surrounding context (e.g. N-gram). We refer to this generative method as Confusion-matrix Method (CFM).

2.2 Discriminative Model for Distractor Generation and Selection

To generate distractors that considers long-distance context and reflects detailed syntactic information of the sentence, we train multiple classifiers for each target word using error-correction pairs extracted from ESL corpus. A classifier for

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

[4] The implementation is available at `https://github.com/tkyf/epair`

[5] Because the Lang-8 corpus does not have POS tags, we assign POS by the NLTK (`http://nltk.org/`) toolkit.

a target word takes a sentence (in which the target word appears) as an input and outputs a verb as the best distractor given the context using following features: 5-gram (±1 and ±2 words of the target) lemmas and dependency type with the target child (lemma). The dependent is normalized when it is a pronoun, date, time, or number (e.g. *he* → *#PRP#*) to avoid making feature space sparse. Table 1 shows an example of features and a class label for the classifier of a target verb (*blame*).

These classifiers are based on a discriminative model: Support Vector Machine (SVM)[6] (Vapnik, 1995). We propose two methods for training the classifiers.

First, we directly use the corrected sentences in the Lang-8 corpus. As shown in Table 1, we use the 5-gram and dependency features[7], and use the original word (misused word by ESL learners) as a class. We refer to this method as DiscESL.

Second, we train classifiers with an ESL-simulated native corpus, because (1) the number of sentences containing a certain error-correction pair is still limited in the ESL corpus and (2) corrected sentences are still difficult to parse correctly due to inherent noise in the Lang-8 corpus. Specifically, we use articles collected from *Voice of America (VOA) Learning English*[8], which consist of 270k sentences. For each target in a given sentence, we artificially change the target into an incorrect word according to the error probabilities obtained from the learners confusion matrix explained in Section 2.2. In order to collect a sufficient amount of training data, we generate 100 samples for each training sentence in which the target word is replaced into an erroneous word. We refer to this method as DiscSimESL[9].

3 Evaluation with Native-Speakers

In this experiment, we evaluate the reliability of generated distractors. The authors asked the help of 3 native speakers of English (1 male and 2 females, majoring computer science) from an author's graduate school. We provide each participant a gift card of $30 as a compensation when completing the task.

Method	**Corpus**	**Model**
Proposed		
CFM	ESL	Generative
DiscESL	ESL	Discriminative
DiscSimESL	Pseudo-ESL	Discriminative
Baseline		
THM	Native	Generative
RTM	Native	Generative

Table 2: Summary of proposed methods (CFM: Confusion Matrix Method, DiscESL: Discriminative model with ESL corpus, DiscSimESL: Discriminative model with simulated ESL corpus) and baseline (THM: Thesaurus Method, RTM: Roundtrip Method).

In order to compare distractors generated by different methods, we ask participants to solve the generated fill-in-the-blank quiz presented in Figure 1. Each quiz has 3 options: (a) only word A is correct, (b) only word B is correct, (c) both are correct. The source sentences to generate a quiz are collected from VOA, which are not included in the training dataset of the DiscSimESL. We generate 50 quizzes using different sentences per each method to avoid showing the same sentence multiple times to participants. We randomly ordered the quizzes generated by different methods for fair comparison.

We compare the proposed methods to two baselines implementing previous studies: Thesaurus-based Method (THM) and Roundtrip Translation Method (RTM). Table 2 shows a summary of each method. The THM is based on (Sumita et al., 2005) and extract distractor candidates from synonyms of the target extracted from WordNet[10]. The RTM is based on (Dahlmeier and Ng, 2011) and extracts distractor candidates from *roundtrip* (pivoting) translation lexicon constructed from the WIT^3 corpus (Cettolo et al., 2012)[11], which covers a wide variety of topics. We build English-Japanese and Japanese-English word-based translation tables using GIZA++ (IBM Model4). In this dictionary, the target word is translated into Japanese words and they are translated back to English as distractor candidates. To consider (local) context, the candidates generated by the THM, RTM, and CFM are re-ranked by 5-gram language

[6] We use Linear SVM with default settings in the scikit-learn toolkit 0.13.1. http://scikit-learn.org

[7] We use the Stanford CoreNLP 1.3.4 http://nlp.stanford.edu/software/corenlp.shtml

[8] http://learningenglish.voanews.com/

[9] The implementation is available at https://github.com/keisks/disc-sim-esl

[10] WordNet 3.0 http://wordnet.princeton.edu/wordnet/

[11] Available at http://wit3.fbk.eu

Method	**RAD (%)**	κ
Proposed		
CFM	94.5 (93.1 - 96.0)	0.55
DiscESL	95.0 (93.6 - 96.3)	0.73
DiscSimESL	**98.3 (97.5 - 99.1)**	0.69
Baseline		
THM	89.3 (87.4 - 91.3)	0.57
RTM	93.6 (92.1 - 95.1)	0.53

Table 3: Ratio of appropriate distractors (RAD) with a 95% confidence interval and inter-rater agreement statistics κ.

model score trained on Google 1T Web Corpus (Brants and Franz, 2006) with IRSTLM toolkit[12].

As an evaluation metric, we compute the ratio of appropriate distractors (*RAD*) by the following equation: $RAD = N_{AD}/N_{ALL}$, where N_{ALL} is the total number of quizzes and N_{AD} is the number of quizzes on which more than or equal to 2 participants agree by selecting the correct answer. When at least 2 participants select the option (c) (both options are correct), we determine the distractor as inappropriate. We also compute the average of inter-rater agreement κ among all participants for each method.

