From Receptive to Productive: Learning to Use Confusing Words through Automatically Selected Example Sentences

Chieh-Yang Huang[1], Yi-Ting Huang[2], Mei-Hua Chen[3] and Lun-Wei Ku[2]

[1] IST, Pennsylvania State University, USA, [2] IIS, Academia Sinica, Taiwan
[3] FLLD, Tunghai University, Taiwan

[1] chiehyang@psu.edu, [2] {ythuang, lwku}@iis.sinica.edu.tw
[3] mhchen@thu.edu.tw

Abstract

Knowing how to use words appropriately has been a key to improving language proficiency. Previous studies typically discuss how students learn receptively to select the correct candidate from a set of confusing words in the fill-in-the-blank task where specific context is given. In this paper, we go one step further, assisting students to learn to use confusing words appropriately in a productive task: sentence translation. We leverage the GiveMeExample system, which suggests example sentences for each confusing word, to achieve this goal. In this study, students learn to differentiate the confusing words by reading the example sentences, and then choose the appropriate word(s) to complete the sentence translation task. Results show students made substantial progress in terms of sentence structure. In addition, highly proficient students better managed to learn confusing words. In view of the influence of the first language on learners, we further propose an effective approach to improve the quality of the suggested sentences.

1 Introduction

In second or foreign language learning, learning synonyms is not uncommon in vocabulary learning (Hashemi and Gowdasiaei, 2005; Webb, 2007). However, clear differentiation and proper use of near-synonyms poses a challenge to many language learners (Laufer, 1990; Tinkham, 1993; Waring, 1997). Researchers have investigated language learners' lexical use problems, e.g., (Chen and Lin, 2011; Hemchua et al., 2006; Yanjuan, 2014; Laufer, 1990; Tinkham, 1993; Waring, 1997; Yeh et al., 2007; Zughoul, 1991) and suggested that discriminating among semantically similar items presents difficulties for learners (Laufer, 1990). For example, Zughoul (1991) analyzed the writings of Arab EFL college students and found that misapplication of near-synonyms was the most common type of word choice error made by his students. Likewise, Hemchua and Schmitt (2006) investigated lexical error types in the writings of Thai college students and found that the use of near-synonyms was the most common error made by their students.

Learners are prone to assuming that synonyms behave identically in all contexts (Martin, 1984). Actually, even though two words may share similar meanings, they may not be fully substitutable in certain scenarios (Edmonds and Hirst, 2002; Karlsson, 2014; Liu and Zhong, 2014; Martin, 1984; Webb, 2007). Synonyms are highly likely to confuse learners (Martin, 1984). For example, both *emphasis* and *stress* describe "special attention or importance". The verbs *lay*, *place*, and *put* can collocate with "emphasis on" and "stress on"; however, "place stress on" is a rare expression (it occurs only once in the British National Corpus). For ESL/EFL learners, correct word usage necessitates not only knowledge of the meaning of a word but also knowledge of its paradigmatic and syntagmatic association. Without usage information, synonyms "usually leave the student mystified" (Martin, 1984). Verbs *construct* and *establish* illustrate the fact that synonyms do not always have the same collocates (Webb, 2007). Although both words share the same meaning of "build", in practice, they are not interchangeable in the collocations "establish contact" and "construct system". Learners must grasp the collocational and syntactic differences to use synonyms effectively in a productive mode (Martin, 1984).

For language learners, to facilitate the use of near-synonyms, confusing words, or collocations, it is not enough to just learn the senses of a single confusing word. This has led to the design of learning materials such as thesauri and dictionaries for confusing and easily-misused words (Room, 1988; Ragno, 2016). Although the

information these reference tools provide is appropriate and instructive, the contents – especially example sentences – are neither rich nor constantly updated.

In view of this, artificial intelligence techniques recently have been widely applied to assist language learning. Applications such as grammar correction (Ng et al., 2014; Napoles and Callison-Burch, 2017) and essay scoring (Alikaniotis et al., 2016; Dong and Zhang, 2016; Zhang and Litman, 2018) are relatively mature. Research on the lexical substitution (McCarthy and Navigli, 2007, 2009; Mihalcea et al., 2010; Melamud et al., 2015) and the detection and correction of collocation errors (Futagi, 2010; Alonso Ramos et al., 2014) have also shown the potential of helping ESL learn similar words, near-synonyms or synonyms. Lexical substitution task try to determine a substitute for a word in a context and preserving its meaning and is possible to help language learners understand the correct meaning of a target word by selecting a lexical substitute. The detection and correction, on the other hand, is an inevitable assistance for ESL learners since, as we know, collocation error is one of the most common lexical misuse problem. However, as interpretation is still challenging for AI models, especially deep learning models (Ribeiro et al., 2016; Doshi-Velez and Kim, 2017), there are fewer applications for tasks involving comparisons and explanations, which is the key to learning confusing words.

