Towards a Better Learning of Near-Synonyms: Automatically Suggesting Example Sentences via Filling in the Blank

Chieh-Yang, Huang[†] Mei-Hua, Chen[‡] Lun-Wei, Ku[†]

[†] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[‡] Department of Foreign Languages and Literature, Tunghai University, Taichung, Taiwan
[†] {appleternity, lwku}@iis.sinica.edu.tw [‡] mhchen@thu.edu.tw

ABSTRACT

Language learners are confused by near-synonyms and often look for answers from the Web. However, there is little to aid them in sorting through the overwhelming load of information that is offered. In this paper, we propose a new research problem: suggesting example sentences for learning word distinctions. We focus on near-synonyms as the first step. Two kinds of one-class classifiers, the GMM and BiLSTM models, are used to solve fill-in-the-blank (FITB) questions and further to select example sentences which best differentiate groups of near-synonyms. Experiments are conducted on both an open benchmark and a private dataset for the FITB task. Experiments show that the proposed approach yields an accuracy of 73.05% and 83.59% respectively, comparable to state-of-the-art multi-class classifiers. Learner study further shows the results of the example sentence suggestion by the learning effectiveness and demonstrates the proposed model indeed is more effective in learning near-synonyms compared to the resource-based models.

Keywords

Natural Language Processing, Computer-assisted Language Learning, BiLSTM, GMM, Data-driven Language Learning

1. INTRODUCTION

As connection to the Web has become part of life, more and more people are looking for answers from the Web. Language learners are no exceptions and thus many resources are available for them on the Web. For example, it is not uncommon for English learners to consult online dictionaries (e.g., Vocabulary.com). Unlike traditional dictionaries, online dictionaries expose learners to a larger number of authentic language examples. Learners are able to discover language rules such as grammatical functions of words and collocations [34, 17]. Such learning process is known as inductive data-driven language learning [16]. While learning a second or foreign language, near-synonyms have been one of the greatest challenges [26]. As a results, learners often count on the reference tools to consult how the synonymous words are appropriately used. However, language learners are not satisfied with the information the dictionaries provide [10, 42]. The example sentences from traditional dictionaries are accurate but limited. On the other hand, some online dictionaries automatically collect online example sentences, which could be too overwhelming and time-consuming [42] for many language learners to induce rules or patterns. Nevertheless, example sentences play a crucial role in appropriate use of near-synonyms. Good examples must demonstrate the difference between near-synonyms; that is, the example sentence must be good for this word but not good for the other to replace the word, whether or not it is grammatically, semantically, or pragmatically appropriate. Bearing this in mind, we develop the `GiveMeExample` system, an example sentence recommendation system.

We design approaches for this system based on two observations from learners who can distinguish a set of synonyms. First, learners should know which word to use in a given context (*Fitness*). Second, learners should know which sentences show clarifying clues and can help them learn these near-synonym words (*Clarification*). Therefore, we adopt a predefined fill-in-the-blank test (*FITB*) and a learner study to evaluate the performance from two aspects: word selection (near-synonym) for the specified context (sentence) and context selection for the specified word. In building an online system, to avoid retraining for different synonym sets, we make the proposed models one(unary)-class learning models, i.e., the models distinguish sentences which belong to a near-synonym word from those which do not; however, we make these models competitive to multi-class models, i.e., models that specify which near-synonym word sentences belong to. In addition, we design a difficulty scorer to select example-oid sentences for learners and minimize their learning load.

To the best of our knowledge, there is no model or service that provides what our system provides for language learners. The main contributions of this work are: (1) We propose for FITB one-class classification models comparable to multi-class ones; (2) We build `GiveMeExample`, the first system to suggest example sentences for learning near-synonyms; (3) We conduct the learner study on the learning effectiveness of the proposed models as well as resource-

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054163

based models, and further show that the former outperform the latter.

2. RELATED WORK

We discuss first from the reason why near-synonyms are difficult for language learners. Then two relevant research streams, example sentence extraction and lexical choice, are described and their corresponding related work are listed. Finally, the vital issue for learners to digest the proposed learning materials, the readability, is mentioned. Its relation to our work is also described.

2.1 Difficulty in Near-synonym Learning

Learning synonyms is "a very common occurrence in vocabulary learning" [40]. Studies show that learning unknown words with known synonyms is "a transfer of knowledge" [40], which not only facilitates learning but also enriches vocabularies (e.g., ([3, 19, 24, 32].) However, appropriate use of synonyms is a challenging task for many language learners [19, 35, 39]. Without sufficient knowledge of the usage of individual synonyms, learners have difficulty determining which synonyms could fit some contexts but not in others [40]. Researchers suggest that effective vocabulary learning occurs in contexts [23] because contextual information equips learners with knowledge such as forms, meanings, grammatical function, collocation, usage constraints [13, 25]. However, the existing reference tools, such as dictionaries or thesauri, appear not to directly and effectively help learners discriminate the nuances of synonymous emotion words [1]. The suggested near-synonyms carry little or no contextual information [22, 20]. Therefore, the synonyms the learner selects are highly likely not to "fit the concept being expressed" [21]. On the other hand, compared with traditional dictionaries, online dictionaries suggest a larger number of example sentences which could be too overwhelming to discover language rules. Such time-consuming inducing process may intimidate language learners [42]. In view of the importance of contextual information (i.e., example sentences) and the limitations of existing reference tools, in this paper, we propose two models (BiLSTM and GMM) to meet learners' needs. Our goal is to automatically suggest representative example sentences from those in online dictionaries to assist learners in differentiating the nuances among near-synonyms.

2.2 Automatic Example Sentence Extraction

Automatic dictionary construction has been widely investigated and applied to various computer-assisted-language-learning (CALL) tasks. However, most dictionary or glossary construction tasks focus on automatic definition extraction and ontology construction. Few researchers have regarded the example sentence as the research target. Kilarriff [18] provides a rule-based approach and clearly defines several criteria for good example sentences. Kilarriff's work, however, only uses features extracted from sentences, such as sentence length, word frequency, and punctuation. Other than this, Didakowski [6] applies natural language processing tools such as part-of-speech taggers and dependency parsers to obtain further syntactic information from sentences, and define additional rules to find high quality sentences based on the retrieved information. A different approach utilizing parallel corpora is introduced in [5], where example sentences are extracted for different senses of a given word. However, these works select appropriate sentences as opposed to clarifying sentences.