Table 3 shows the results of the first experiment; RAD with a 95% confidence interval and inter-rater agreement κ. All of our proposed methods outperform baselines regarding RAD with high inter-rater agreement. In particular, DiscSimESL achieves 9.0% and 4.7% higher RAD than THM and RTM, respectively. These results show that the effectiveness of using ESL corpus to generate reliable distractors. With respect to κ, our discriminative models achieve from 0.12 to 0.2 higher agreement than baselines, indicating that the discriminative models can generate sound distractors more effectively than generative models. The lower κ on generative models may be because the distractors are semantically too close to the target (correct answer) as following examples:

> The coalition has **published/<u>issued</u>* a report saying that ... .

As a result, the quiz from generative models is not reliable since both *published* and *issued* are correct.

Method	r	**Corr**	**Dist**	**Both**	**Std**
Proposed					
CFM	0.71	56.7	29.6	13.5	11.5
DiscESL	0.48	62.4	27.9	10.4	12.8
DiscSimESL	**0.76**	64.0	20.7	15.1	13.4
Baseline					
THM	0.68	57.2	28.1	14.6	10.7
RTM	0.67	63.4	26.9	9.5	13.2

Table 4: (1) Correlation coefficient r against participants' TOEIC scores, (2) the average percentage of correct answer (Corr), incorrect answer of distractor (Dist), and incorrect answer that both are correct (Both) chosen by participants, and (3) standard deviation (Std) of Corr.

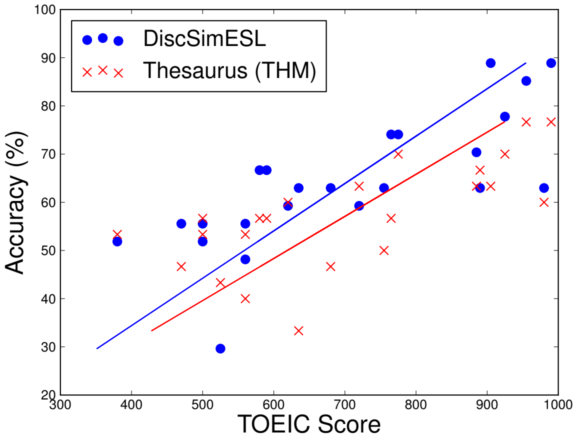

Figure 3: Correlation between the participants' TOEIC scores and accuracy on THM and DiscSimESL.

4 Evaluation with ESL Learners

In this experiment, we evaluate the validity of generated distractors regarding ESL learners' proficiency. Twenty-three Japanese native speakers (15 males and 8 females) are participated. All the participants, who have taken at least 8 years of English education, self-report proficiency levels as the TOEIC scores from 380 to 990[13]. All the participants are graduate students majoring in science related courses. We call for participants by e-mailing to a graduate school. We provide each participant a gift card of $10 as a compensation when completing the task. We ask participants to solve 20 quizzes per each method in the same manner as Section 3. To evaluate validity of distractors, we use only reliable quizzes accepted in Section 3. Namely, we exclude quizzes whose options are both correct. We evaluate correlation between learners' accuracy for the generated quizzes and the TOEIC score.

Table 4 represents the results; the highest corre-

[12] The irstlm toolkit 5.80 `http://sourceforge.net/projects/irstlm/files/irstlm/`

[13] The official score range of the TOEIC is from 10 to 990.

lation coefficient r and standard deviation on DiscSimESL shows that its distractors achieve best validity. Figure 3 depicts the correlations between the participants' TOEIC scores and accuracy (i.e. Corr.) on THM and DiscSimESL. It illustrates that DiscSimESL achieves higher level of positive correlation than THM. Table 4 also shows high percentage of choosing "(c) both are correct" on DiscSimESL, which indicates that distractors generated from DiscSimESL are difficult to distinguish for ESL learners but not for native speakers as a following example:

> ..., she found herself on stage ... **playing/performing* a number one hit.

A relatively lower correlation coefficient on DiscESL may be caused by inherent noise on parsing the Lang-8 corpus and domain difference from quiz sentences (VOA).

5 Conclusion

We have presented methods that automatically generate semantic distractors of a fill-in-the-blank quiz for ESL learners. The proposed methods employ discriminative models trained using error patterns extracted from ESL corpus and can generate reliable distractors by taking context of a given sentence into consideration. The human evaluation shows that 98.3% of distractors are reliable when generated by our method (DiscSimESL). The results also demonstrate 0.76 of correlation coefficient to their TOEIC scores, indicating that the distractors have better validity than previous methods. As future work, we plan to extend our methods for other POS, such as adjective and noun. Moreover, we will take ESL learners' proficiency into account for generating distractors of appropriate levels for different learners.

Acknowledgments

This work was supported by the Microsoft Research Collaborative Research (CORE) Projects. We are grateful to Yangyang Xi for granting permission to use text from Lang-8 and Takuya Fujino for his error pair extraction algorithm. We would also thank anonymous reviewers for valuable comments and suggestions.

References

Charles Alderson, Caroline Clapham, and Dianne Wall. 1995. *Language Test Construction and Evaluation*. Cambridge University Press.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Corpus version 1.1. *Technical report, Google Research.*

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3 : Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Associattion for Machine Translation (EAMT)*, pages 261–268, Trent, Italy, May.

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., July.

Ayako Hoshino and Hiroshi Nakagawa. 2005. A Real-Time Multiple-Choice Question Generation for Language Testing — A Preliminary Study —. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, pages 17–20, Ann Arbor, June.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel R. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan & 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 2nd Workshop on Building Educational Applications Using NLP*, pages 1–8, Ann Arbor, June.

Ruslan Mitkov, Le An Ha, and Nikiforos Karamanis. 2006. A Computer-Aided Environment for Generating Multiple-Choice Test Items. *Natural Language Engineering*, 12:177–194, 5.

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 2nd Workshop on Building Educational Applications Using NLP*, pages 61–68, Ann Arbor, June.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.