GiveMeExample (Huang et al., 2017) is one of the few systems. It offers students suggestions of example sentences for confusing words and helps them to choose proper words for fill-in-the-blank multiple-choice questions. GiveMeExample aims to provide opportunities for learners to self-learn the nuances between confusing words by comparing and contrasting the suggested example sentences. However, the fill-in-the-blank multiple-choice format has its limitations. First, it decreases learning efficiency: students look for hints (such as prepositions or collocations) from the example sentences to match the words adjacent to the blank instead of reading and comparing these example sentences thoroughly. Also, as answering multiple-choice questions is a discriminative task, students attempt to select the most possible candidate among all choices instead of learning to properly use the confusing words in question.

To improve the learning effect, we adopt Give-

MeExample but deploy it using a carefully designed sentence translation task. Studies (Uzawa, 1996; Prince, 1996; Laufer and Girsai, 2008) have investigated the effect of using translation tasks in language learning. With the integration of the translation task, learners were asked to produce a second language (L2) text conditioned on a given first language (L1) sentence. It is one of effective ways to learn word usage by producing a good translation. In other words, we intentionally move from a receptive to a productive learning task. Generating sentences using confusing words requires a better understanding of the words: with this task we hope to discover how to better assist language learners to learn to differentiate confusing words.

2 Automatic Example Sentence Selection

In this study, we seek to use the GiveMeExample system (Huang et al., 2017) as a basis to improve the automatic example sentence selection task which aims to select sentences that clarify the differences between confusing words. GiveMeExample proposes a clarification score to represent the ability of a sentence to clear up confusion between the given words. In this section, we describe the three main steps to build the automatic example sentence selection model: the definition of the clarification score, the word usage model, and the dictionary-like sentence classifier.

2.1 Problem Definition

Here we define the task more clearly. Given a confusing word set $W = \{w_1, w_2, ..., w_n\}$ and their corresponding sentence sets $\{S_1, S_2, ..., S_n\}$, each sentence set contains a set of sentences $S_t = \{s_{t1}, s_{t2}, ..., s_{tm}\}$. The target is to choose k sentences from each sentence set that clarify the differences among the words in the confusing word set. The desired results are thus sentence sets which clarify W, $\{S'_1, S'_2, ..., S'_n\}$, where $S'_t = \{s'_1, s'_2, ..., s'_k\}$.

2.2 Workflow

Given a word set and the corresponding sentence sets, GiveMeExample selects sentences by (1) building a word usage model for each word, (2) selecting learning-suitable sentences using a dictionary-like sentence classifier, and (3) ranking sentences by computing clarification scores with the help of the word usage model. The top

Number	Word	Example sentence
1	refuse	I was expecting you to <u>refuse</u> to leave the house.
2	refuse	She declined to serve as an informant and <u>refused</u> his request that she keep their meeting secret.
3	reject	In July, a judge in Australia <u>rejected</u> his request for a suppression order.

Table 1: Example sentences that illustrate clarification

five sentences for each word are selected to show learners.

2.3 Clarification Score

To understand the definition of clarification, we start from the confusing word set {*refuse*, *reject*} in Table 1. The first sentence clarifies the differences better than the second sentence, as the usage of *refuse* in "refused his request" from the second sentence is the same as that for *reject* in "rejected his request" in the third sentence. This illustrates two properties of clarification: the fitness score and the relative closeness score. The fitness score measures how well a sentence s illustrates the usage of word w_1: in this sentence the word should be used in a common way instead of a rare way. The relative closeness score, in turn, measures how well a sentence s for word w_1 highlights the difference between w_1 and the other words $\{w_2, ..., w_n\}$: it must be appropriate for w_1 but inappropriate for $\{w_2, ..., w_n\}$. Namely, when we replace w_1 with $\{w_2, ..., w_3\}$ in s, this sentence should become a wrong sentence. As a result, given a function $P(s|w)$ that estimates the fitness between a sentence s and a word w, we define the clarification score as

$$score(s|w_i) = P(s|w_i) * (\sum_{w_j \in W - w_i} P(s|w_i) - P(s|w_j))$$

(1)

which is the multiplication of the fitness score and the relative closeness score.

2.4 Word Usage Model

The word usage model represents the distribution of the usage and the context for a given word, that is, the fitness score $P(s|w)$. GiveMeExample includes two word usage models: a Gaussian mixture model (GMM) and a bidirectional long-short-term-memory model (BiLSTM), described as follows. Notice that the word usage model is trained as a classifier per word.