2.3 Lexical Choice

The lexical choice task is another research topic highly related to the example sentence recommendation. The intuition is that if making the right lexical choice means knowing the word is right for the sentence, we hence infer that the sentence is also right for the word. Much work has been done on the lexical choice problem. The goal of `GiveMe-Example` is to clearly explain the differences among a set of near-synonyms. Therefore, the lexical choice for non-near-synonyms is beyond the scope of discussion.

The near-synonym lexical choice problem focuses on identifying subtle differences between near-synonyms. To solve this problem, a lexical co-occurrence network including second-order co-occurrence is introduced in [7], where Edmonds suggests a fill-in-the-blank (FITB) test on the 1987 Wall Street Journal (WSJ), which provides a benchmark for evaluating lexical choice performance on near-synonyms. This benchmark is automatically generated by covering up the original word in the sentence, leaving a blank, so although there is potentially more than one option in the choice list, only the original one is considered correct. Approaches utilizing pointwise mutual information (PMI) and support vector machines (SVMs) are proposed in [14]. To exploit the power of large web corpora, Islam and Inkpen[15] build a 5-gram language model with the Google 1T 5-gram database. Wang and Hirst [38] use latent semantic analysis to represent words in latent semantic space and use SVM to learn the subtle differences between near-synonyms. To clearly discriminate near-synonyms, most approaches suggest that the nuances between near-synonyms are hiding in the local context, that is, in the immediately surrounding words. Moreover, most high-performance approaches are multi-class classifiers. In this paper, we introduce a Gaussian mixture model and a Bidirectional Long Short Term Memory model [8] to capture the contextual semantics located within the sentence but make them one-class classifiers.

2.4 Readability

Even if we have selected explanatory example sentences, for language learners, their preference is still for simple sentences. Therefore, we look for assistance from related work on readability measurements. Early work on readability measurement utilized a few simple features such as word frequency. Dale and Chall [4] determined 3,000 common words and used the percentage of rare words to assess the lexical difficulty. Recently, however, researchers have leveraged the power of syntactic parsers to build more robust measurement methods. In [30, 12, 28], machine learning algorithms are proposed to combine language model features, which are grammatical features extracted from parsers, together with traditional features. In this paper, to filter out difficult and inappropriate sentences, we build a difficulty scorer based on the work of [28] but to select informative physical dictionary example-oid sentences, i.e., those sentences similar to expert-generated example sentences for near-synonyms.

3. EXAMPLE SENTENCE SUGGESTION

In this paper, we propose the novel task of example sentence suggestion, aiming at helping learners learn the nuance between near-synonyms. In this section, we first define the

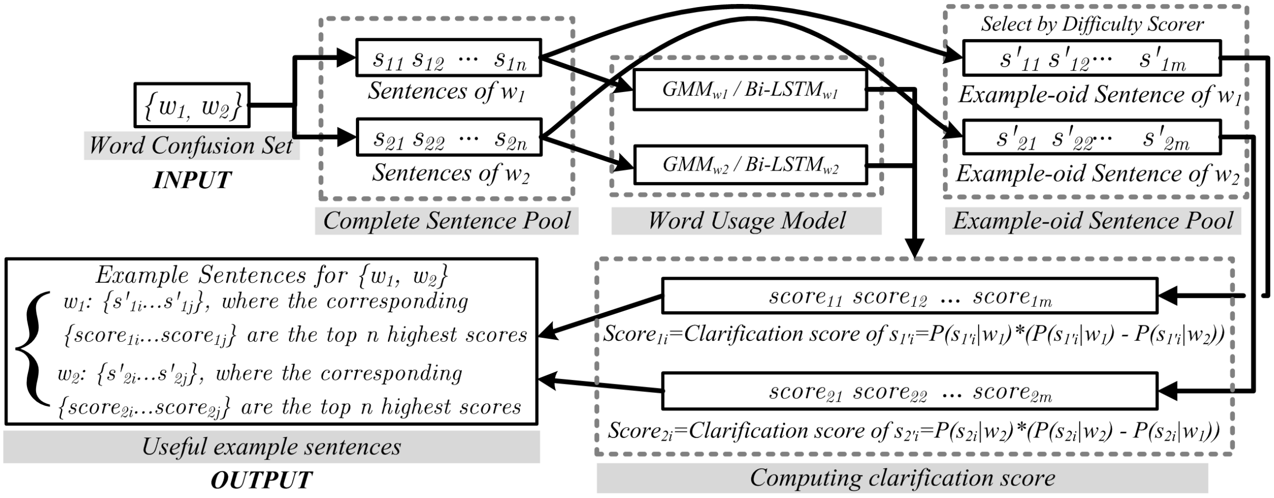
{w1, w2}
Word Confusion Set
INPUT
s11 s12 ··· s1n
Sentences of w1
s21 s22 ··· s2n
Sentences of w2
Complete Sentence Pool
GMMw1 / Bi-LSTMw1
GMMw2 / Bi-LSTMw2
Word Usage Model
Select by Difficulty Scorer
s'11 s'12··· s'1m
Example-oid Sentence of w1
s'21 s'22··· s'2m
Example-oid Sentence of w2
Example-oid Sentence Pool
score11 score12 ... score1m
Score1i=Clarification score of s1'i=P(s1'i|w1)*(P(s1'i|w1) - P(s1'i|w2))
score21 score22 ... score2m
Score2i=Clarification score of s2'i=P(s2i|w2)*(P(s2i|w2) - P(s2i|w1))
Computing clarification score
Example Sentences for {w1, w2}
w1: {s'1i...s'1j}, where the corresponding {score1i...score1j} are the top n highest scores
w2: {s'2i...s'2j}, where the corresponding {score2i...score2j} are the top n highest scores
Useful example sentences
OUTPUT

Figure 1: Workflow of GiveMeExample, when process the input synonym set $\{w_1, w_2\}$.

research problem. Then we design a workflow which breaks down this task into three stages to accomplish it. In the designed workflow, we need to collect candidate example sentences for suggestion, to learn the usage of each near-synonym to suggest appropriate sentences, and to measure the clarification power of each sentence for a set of near-synonyms. The designed models and scoring functions in this workflow are also introduced here.