2.4.1 GMM with Local Contextual Features

The idea of the GMM is to turn words around the target word, namely, its context, into embeddings and then model the distribution with a

Gaussian mixture model (Xu and Jordan, 1996). Empirically, taking words within a window of size two provides the best results. Therefore, given a sentence $s = \{w_1 \cdots w_t \cdots w_n\}$ where w_t is the target word, the features are $f = \{e_{w_{t-2}}, e_{w_{t-1}}, e_{w_{t-2}} + e_{w_{t-1}}, e_{w_{t+1}}, e_{w_{t+2}}, e_{w_{t+1}} + e_{w_{t+2}}\}$. Note that the features contain not only the corresponding word embeddings, but also the summation of two adjacent words to leverage the meaning. Since the word embedding contains both word identity information and semantic information, the GMM model[1] therefore learns the distribution of both usage and semantic meaning.

2.4.2 BiLSTM

As the confusing words can diverge widely from the target word itself, or could involve long-term dependencies, GMM with local contextual features do not always capture enough information. The BiLSTM model thus utilizes the whole sentence as a feature. The BiLSTM model consists of a forward LSTM and a backward LSTM, which take the words preceding and following the target word as features respectively. The output vectors of these two LSTMs are concatenated to form a sentence embedding. After passing through two dense layers, the BiLSTM model is then built as a binary classifier that decides whether the given sentence is the sentence of the target word or not. In contrast to the generative GMM model, negative samples are needed to train the BiLSTM. As a result, sentences from the corpus are randomly sampled as negative samples[2].

2.5 Dictionary-like Sentence Classification

The given sentences are not always suitable for language learning. For example, a 40-word-long sentence could be too complicated and distracting to learn, and a short sentence such as "It is sophisticated" is not suitable for language learning due to its lack of information. GiveMeExample is equipped with a dictionary-like sentence classifier to select sentences that are simple but informa-

---

[1] Each GMM model is trained on 5,000 instances.

[2] Each BiLSTM model is trained on 5,000 positive instances and 50,000 negative instances.

Figure 1: Example questions for translation experiment. Participants click the readmore button to retrieve more example sentences (the maximum number of sentences for each word is five). Also, *introverts* and *extraverts* are two tips that we provide, as they are more difficult but not directly related to *social* and *sociable*.

tive. GiveMeExample collects sentences from the COBUILD English Usage Dictionary (Sinclair, 1992) to train the dictionary-like sentence classifier with syntactic features (Pilán et al., 2014) and a logistic regression model (Walker and Duncan, 1967). Hence, it tends to select sentences similar to those in the COBUILD dictionary.

3 Deployment: Sentence Translation

The sentence translation experiment was separated into a pre-test and a post-test. In both of the tests, participants were asked to translate ten sets of questions from Mandarin to English. In each set, there were four translation questions corresponding to a specific set of confusing words. In addition to answering the question, participants could refer to the example sentences suggested by Give-MeExample in the post-test. In the following paragraph, we describe the experiment in detail.

3.1 Building Translation Questions

In the sentence translation task were 15 confusing word sets selected from Collins COBUILD English Usage (Sinclair, 1992) and the Longman Dictionary of Common Errors (Turton and Heaton, 1996). These two books identify errors in word usage commonly made by language learners and then clear up the confusion. Thus the word sets provided in the books were used as the desired confusing words. A word set contained two or three words. After selecting the confusing word set, we extracted sentences that contain these words from the parallel corpora Chinese English

News Magazine Parallel Text (LDC2005T10) and Hong Kong Parallel Text (LDC2004T08). These sentences were used as candidate questions. Since many sentences in the parallel corpora were long and complicated, we removed sentences whose Chinese translation contains more than 40 words. In the last step, we manually chose appropriate sentences for testing the confusing words. In the end, 15 confusing word sets were determined, each of which contains four questions to be translated resulting a total of 60 questions. Note that some difficult words in the question, such as "introverts" and "extraverts" in Figure 1, were provided as they were unrelated to testing learner use of confusing words.

3.2 Recommending Example Sentences

To recommend sentences, we first collected sentences from Vocabulary[3], an online dictionary. The example sentences in Vocabulary mainly come from formally-written news articles. We collected 5,000 sentences for each word and used all of them to train the GMM and BiLSTM word usage models. When recommending example sentences, we used only the qualified sentences which were filtered by the dictionary-like sentence classifier. The pretrained 300-dimension GloVe (Pennington et al., 2014) embeddings were used in both GMM and BiLSTM. We selected the last five sentences from Vocabulary as a baseline setting.