3.1 Problem Setting

To clearly explain the research problem in this paper, we define it as follows. Given an input near-synonym set $W = \{w_1, w_2, ..., w_n\}$ where $n >= 2$, the output sentence set $S = \{s_1, s_2, ..., s_n\}$ should best clarify the difference among W, where s_n is the example sentence set of w_n. As learners would need more sentences to help them learn the difference, we aim to find effective sentences in s_n. In this paper, we provide up to 5 different example sentences for each near-synonym in the learner study; in real `GiveMeExample` web service, the system provides up to 10 different example sentences.

3.2 Workflow

A three-stage workflow is proposed to solve the defined research problem: building the sentence pool, learning word usage models and measuring the clarification ability. Figure 1 shows the workflow of `GiveMeExample`. In the next paragraph, we use an example to go through all steps in this figure to explain the whole workflow.

Given an input synonym set $\{w_1, w_2\} = \{refuse, disagree\}$, we first collect their corresponding example sentences to build two **complete sentence pools**. Each of them contains 5,000 example sentences. All the sentences in the complete sentence pool are then used to train **word usage models**. At the same time, we build the **example-oid sentence pools** of "*refuse*" and "*disagree*" by filtering out difficult sentences using the **automatic difficulty scorer**. In this step, each example-oid sentence pool contains 500 example-oid sentences. Next, we calculate the **clarification score** for all sentences in the example-oid sentence pool using the **fitness scores** estimated by the word usage models. For example, for "*refuse*", the score for each sentence in its example-oid sentence pool ($score_{1i}$) is calculated by estimating the fitness scores by word usage models of "*refuse*" and "*disagree*". Finally, we propose example sentences with the top n highest scores upon request, as the final result in s_n. In the following sections, we describe the three stages in details in the workflow sequentially.

Table 1: Features of difficulty scorer.

	Syntactic Feature		Lexical Feature
1.	Sentence length	1.	Avg. token length
2.	Avg. dependency depth	2.	Percentage of words longer than 6 characters
3.	# Dependency arcs deeper than 4	3.	Type-token ratio
4.	Deepest dependency / sentence length	4.	Avg. word frequency
5.	Ratio of right dependency arcs	5.	# Words above Taiwan high school level[1]
6.	# Modifiers	6.	Lexical density
7.	# Subordinates	7.	Nouns / verbs
8.	# Prepositional complements	8.	Modal verbs / verbs
		9.	Participles / verbs
		10.	# Relative pronouns
		11.	Pronouns / nouns
		12.	Avg. # Senses per word

3.3 Sentence Pool

Two sentence pools, the complete sentence pool and the example-oid sentence pool are built for each near-synonym. The complete sentence pool is assembled using example sentences from Vocabulary.com[2]. These sentences are from web news organizations such as The Washington Times, The New York Times, BBC, and Reuters, and are good for language learning, as news articles have been widely used for this purpose for their high quality writing and sufficient quantity. However, as news articles often include new words and long sentences, they are often too difficult and need further processing when used with language learners. Therefore, this complete sentence pool built of all sentences with near-synonyms from Vocabulary.com is used only to train

[1] `http://www.ceec.edu.tw/Research/paper_doc/ce37/ce37.htm`

[2] Text from Vocabulary.com (`https://www.vocabulary.com`), Copyright ©1998–2016 Thinkmap, Inc. All rights reserved.

word usage models. When choosing useful example sentences for learners, we use the example-oid sentence pool, in which difficult sentences are removed by the automatic difficulty scorer, which is built to extract dictionary example-oid sentences.

Example-oid sentences are selected from the complete sentenc pool. To build the example-oid sentence pool, the difficulty scorer needs to be created first. To train the scorer, a total of 8,743 example sentences were collected from the website of COBUILD English Usage dictionary[3] [33] and used as positive samples. However, it is difficult to have the gold negative samples. As mentioned, very often sentences from Vocabulary.com are from news articles and can be said to be difficult according to the rules from Pilán's work [28]: they are relatively long and they contain many proper nouns which learners may not know. There could be physical dictionary example-oid sentences in Vocabulary.com but they should be the minority. Hence, we randomly select the same amount (8,743) of example sentences from available sentences in Vocabulary.com as negative samples. The features we used are also based on work of Pilán et al. [28] but with several modifications as their readability assessment is built for Swedish. The detail features are all listed in Table 1. Then Logistic Regression [37] is adopted to learn the difficulty scorer. Finally, 500 example-oid sentences are selected for each near-synonym to form its example-oid sentence pool.

3.4 Word Usage Model

The word usage model is used to estimate the appropriateness of filling a word into a sentence. To give a clearer description, the problem here would be to estimate $P(s|w)$, where w is the target word and $s = w_1, w_2, \cdots, w_{t-1}, \underline{\quad}, w_{t+1}, \cdots, w_{n-1}, w_n$ stands for the sentence with a slot. When training the word usage model for w, we collect thousands of sentences (5,000 in this paper) which contain the near-synonym w, and then utilize the context of the near-synonym w, i.e., s, to learn a model of the appropriate context.

The word usage model is built as an **one-class classifier** to recognize target samples from an unknown sample space. Although multi-class classifiers, such as Support Vector Machine (SVM), usually perform better in discriminating difference among different classes, it is infeasible for us to train all combinations of multi-class classifiers as our goal is to deal with any near-synonym set. For example, if the vocabulary size is n, an one-class approach needs to train only n models but a multi-class approach would need C_2^n or C_3^n models to cover all possible input near-synonym sets (if one set contains only 2 words or 3 words respectively. Having more words in a near-synonym set will make the number of models grows exponentially.) The output probabilities of the word usage model should indicate the probability of the observed data to be classified as a target sample, that is, in our case, the probability that s is appropriate as w's context. We refer to this as the sentence's *fitness score*. For this we use two models: the Gaussian mixture model (**GMM**) [41] and the bi-directional long short-term memory (**BiLSTM**) neural network model.

To train a one-class GMM classifier for w, we utilized as the training samples the 5,000 most updated sentences containing w on Vocabulary.com. For each one-class BiLSTM classifier for w, the same 5,000 sentences were used as the positive samples; we randomly selected another 50,000 sentences which did not contain w as the negative samples for training.

3.4.1 GMM with contextual feature

For the GMM model, we define contextual features which focus on local information within a specific window of size k. Given $sentence = w_1 \cdots w_{i-k} \cdots w_i \cdots w_{i+k} \cdots w_n$, where w_i is the target word and k is the window size, we take the k words preceding and following the target word and represent them as well as their adjacent combinations in sequence using their summation of word embeddings [27]. Empirically, we found the best performance was achieved with $k = 2$ for our task, which yields the contextual features $\{e_{w_{i-2}} e_{w_{i-1}} e_{w_{i-2,i-1}} e_{w_{i+1}} e_{w_{i+2}} e_{w_{i+1,i+2}}\}$, where e_w denotes the summation of word embeddings of word sequence w. Figure 2 also illustrates the contextual features.

We can elaborate how GMM works in the FITB task by its ability of modeling the distribution and semantic of the contexts. For the former, GMM learns the distribution of the given samples. As a result, given the words in four adjacent contextual slots as features, GMM will capture the distribution of these contexts co-occurred with the target word w_i. For the latter, as word embedding is proven to be able to represent the semantic of words, it enables GMM to obtain the semantic of w_i's context. For example, when "he", "she" and "John" appear very often within the window, GMM will understand that words semantically relevant to "human" could appear in w_i's context as the word embeddings of "he", "she", "John" and "human" are close to each other in the semantic space.

We adopt the Bayesian information criterion (BIC) [31] to decide the number of Gaussian mixtures n. BIC introduces a penalty term to the number of parameters in the model to balance underfitting and overfitting. We found the best n was around 50 after several runs of BIC, which was hence set as the value of n in our approach.

3.4.2 Bidirectional LSTM (BiLSTM)

The other approach we proposed to build the word usage model is the bidirectional LSTM neural network [29, 11, 9]. Likewise, we sought to consider the words preceding and following the target word. An advantage of using LSTM is that it utilizes information in the whole sentence instead of being limited to a small window around the target word; GMM is less suitable for this due to its memory and computation demands. Moreover, LSTM's forgetting of distant information emphasizes the neighboring context. Figure 3 shows the BiLSTM model architecture. The preceding and following words are passed into the forward LSTM and backward LSTM respectively, after which the output vectors of the two LSTMs are concatenated together to form the sentence embedding, which is used as the given sentence's feature. We add two fully connected layers to generate the final one-class classifier which predicts whether the word w_i is appropriate for the slot.

LSTM is usually used to capture the information from a sequence such as a sentence according to the input order of its composite components. LSTM is also believed to be able to automatically identify the key components in the given sequence for the current task. As a result, a BiLSTM model could be regard as a model that remembers the preceding and the following sequences, and further optimizes

[3] http://www.wordreference.com/EnglishUsage/

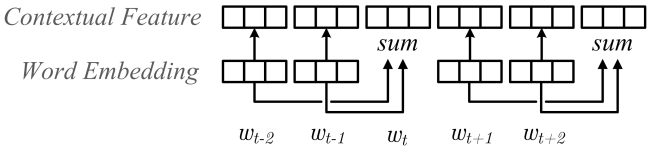

Figure 2: Contextual feature extracted for GMM

the weights of the components (i.e., words) in these two sequences in order to predict w_i.

3.5 Measuring Sentence Clarification Ability

We use a clarification scoring function to model the clarification ability of example sentences. When searching for useful example sentences for the target word w_i in a synonym set W (containing near-synonyms), we consider two factors:

(1) **Fitness**: $P(s|w_i)$, whether word w_i is appropriate for the example sentence s given a slot for w_i. A higher score denotes a more suitable sentence for the word. As Table 2 shows, the context of the near-synonym should be appropriate for it. If the score is low, it probably means that the current usage of the word is rarely seen and thus should not be selected. The probability function $P(s|w_i)$ measures how suitable the word is given the context in the sentence, as discussed in Section 3.4.

(2) **Relative closeness**: $\sum_{w_j \in W - w_i} P(s|w_i) - P(s|w_j)$, the summation of difference of probabilities between $P(s|w_i)$ and $P(s|w_j)$. A high score denotes a better fit of s to w_i and a worse fit to $W - w_i$. For example, in Table 2, the first sentence is suitable only for "refuse" because the word "to" follows the slot but "disagree" would not be followed by it. In this case, the relative closeness of the sentence should be high ($P(s|refuse)$ is high and $P(s|disagree)$ is low), which indicates its good ability to clarify the difference between "refuse" and "disagree". The third sentence is a similar case. However, the second and the forth sentences are confusing as the slots are suitable for both "refuse" and "disagree". In this case, the relative closeness should be low to indicate their slight differentiation power.

We believe these two scores help to retrieve example sentences that elaborate the right usage for w_i (high fitness score) and represent the difference between w_i and other words (high relative closeness). Therefore, we use their product as the clarification scoring function:

$$score(s|w_i) = P(s|w_i) * (\sum_{w_j \in W - w_i} P(s|w_i) - P(s|w_j)) \qquad (1)$$

where $score(s|w_i)$ denotes the clarification score of the example sentence s for w_i. We generate recommendations by ranking sentences in the example-oid sentence pool by their clarification scores. We repeat this procedure for all words in synonym set W to find useful example sentences. Next, we describe how to calculate probability $P(s|w_i)$.

4. EXPERIMENT

We conduct two experiments to evaluate the whole example sentence suggestion framework. The first experiment is to see whether we can find sentences like those handcraft ones included in the dictionaries. The second experiment is to know whether the designed models can find appropriate example sentences for each near-synonym by the FITB task. We detail the experiment procedure in this section.

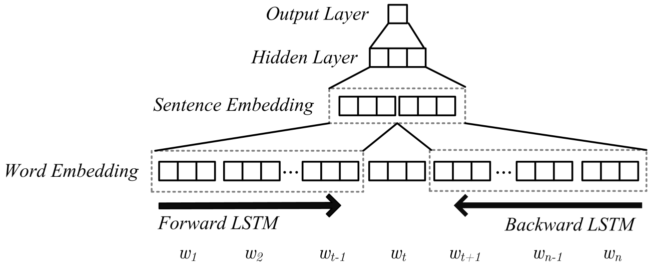

Figure 3: Architecture of the bidirectional LSTM word usage model. The model outputs a probability to measure whether the input context fits the word.