3.3 Experimental Setup

Sixteen college students were recruited for this translation experiment. As the translation of total 60 questions may not be done in one class, each participants was asked to complete ten randomly-assigned question sets, each of which contained four questions. Thus a total of 40 translation questions were given. This process guarantees that every questions is translated by the same number of participants. The testing period was about 45 minutes, leaving participants about five minutes for each question set. In addition to translating, five example sentences were provided for each word in the post-test. To ensure the students read the suggested sentences, only one example sentence was displayed in the beginning, a "readmore" button was designed for retrieving more example sentences (the maximum number of example sentences is five for each word). The "read-

---

[3]https://www.vocabulary.com/

Category	Example	Grade
Appropriateness	There is a small ~~opportunity~~ **possibility** that she had actually met such a person.	0
Local grammar	What are you going to do if we <u>refuse</u> to ~~following~~ **follow** you?	3
Global grammar	This building ~~is~~ **was** destroyed by the earthquake.	3
Structure	The accident was caused ~~by error~~. (The error is made by human, so it should be "by human error.")	1
Meaning	To a <u>skillful</u> pilot, ~~it's lucky to say that landing in torrential rain.~~ (The meaning is wired and the correct sentence should be "Landing safely in torrential rain can only be a matter of luck for the most skilled pilot.")	1

Table 2: Examples of grade criteria. The underlined word is the target confusing word.

more" activities were logged for further investigation. The pre- and post-tests were administered in two different weeks to reduce short-term memory effects. Figure 1 shows a screen-shot of a post-test with the confusing word set *social* and *sociable*.

The example sentences provided were suggested by the GMM and BiLSTM models or selected from the Vocabulary website. Note that to discourage participants from guessing specific patterns, the example sentences from one of the three sources were presented randomly. For instance, as GMM takes contexts within a window as features, the most significant difference exists only within this window. However, we do not expect participants to look only at this small piece of text. Also, sentences from Vocabulary are generally more difficult than those from GMM or BiLSTM, but participants who are consistently presented with difficult sentences may stop considering these example sentences to be useful resources. As the source is assigned randomly for each proposed example sentence, the total number of sentences for each source is set to the number that can best distribute sentences from different sources evenly.

3.4 Grading

Grading was done by an English native speaker who is professional in language learning and teaching. The grading criteria takes into account appropriateness, grammar, and completeness. Appropriateness measures whether the correct word is used or not, so the score here is either zero or one point. Grammar involves local grammar as well as global grammar. All the grammar errors relating directly to the target confusing word belong to local grammar; the remaining grammar errors throughout the sentence belong to global grammar. The initial points for both grammar parts are four points; each grammar error results in a one-point deduction. Completeness, which eval-

uates whether the student's translation represents all of the meanings, takes into account structure and meaning. If a student missed content such as adverbial phrases, points were deducted in terms of structure. Similarly, if a student's translation was different from the original meaning, points were deducted in terms of meaning. Both structure and meaning started with two points. Examples are listed in Table 2. Given our focus on examining whether students can learn how to differentiate and use confusing words, we computed a weighted sum for reference as follows:

$$WeightedSum = 5*Appropriateness + LocalGrammar \tag{2}$$

which is the sum of the appropriateness scores, weighted by 5, and the local grammar scores.

4 Results and Discussions

The pre and post scores for the grading categories are summarized in Table 3. Student are separated into Highly proficient group and Less proficient group evenly by an external collocation test score (Chen and Lin, 2011). In general, the suggested example sentences helped students make substantial progress in terms of sentence structure. It is worth noting that students were able to comprehend the meaning of confusing words in the given sentences selected from both of the BiLSTM and GMM models. Students performed significantly better in *appropriateness*, *local grammar*, and *structure* when the sentences were suggested by BiLSTM; while the GMM model was good at presenting the structures of sentences and demonstrating the meaning of confusing words.

Highly proficient students learned confusing words better from the suggested example sentences. The findings showed that BiLSTM helped them gain a better understanding of *appropriateness*, *local grammar*, and *structure*, and GMM

Group	Model	Appropriateness			Local grammar			Weighted sum			Global grammar			Structure			Meaning		
		pre	post	t-test	pre	post	t-test	pre	post	t-test	pre	post	t-test	pre	post	t-test	pre	post	t-test
H	Vocabulary	**0.714**	0.571	0.302	**3.429**	3.143	0.178	**7.000**	6.000	0.237	**2.429**	2.143	0.229	0.714	**1.000**	0.229	1.000	1.000	0.500
H	GMM	0.444	0.444	0.500	2.444	**3.000**	0.123	4.667	**5.222**	0.347	**1.667**	1.556	0.364	0.667	**1.222**	0.025*	0.333	**1.111**	0.004*
H	BiLSTM	0.273	**0.545**	0.041*	2.364	**3.273**	0.008*	3.727	**6.000**	0.011*	**1.545**	1.364	0.276	0.818	**1.182**	0.052	0.545	**0.909**	0.052
L	Vocabulary	0.182	**0.364**	0.170	2.182	**2.909**	0.098	3.091	**4.727**	0.056	0.818	**1.091**	0.247	0.364	**1.000**	0.013*	0.455	**0.636**	0.220
L	GMM	0.417	**0.583**	0.169	2.333	**2.917**	0.066	4.417	**5.833**	0.072	0.750	**1.500**	0.028*	0.500	**1.167**	0.012*	0.333	**1.083**	0.010*
L	BiLSTM	0.429	**0.524**	0.165	2.667	**2.714**	0.443	4.810	**5.333**	0.169	1.238	**1.571**	0.116	0.762	**1.143**	0.004*	0.524	**0.857**	0.025*

Table 3: Result of translation experiment. The number of translated questions for each model ranges from 7 to 21, with the average number 11.8, depending on the number of early leave and absence we encountered in the experiment day. The pre- and post- numbers correspond to the average score for pre-test and post-test respectively and the t-test stars represent significance. The participants were separated into highly proficient (H) and less proficient (L) groups.

helped with *structure* and *meaning*. Although it was difficult for less-proficient students to recognize the difference (small improvement in *appropriateness* and *local grammar*), the GMM model significantly facilitated their comprehension in terms of *structure*, *global grammar* and *meaning*.

The "readmore" logs show that most of the students clicked the button and expand all the example sentences immediately. This might imply that students did read all the example sentences and could refer to them when producing translations.

We analyzed the translation tasks to identify possible problems. Below we discuss three possible explanations in terms of test items, learner behavior, and the suggested example sentences.

First, in the proposed translation task, participants sometimes focused on the wrong segment of the test item to translate with the confusing words. This may be because in this productive testing process, we do not specifically tell participants which source word should be aligned to the target confusing word. For instance, in "For a person to become so poor, if it's not because they didn't work hard in their youth then its because they have truly had hard luck", participants should have translated the source words "hard luck" to English using the appropriate word in the confusing word set. However, the students showed confusion in their focusing on translating the source word *poor* into one of *hard*, *difficult*, and *tough* as opposed to the source word *hard* in *hard luck*. One example translation made by a participant is "The reason why a person's life is tough might because he/she was lazy when he/she was young or he/she had a bad luck". In such cases, the learning effect cannot be correctly evaluated.

We seek to find the best example sentences for word sets where the words are confusing for learners. Hence regarding the suggested example sentences, the example sentences were extracted as long as the confusing words shared the least familiar senses. However, this led to words being chosen in example sentences with different senses and/or even different parts of speech, which is how we wanted to compare them. The words *hard* and *difficult* exemplify these issues. First, according to WordNet, *hard* in this case indicates "resisting weight or pressure" in the example sentence "Such uncertainty can be hard on families, too", whereas *difficult* means "needing skill or effort" in the sentence "But other stories are more difficult to explain". On the other hand, *hard* is an adverb in "Banks will have to work harder to make profits", while *difficult* is an adjective in "But other stories are more difficult to explain".

Student behavior also affected the performance of this study. Some highly proficient students were observed skipping the example sentences and thus not learning from them how to differentiate the confusing words, which led to inappropriate translations similar to those made in the pre-test. It could be that these highly proficient students were more confident of their command of certain confusing words. For example, when required to choose from *beat*, *defeat*, and *win* to translate "Emmanuel Macron beats Marine Le Pen in both rounds of the French presidential election", one highly proficient student made these translations in the pre- and post-tests, respectively: "Emmanuel Macron won over Marine Le Pen for two rounds of presidential election", and "Emmanuel Macron won over Marine Le Pen for presidential election for two rounds", whereas *win over* is not a usage suggested by example sentences. In addition, from this example we can see that though they rarely read example sentences, they did try to translate in other words in the post-test, which results in the unstable scores of *global grammar* that are less relevant to the near-synonym recognition but to the translation instead.

466

These three limitations partially explain learner performance in the translation task. Thus we attempted to refine the method for example sentence extraction. Improving the test items and controlling student learning behavior is beyond the scope of this study.