Table 2: Example sentences of *refuse* and *disagree*.

	synonym	context of synonym
1.	refuse	The council pointedly ___ to undertake a careful or studied approach.
2.	refuse	Authorities say they shot him after he ___.
3.	disagree	The state Supreme Court ___ with all of them.
4.	disagree	But some Republicans in his home state ___.

4.1 Evaluation for Finding Example-oid Sentences

To evaluate the ability of finding the dictionary example-oid sentences, we utilize the previously collected 8,743 positive and 8,743 negative example sentences for experiments. We set the testing set ratio to 0.2, 0.25, and 0.5 respectively to see whether the accuracy will vary accordingly. For each testing set ratio, we run 10 experiments by randomly selecting the testing instances from the whole sentence pool each time. The average accuracy of 10 experiments for each testing ratio is reported in Table 5. Results show that the performance is stable for different testing set ratios and the average accuracies are all over 90%, which confirms the reliability of the proposed difficulty scorer on selecting example-oid sentences.

4.2 Evaluation for Fitness by FITB

The fill-in-the-blank (FITB) test was adopted to assess whether the proposed fitness score identifies the appropriate context for a given word. A FITB question contains a sentence with a blank word, with several near-synonym candidate answers. Edmonds [7] suggest the FITB test on 1987 Wall Street Journal (WSJ) and used 1988–1989 WSJ as a training data set. Unlike other multi-class approaches whose outputs are directly the answer, both proposed models solve the FITB question s_{FITB} by comparing the output probabilities of the word usage model of the corresponding candidate words W, and suggest the one with the highest probability.

The accuracy results of two proposed methods and other related work are shown in Table 3. Although SVMs on latent vectors [38] achieved better results, applying multi-class SVM (or other multi-class learning models) in our system increases the complexity from n to n^2 when training C_2^n models (if only two-word synonym sets are considered). As n is

Table 3: Fill-in-the-blank results (accuracy) on WSJ near-synonym sets

	difficult, hard, tough	error, mistake, oversight	job, task, duty	responsibility, burden, obligation, commitment	material, stuff, substance	give, provide, offer	settle, resolve	Avg
Co-occur network	47.90%	48.90%	68.90%	45.30%	64.60%	48.60%	65.90%	55.70%
SVM on latent vectors	61.70%	**82.50%**	**82.40%**	63.50%	78.50%	**75.40%**	77.90%	**74.50%**
5-gram LM	63.20%	78.70%	78.20%	**72.20%**	70.40%	55.80%	70.80%	69.90%
GMM	63.63%	67.89%	81.36%	63.70%	**78.88%**	61.65%	74.41%	70.26%
BiLSTM	**64.14%**	73.66%	74.57%	70.03%	77.29%	72.83%	**78.85%**	**73.05%**

Table 4: Fill-in-the-blank results (accuracy) on 24 word pairs.

	GMM	BiLSTM		GMM	BiLSTM		GMM	BiLSTM
postpone, delay	63.65%	72.89%	sociable, social	99.40%	94.69%	convince, persuade	60.06%	65.57%
finally, eventually	64.55%	72.75%	briefly, shortly	85.97%	90.27%	possibility, opportunity	85.11%	85.96%
manage, arrange	76.56%	82.49%	disagree, refuse	92.78%	93.64%	spoil, destroy	85.66%	92.71%
realize, understand	67.41%	76.41%	advertisement, announcement	81.28%	93.24%	strange, unusual	79.20%	88.25%
usual, ordinary	78.31%	90.01%	fault, mistake	82.47%	80.39%	disease, illness	72.89%	80.44%
feminine, female	70.99%	87.16%	scenery, landscape	72.15%	87.06%	skilled, skillful	78.54%	80.66%
safe, secure	78.69%	84.25%	happen, occur	75.54%	78.95%	embarrassed, ashamed	72.38%	79.05%
floor, storey	95.90%	76.06%	scarce, rare	80.70%	82.18%	alternate, alternative	75.69%	91.04%
						Average	78.16%	83.59%

Table 5: The accuracy of predicting dictionary example-oid sentences by the difficulty scorer.

	Testing Set Ratio		
	0.2	0.25	0.5
Average	91.05%	90.89%	91.17%

Table 6: Example test question set for the near-synonyms skilled and skillful.

skilled vs. skillful
1. It takes four years to train a _____ engineer.
2. As an artist, he was very _____ with a pencil.
3. Weaving was a very _____ job, requiring a five-year apprenticeship.

the vocabulary size, it is generally infeasible to use multi-class learning models. BiLSTM exploits comparably more context and achieves better results, but GMM still achieves satisfactory results. The proposed GMM and BiLSTM both outperform the only one-class related work, the 5-gram language model [15] trained on Google 5-gram 1T and taking into account the whole sentence. This may indicate that remote words still contain useful information for differentiating near-synonyms.

For the later user study, we create a private dataset which contains 24 near-synonym pairs. The details are described later in section 5.2. We also conducted the FITB experiments on these 24 near-synonym pairs using the proposed GMM and BiLSTM models trained by Vocabulary.com for reference. All sentences in WSJ containing these 48 near-synonyms were selected and then the near-synonyms were removed from these sentences to generate the FITB questions for testing. The models need to fill the original near-synonym back into the blank by selecting from two choices (near-synonyms). Results are shown in Table 4, where BiLSTM still outperforms GMM with the accuracy 83.59%. Hence, we conclude that the proposed BiLSTM succeeds in capturing vital features from long-distance words to find nuances among synonyms.

5. LEARNER STUDY

In addition to the proposed techniques, we also care about the effect of learning. Therefore, we include a learner study in this section in order to get insights of the relevance between the performance of the FITB task and the performance of learners. From the results of this learner study, we further summarize our observations.