5 Leveraging First Language for Better Example Sentence Selection

From the results of the translation experiment, we observed that some words were confusing to students due to language transfer from L1 (native language) to L2 (foreign language). Some students learn English such that they only remember how to spell words and their L1 definitional glosses, rather than understanding their context or usage. For example, the confusing words *hard* and *difficult* are very similar and almost interchangeable. If these words are memorized only by memorizing the L1 definitional glosses, *not easy*, students may fail to recognize the slight difference between them. In other words, example sentences containing words that translate into similar glosses in L1 are the sentences that indeed contain confusing senses, and thus are the target candidates for the GiveMeExample system to consider for suggestion. We follow this line of thinking to improve the example sentences.

In the new setting, the GiveMeExample system groups example sentences by the L1 definitional glosses of confusing words before proceeding to automatic sentence selection with the BiLSTM or GMM word usage model. When a word has multiple senses, this step helps to identify the confusing sense, under the assumption that words with similar L1 definitions are confusing. Take for example *hard* and *difficult*: *hard* as an adjective has multiple meanings – "not easy, requiring great physical or mental effort to accomplish, resisting weight or pressure, hard to bear", etc; whereas *difficult* has the meanings "not easy, requiring great physical or mental effort to accomplish, and hard to control". The common sense in L1 is *not easy, requiring great physical or mental effort to accomplish.* Sentences containing confusing words whose L1 translations share these two senses are selected for later processing and suggestion.

To identify these sentences, we need each word in the sentence and its corresponding L1 translation. For this purpose, parallel texts from two corpora – Chinese English News Magazine Parallel Text (LDC2005T10) and Hong Kong Parallel Text (LDC2004T08), that is, a total of 2,682,129 English-Chinese sentence pairs – are utilized to learn the word alignment between L1 and L2 parallel sentences. To align example sentences from Vocabulary, first they were all translated into Traditional Chinese using Googletrans[4]. Then we used NLTK[5] to tokenize English sentences and CKIP (Chen and Liu, 1992) to segment Chinese sentences respectively. After that, the word alignment model GIZA++ (Och and Ney, 2003), a toolkit that implements several statistical word alignment models, was adopted to align English words to their corresponding Chinese words. After alignment, the L1 translations of confusing words were recognized, after which the sentences in the example sentence pool of the confusing words in the same set were clustered with respect to their L1 translation. There were 12 confusing word sets with more than one common L1 translation. Only words in three confusing word sets (*possibility* vs. *opportunity*, *social* vs. *sociable*, and *unusual* vs. *strange*) had all different L1 translations. When a common L1 translation was found for a set of confusing words, GiveMeExample passed through only those sentences containing confusing words with the same L1 translation to the sentence selection component.

5.1 Human Evaluation

We employed Amazon Mechanical Turk crowd-workers to give their perspectives on the suggested sentences considering the L1 of learners. Twelve sets of confusing words with common L1 translations were evaluated. GiveMeExample in both the original and the new settings suggested respectively five sentences using the BiLSTM and GMM models for each word in the twelve sets. In this new setting, six words – (*briefly*, *duty*, *ordinary*, *sight*, *shortly*, and *unusual*) – had less than five sentences.

Figure 2 shows a screenshot of two versions of the suggested example sentences presented side-by-side. Crowd-workers were given no information about the settings or the sentence selection models (BiLSTM or GMM). For each task, participants were to read several sentences suggested by the two versions of the GiveMeExample system and then answer the following four questions.

---

[4]https://pypi.org/project/googletrans/
[5]https://www.nltk.org/

Figure 2: An example survey for crowd-workers to compare GiveMeExample with different settings. In this specific example, *first* represents sentences suggested by the new setting; *second* represents those from the original.

Q1: Is Mandarin your first language (y/n)?

Q2: Are these words confusing to you (y/n)?

Q3: Which set of example sentences you think is more useful for learning these words (1/2)?

Q4: In what aspect you think they are more useful (choose one)? (a) clarifying their meaning (e.g., *social* encounter vs. *sociable character*) (b) demonstrate their usage (e.g., *as usual* but not *as common*) (c) showing correct grammar (e.g., *The proposal was narrowly defeated in a January election*, but *Obviously we want to continue to win games*.)

The purpose of Q1 and Q2 is to understand the background of turkers, Q3 is to compare the new setting with old setting among two models, and Q4 is to investigate the effect of considering L1 translation. We also consulted a native speaker who works as an expert editor. This expert completed the surveys under the same conditions as the crowd-workers.

5.2 Results and Analysis

Sixty-one crowd-workers participated in the evaluation. Mandarin was the first language of 12 (19.67%) of them. On average, each worker completed six tasks (SD=8.17). For each set out of 12 sets, 15 workers were asked to answer the questions. We tested the example sentences suggested by both GMM and BiLSTM models, collecting in total 360 ratings from workers. It was an interesting finding that only 5% of the confusing word sets were labeled by workers as confusing no matter they were native speakers or not [6]. Details are shown in Table 4.

Table 4 shows the feedback on Q3 and Q4 from workers and the expert on each confusing word set. Results from the expert confirm that when considering L1, our approach could provide better example sentences. However, results from the crowd-workers were mixed.

Several interesting observations were gleaned from this experiment. First, when considering the L1 translation and grouping sentences by their L1 sense, the example sentences containing confusing words with different senses were excluded. Therefore, learners could focus more on the confusing sense to be learned. For example, *work hard* is a commonly seen phrase in the example sentences suggested by the original setting. When students learned the confusion set containing *hard*, *difficult*, and *tough*, the sentences containing *work hard* were of little help, as the meanings were irrelevant to the confusing sense in this set. However, in the new setting, the example sentences for *hard* were more semantically related to *difficult* and *tough*. We can say that in this task, consideration of L1 amounted to implicitly performing word sense disambiguation (WSD).

The exclusion of sentences that did not contain words with the confusing sense has additional benefit. That is, the suggested sentences are more likely to focus on the demonstration of the confusing sense. This has the advantage that the confusing words in the suggested sentences are diverse in their part of speech and pragmatic domain. For instance, in the confusion set *defeat*, *win*, and *beat*, the common L1 sense among them is "to conquest" and "victory". Under these certain meanings, only *win* can be used as a verb or a noun whereas the other two words can only function as a verb. This illustrates the power of grouping sentences by L1 translation Another example is *destroy* in the confusion set *destroy*, *ruin*, and *spoil*. In the original setting, *destroy* is used in only the military domain and thus is misleading. When using the GMM model which considers only the local context, the issue is even more serious. This is mitigated in the new setting, especially for the GMM model.

Following the above, in some cases workers indeed tended to prefer example sentences of some

---

[6]The expert had a clear understanding of these words.

Confusing word set	Q2 (turkers)		Q3 (turkers)		Q3 (expert)		Q4 (turkers)					
							BiLSTM			GMM		
	No	Yes	BiLSTM	GMM	BiLSTM	GMM	(a)	(b)	(c)	(a)	(b)	(c)
ordinary / usual / common	97%	3%	N	N	O	N	46%	40%	14%	46%	40%	14%
skillful / skilled	93%	7%	O	O	O	N	46%	46%	6%	34%	46%	20%
alternative / alternate	97%	3%	O	O	N	O	34%	60%	6%	26%	46%	26%
destroy / ruin / spoil	100%	0%	O	N	N	N	40%	40%	20%	20%	74%	6%
scarce / rare / unusual	97%	3%	O	O	O	O	14%	74%	14%	0%	74%	26%
defeat / win / beat	100%	0%	N	N	N	N	40%	46%	14%	20%	60%	20%
sight / landscape / scenery	93%	7%	N	O	N	O	40%	46%	14%	34%	46%	20%
briefly / shortly / concisely	97%	3%	O	N	O	O	14%	66%	20%	14%	60%	26%
hard / difficult / tough	90%	10%	O	N	O	O	14%	80%	6%	20%	54%	26%
error / mistake / oversight	90%	10%	O	N	N	N	26%	60%	14%	20%	66%	14%
duty / job / task	97%	3%	N	N	O	N	46%	46%	6%	14%	66%	20%
obligation / responsibility / commitment	93%	7%	N	N	N	N	26%	66%	6%	46%	40%	14%
Mean	95%	5%	42%(N)	67%(N)	50%(N)	58%(N)	32%	56%	12%	24%	56%	20%

Table 4: Results from the human evaluation. N represents the example sentences from the new setting, and O are from the original one. In addition, the expert annotated that ALL of the suggested sentences were useful for demonstrating their usage (b).

pattern. For example, in the set *scarce*, *rare*, and *unusual*, confusing words in the example sentences that shared the L1 translation *very hardly* resulted in example sentences containing confusing words functioning as adverb, adjective, and adjective, respectively; however, in the original setting where context is considered before sense, they all function as adjectives. This interesting result reveals that there is overhead when learning from materials without patterns, which could also be why only highly proficient students can learn the appropriateness.

6 Conclusion

In this paper, we leverage GiveMeExample, an AI system which automatically suggests example sentences to help ESL learners better learn to differentiate confusing words. To evaluate the system effectiveness, we designed a sophisticated sentence translation task around the problem of students not really learning via the previously designed receptive task, i.e., multiple-choice selection. This approach was evaluated using college students; results show that students made substantial progress with assistance of the system. Specifically, after learning the example sentences, students produced more structural sentences. However, learning to use appropriate words is a demanding task which requires higher language proficiency.