5.1 Participants

A total of 16 Chinese-speaking EFL college freshmen in a university in Taiwan were recruited from an intact three-credit class entitled ”Freshmen English”. This course met three hours per week for 18 weeks in the semester. The participants - six males and 10 females - were aged between 18 and 19. They had at least six years of formal instruction, from junior to senior high school, and were estimated to be intermediate learners of English, as measured by the proficiency test [2] taken at the beginning of the experiment. Grounded in the statistical analyses, the findings are presented below.

5.2 Materials

To evaluate the effectiveness of the example sentences suggested by our two models (i.e., BiLSTM and GMM) on near-synonym learning, we designed a fill-in-the-blank task. To develop the task, we needed to collect near-synonyms and generate example sentences. First, we selected 85 pairs of near-synonyms from Collins COBUILD English Usage dictionary [33] only if each selected word has more than two example sentences appearing in the dictionary; besides, the total number of the example sentences for the two synonyms should not be less than five. Next, after scrutinizing all synonym pairs and excluding similar-looking words (e.g., ”although” and ”though”), 35 pairs of near-synonyms were collected. Although similar-looking words could be confusing for language learners, they are not near-synonyms. Thus, they were not included in this study. Then,a university English professor examined all the pairs and removed some extremely easy or difficult near-synonyms. Finally, a total of 24 pairs of near-synonyms were determined for developing pedagogical materials.

Regarding example sentences, to assess the effectiveness of the proposed BiLSTM and GMM models on suggesting example sentences, we adopted two resources as additional models (resource-based models) for comparison: two paper-based dictionaries (i.e., Collins COBUILD English Usage dictionary [33] and Longman dictionary of common errors [36]) and an online dictionary (i.e., Vocabulary.com). Specifically, compiled by lexicographers, the example sentences from traditional dictionaries are accurate but limited. On the other hand, Vocabulary.com automatically collects example sentences online, and hence it provides a larger number of example sentences. The purpose of adopting Collins COBUILD English Usage dictionary and Vocabulary.com is to simulate learner behavior of consulting dictionaries. Regarding the test development, we designed a fill-in-the-blank task to examine whether the students were able to use appropriate near-synonyms. For each pair of near-synonym, we designed three fill-in-the-blank questions (as seen in Table 6) using the example sentences of Collins COBUILD English Usage dictionary. Students had to select one synonym to complete the sentence. Note that the odd number of questions is to reduce the possibility of guessing answers. To avoid fatiguing learners, we determined three questions instead of more for each pair of near-synonyms.

5.3 Experimental Design

Concerning the treatment phase, the example sentences suggested by the four models were displayed. To objectively evaluate the performance of the proposed models and keep the largest number of participants for each model, we conducted pairwise comparisons on the four models to determine which of each model had a greater performance. Thus, a total of six pairwise comparisons were conducted. For each comparison, the students were randomly divided into two groups and were provided with the example sentences suggested by two different models. Importantly, to avoid an information deluge, five example sentences at most were prepared for individual words of the 24 pairs of near-synonyms.

A complete administration of the fill-in-the-blank task took 80 minutes. First, all the students were given a pre-test and a background questionnaire (30 minutes). Then five minutes were taken to introduce the `GiveMeExample` system and had students familiarize themselves with the system. The next stage consisted of the treatment phase and a post-test (45 minutes). The students were asked to learn the 24 pairs of near-synonyms by reading the suggested example sentences. Once they finished reviewing all example sentences of individual pairs of near-synonyms, the students had to complete the fill-in-the-blank questions. As each pair connected with 3 questions, the students were to complete a total of 72 questions. Note that the test items in the pre-test and post-test were identical except for their order. The students were then asked to complete a reflection questionnaire.

5.4 Results and Discussions

The results of learner study are shown in Table 7. For each near-synonym set, there are 4 columns. The column "pre" indicates the number of correctly answered FITB questions by learners in the pre-test; "diff" indicates the difference between the numbers of correctly answered questions in the post- and pre-test ("pre"), i.e., the number of correct questions gained after learning from example sentences; "rspd" indicates the learning time, which is the time difference to answer the same question in the pre- and post- test, i.e., the post-test time minus the pre-post time. The learning time will be a little bit under-estimated but very close to the real time to digest the example sentences as learners should need less time to read the same question in the post-test. The last column "#sen" indicates the average number of sentences read by learners before answering the question requested by them. In Table 7, the four models, i.e., Dictionary, Vocabulary, Give-GMM, and Give-BiLSTM, are compared. Results show that the best model is Give-GMM, then Vocabulary, Give-BiLSTM, and then the worst model Dictionary. From this performance order, we can see that overall the digitization is helpful for learning near-synonym words from sentences as Dictionary provides the least assistance. In addition, from results we find observations that show the effectiveness and the efficiency of the proposed models and system. We look into near-synonyms as well as their example sentences, and summarize four key points worth further discussions in the following sections.

5.4.1 *Time to Digest*

Question: Did GiveMeExample select sentences easy for learning?

Efficiency wise, the average learning time when learning with two automatic models is largely shorter than the time of the resource-based model Vocabulary (see (a):Give-BiLSTM rspd=56.07 vs. Vocabulary rspd=74.56; (d):Give-GMM rspd=19.01 vs. Vocabulary rspd=52.27). Note that Dictionary is not compared with the other three models here as it may provide less example sentences which need less time to read. This shows that the automatically proposed example sentences are easier for learners to read, understand, and conclude the appropriate usages of near-synonyms. In other word, with the two major features we design for the system, providing dictionary example-oid and difference-descriptive sentences, suggested sentences can shorten the learning process of near-synonyms.