The learner's first language may lead to confusion in different areas: this is also taken into account with a novel approach. Overall, the example sentences in the refined list were considered more useful for learning by Amazon mechanical turkers and the expert English editor. However, for ESL learners such as students and some of the turkers, they tended to prefer example sentences with similar patterns to mitigate cognitive overhead. Thus, future work will focus on providing example sentences with similar patterns but diverse contexts.

Acknowledgments

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Margarita Alonso Ramos, Marcos García Salido, and Orsolya Vincze. 2014. Towards a collocation writing assistant for learners of spanish.

Keh-Jiann Chen and Shing-Huan Liu. 1992. Word identification for Mandarin Chinese sentences. In *Proceedings of the 14th Conference on Computational linguistics (COLING '92) - Volume 1*, pages 101–107.

Mei-Hua Chen and Maosung Lin. 2011. Factors and analysis of common miscollocations of college students in Taiwan. *Studies in English Language and Literature*, (28):57–72.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational linguistics*, 28(2):105–144.

Yoko Futagi. 2010. The effects of learner errors on the development of a collocation detection tool. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 27–34. ACM.

Mohammad Reza Hashemi and Farah Gowdasiaei. 2005. An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC journal*, 36(3):341–361.

Saengchan Hemchua, Norbert Schmitt, et al. 2006. An analysis of lexical errors in the English compositions of Thai learners.

Chieh-Yang Huang, Mei-Hua Chen, and Lun-Wei Ku. 2017. Towards a better learning of near-synonyms: Automatically suggesting example sentences via fill in the blank. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 293–302. International World Wide Web Conferences Steering Committee.

Monica Karlsson. 2014. Advanced Students' L1 and L2 Mastery of Lexical Fields of Near Synonyms. *World Journal of English Language*, 4(3):1.

Batia Laufer. 1990. Ease and difficulty in vocabulary learning: Some teaching implications. *Foreign Language Annals*, 23(2):147–155.

Batia Laufer and Nany Girsai. 2008. Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied linguistics*, 29(4):694–716.

Dilin Liu and Shouman Zhong. 2014. L2 vs. L1 use of synonymy: An empirical study of synonym use/acquisition. *Applied linguistics*, 37(2):239–261.

Marilyn Martin. 1984. Advanced vocabulary teaching: The problem of synonyms. *The Modern Language Journal*, 68(2):130–137.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language resources and evaluation*, 43(2):139–159.

Oren Melamud, Omer Levy, and Ido Dagan. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 9–14. Association for Computational Linguistics.

Courtney Napoles and Chris Callison-Burch. 2017. Systematically adapting machine translation for grammatical error correction. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 345–356.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 174–184.

Peter Prince. 1996. Second language vocabulary learning: The role of context versus translations as a function of proficiency. *The modern language journal*, 80(4):478–493.

Nancy Ragno. 2016. *Use the Right Word: Your Quick & Easy Guide to 158 Words Most Often Confused or Misused*. Nancy Ragno.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Adrian Room. 1988. *Dictionary of confusing words and meanings*. Dorset Press.

John Sinclair. 1992. *Collins COBUILD English Usage*. Collins.

Thomas Tinkham. 1993. The effect of semantic clustering on the learning of second language vocabulary. *System*, 21(3):371–380.

Nigel D Turton and John Brian Heaton. 1996. *Longman Dictionary of Common Errors*. Longman.

Kozue Uzawa. 1996. Second language learners' processes of l1 writing, l2 writing, and translation from l1 into l2. *Journal of second language writing*, 5(3):271–294.

Strother H Walker and David B Duncan. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179.

Robert Waring. 1997. The negative effects of learning words in semantic sets: A replication. *System*, 25(2):261–274.

Stuart Webb. 2007. The effects of synonymy on second-language vocabulary learning. *Reading in a Foreign Language*, 19(2):120–136.

Lei Xu and Michael I Jordan. 1996. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation*, 8(1):129–151.

HUO Yanjuan. 2014. BNC-Based Design of College English Vocabulary Teaching for Chinese College Students. *Studies in Literature and Language*, 8(3):122–125.

Yuli Yeh, Hsien-Chin Liou, and Yi-Hsin Li. 2007. Online synonym materials and concordancing for EFL college writing. *Computer Assisted Language Learning*, 20(2):131–152.

Haoran Zhang and Diane Litman. 2018. Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 399–409.

Muhammad Raji Zughoul. 1991. Lexical choice: Towards writing problematic word lists. *International Review of Applied Linguistics*, 29(1):45–60.