The system can assist learners even they have different ways to utilize the system. Most of the participants used the proposed sentences to double confirm their answers. However, there are some confident learners who only read example sentences when they are not sure about the answer. In (f) we find negative learning time (rspd < 0), which comes from the negative learning time of one confident learner. This negative number does not indicate that the learner cannot learn with the system, but instead this learner only learns when necessary. We find that this learner did not read the example sentences and answered directly in the post-test for most questions he felt confident, which saved reading time for both the example sentences and the question and hence lead to a negative learning time. He only requested for example sentences for confusing near-synonyms. This is an even more smart and natural way to use the system and

Table 7: Result of Learner Study

(a)	BiLSTM				Vocabulary			
	pre	diff	rspd	#sen	pre	diff	rspd	#sen
postpone, delay	17	-5	83.72	7.6	13	0	127.54	7.0
finally, eventually	16	-8	61.37	8.9	14	0	81.27	8.3
manage, arrange	16	-1	31.14	9.1	20	-4	54.87	7.3
realize, understand	14	2	48.06	9.5	19	2	34.55	6.8
Average	15.8	-3	56.07	8.8	16.5	-0.5	74.56	7.3

(b)	Dictionary				Vocabulary			
	pre	diff	rspd	#sen	pre	diff	rspd	#sen
usual, ordinary	18	-9	26.71	4.0	19	4	34.81	7.6
feminine, female	19	1	26.87	5.9	22	-4	13.38	5.8
safe, secure	21	-1	9.19	5.9	19	-1	42.02	7.0
storey, floor	12	4	13.11	3.9	10	6	15.97	7.3
Average	17.5	-1.3	21.47	4.9	17.5	1.3	26.54	6.9

(c)	BiLSTM				Dictionary			
	pre	diff	rspd	#sen	pre	diff	rspd	#sen
sociable, social	21	1	20.01	7.6	19	3	3.40	2.8
briefly, shortly	21	3	36.27	6.8	23	-1	20.02	4.4
disagree, refuse	22	1	5.15	6.3	22	2	17.74	9.1
advertisement, announcement	16	3	4.56	6.4	20	1	7.03	2.8
Average	20	2	16.50	6.8	21	1.3	12.05	4.8

(d)	GMM				Vocabulary			
	pre	diff	rspd	#sen	pre	diff	rspd	#sen
fault, mistake	8	4	25.03	7.0	13	-3	58.39	8.5
scenery, landscape	19	3	15.96	6.9	17	-1	52.18	8.6
happen, occur	19	0	20.76	7.6	13	6	60.14	8.0
scarce, rare	15	1	14.27	7.5	13	2	38.39	8.1
Average	15.3	1.8	19.01	7.3	14	1	52.27	8.3

(e)	BiLSTM				GMM			
	pre	diff	rspd	#sen	pre	diff	rspd	#sen
convince, persuade	21	2	7.89	4.3	20	-2	12.98	7.0
possibility, opportunity	23	-4	4.06	5.3	21	0	16.75	6.1
spoil, destroy	15	1	34.82	7.9	15	3	18.83	7.9
strange, unusual	16	-1	9.76	6.4	15	1	10.09	9.5
Average	18.8	-0.5	14.13	5.9	17.8	0.8	14.66	7.6

(f)	Dictionary				GMM			
	pre	diff	rspd	#sen	pre	diff	rspd	#sen
disease, illness	16	-4	-1.48	2.6	12	6	13.67	8.5
skilled, skillful	10	-1	-3.18	2.8	11	1	23.34	9.4
embarrassed, ashamed	17	1	11.35	4.0	20	1	32.96	9.0
alternate, alternative	10	6	11.92	5.0	12	3	30.57	9.9
Average	13.3	0.5	4.65	3.6	13.8	2.8	25.14	9.2

this learner indeed succeeded to answer more questions correctly in the post-test.

5.4.2 *Adaquate Quantity*

Question: Can GiveMeExample suggest enough example sentences to learners?

Sentences from Dictionary should have the strongest power to demonstrate the difference between near-synonyms as they are designed by human experts. However, in the learner study, they turn out to be the least effective.

In the learner study, we made the setting of four models as similar as possible. However, there was a natural restriction for Dictionary that the available example sentences were limited to those in the physical dictionaries and could be less than 5. Results show that the number of requested example sentences (#sen) is under 5 for Dictionary, while this number for the other models exceeds 5 and is usually around 7 to 8 (the maximum number is 10, 5 for each near-synonym). From this observation, we can say that in order to get some insight to use near-synonyms appropriately, on average at least 3 example sentences are necessary for each near-synonym. Dictionary fails to provide that many, so its learning performance is the least effective. On the other hand, stats also show that the sentences provided by the system are sufficient as the value of most #sen are not yet close to the maximum value, i.e., 9 or 10. Learners stopping requesting for more example sentences suggests they have learned or they start to feel overwhelmed.

5.4.3 *Sentence Diversity*

Question: How are the example sentences proposed by GiveMeExample different from the conventional resource-based models?

Interestingly, Vocabulary has the performance between that of the two automatic models, Give-GMM and Give-BiLSTM. We find that this is because automatically providing example sentences focusing on the same aspect for learning is sometimes misleading. As aiming at emphasizing major one or two differences for learners to capture them easily, sentences from automatic models are basically less diverse, compared to those selected with no criteria but only up-to-date example sentences from the Vocabulary website. As we have shown by the learning time (rspd), this helps learners learn efficiently. However, this also brings us disadvantages sometimes. If the focused difference by the system is not the one learners should pay attention to for the question or it is vague for learning the usage of near-synonyms, the example sentences may lead learners to the wrong answer. Table 8 shows the misfocused usage by Give-GMM. Here the example sentences show the learners of the usage ”at fault”, while later to give the correct answer learners should know using fault is appropriate for machines. In this example, as in the question the word before the blank is not ”at”, examples mislead learners into selecting ”mistake”. It is true that people use ”at fault” but not ”at mistake”, but this information is not helpful for answering the question correctly. When this misleading happened, some near-synonyms like (a)postpone, delay (diff=-5 by BiLSTM) and (a)finally, eventually (diff=-8 by BiLSTM) suffered. We will see in the next section that compared to Give-GMM, Give-BiLSTM also suffered from providing vague example sentences to learners. Therefore, its learning effectiveness was harmed and hence overall Give-BiLSTM cannot outperform Vocabulary.

Table 8: Misfocused Example Sentences from Give-GMM

	Give-GMM
Fault	
1.	But the benchmark itself may be at fault.
2.	Google has said that self-driving cars were never at fault.
3.	But they do now: The United States is at fault.
Mistake	
1.	My second mistake: not addressing meat eaters here.
2.	However, there are a number of technical mistakes.
3.	It’s the country’s policy mistakes of the past six months.
1.	**Fault vs. Mistake** The machine has developed a ______.

5.4.4 Information Scope for Learning

Question: How the example sentences from two automatic models different from each other?

Give-GMM and Give-BiLSTM achieve similar performance in the FITB task and Give-BiLSTM even performs better (see Table 3 and 4.) However, the learning effectiveness of Give-GMM largely outperforms that of Give-BiLSTM.

To understand the reason behind this phenomenon, we examine the proposed example sentences from Give-BiLSTM and Give-GMM. We found that the example sentences suggested by Give-BiLSTM, which considered the whole sentences, demonstrated the difference of two whole sentences, while those proposed by Give-GMM only emphasized the difference surrounding the near-synonyms. Hence, the differences Give-GMM demonstrated were more focused and local, while those from Give-BiLSTM are global. This leads to the results that learners felt easier to capture the difference between near-synonyms with Give-GMM, while Give-BiLSTM itself learned better for the FITB task but failed to teach learners how to do it.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed the GMM and BiLSTM models to recommend example sentences for learning near-synonyms via the FITB task. Results demonstrated two proposed models, constructed by many one-class classifiers, achieve a comparable performance to the multi-class classifier, and show that we have successfully taken the first step toward solving the example sentence suggestion problem through the proposed `GiveMeExample` system. The online `GiveMeExample` system is available at `http://givemeexample.com/GiveMeExample/`.

We have performed a learner study in this paper and reported the results. They also confirmed that the proposed models can assist language learners on learning near-synonyms from both the effectiveness and efficiency aspects. An interesting and inspiring observation is that the good performance of the model is not always equal to the learning effectiveness the model can bring in. In the future, we will explore linguistic aspects for which our models are not so confident and pay more attention to the effective transformation from the advanced technology to the improvement of the learning result. We believe this is a worthy research direction.

Acknowledgements

Research of this paper was partially supported by Ministry of Science and Technology, Taiwan, under the contract 105-2221-E-001-007-MY3.

7. REFERENCES

[1] M.-H. Chen, S.-T. Huang, J. Chang, and H.-C. Liou. Developing a corpus-based paraphrase tool to improve efl learners' writing skills. *Computer Assisted Language Learning*, 28(1):22–40, 2015.

[2] M.-H. Chen and M. Lin. Factors and analysis of common miscollocations of college students in taiwan. *Studies in English Language and Literature*, 2011.

[3] M. E. Curtis. The role of vocabulary instruction in adult basic education. *Comings, J., Garner, B., Smith, C., Review of Adult Learning and Literacy*, 6:43–69, 2006.

[4] E. Dale and J. S. Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54, 1948.

[5] G. De Melo and G. Weikum. Extracting sense-disambiguated example sentences from parallel corpora. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 40–46. Association for Computational Linguistics, 2009.

[6] J. Didakowski, L. Lemnitzer, and A. Geyken. Automatic example sentence extraction for a contemporary german dictionary. In *Proceedings EURALEX*, pages 343–349, 2012.

[7] P. Edmonds. Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 507–509. Association for Computational Linguistics, 1997.

[8] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[9] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[10] K. Harvey and D. Yuill. A study of the use of a monolingual pedagogical dictionary by learners of english engaged in writing. *Applied Linguistics*, 18(3):253–278, 1997.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] Y.-T. Huang, H.-P. Chang, Y. Sun, and M. C. Chen. A robust estimation scheme of reading difficulty for second language learners. In *Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on*, pages 58–62. IEEE, 2011.

[13] T. Huckin and J. Coady. Incidental vocabulary acquisition in a second language. *Studies in second language acquisition*, 21(02):181–193, 1999.

[14] D. Inkpen. A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):2, 2007.

[15] A. Islam and D. Inkpen. Near-synonym choice using a 5-gram language model. *Research in Computing Sciences*, 46:41–52, 2010.

[16] T. John. Should you be persuaded: Two examples of data-drivenlearning. *Johns TF, King P. Classroom Conlcor-ldanlcing. Birmingham: ELR*, 1991.

[17] T. Johns. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *Perspectives on pedagogical grammar*, 293, 1994.

[18] A. Kilgarriff, M. Husák, K. McAdam, M. Rundell, and P. Rychlỳ. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*, pages 425–432, 2008.

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

[20] D. Liu. Salience and construal in the use of synonymy: A study of two sets of near-synonymous nouns. *Cognitive Linguistics*, 24(1):67–113, 2013.

[21] D. Liu and S. Zhong. L2 vs. l1 use of synonymy: An empirical study of synonym use/acquisition. *Applied Linguistics*, page amu022, 2014.

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

[23] W. Nagy and D. Gentner. Semantic constraints on lexical categories. *Language and Cognitive Processes*, 5(3):169–201, 1990.

[24] I. Nation. Vocabulary size, growth, and use. *The bilingual lexicon*, pages 115–134, 1993.

[25] I. S. Nation. *Learning vocabulary in another language*. Ernst Klett Sprachen, 2001.

[26] P. Nation and J. Newton. Teaching vocabulary. *Second language vocabulary acquisition*, pages 238–254, 1997.

[27] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[28] I. Pilán, E. Volodina, and R. Johansson. 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, 2014.

[29] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[30] S. E. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics, 2005.

[31] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[32] Z. Shiraz and M. Yamini. Investigating the interface between depth of vocabulary knowledge and efl learners' strategy use. *World Applied Sciences Journal*, 14(5):666–673, 2011.

[33] J. Sinclair. *Collins COBUILD English Usage*. Collins, 1992.

[34] J. M. Sinclair. *How to use corpora in language teaching*, volume 12. John Benjamins Publishing, 2004.

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

[36] N. D. Turton and J. B. Heaton. *Longman dictionary of common errors*. Longman, 1996.

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

[38] T. Wang and G. Hirst. Near-synonym lexical choice in latent semantic space. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1182–1190. Association for Computational Linguistics, 2010.

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

[40] S. Webb. The effects of synonymy on second-language vocabulary learning. *Reading in a Foreign Language*, 19(2):120, 2007.

[41] L. Xu and M. I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

[42] Y. Yeh, H.-C. Liou, and Y.-H. Li. Online synonym materials and concordancing for efl college writing. *Computer Assisted Language Learning*, 20(2):131–152, 2007.