Carrier Sentence Selection with Word and Context Embeddings

Chak Yan Yeung, John Lee and Benjamin Tsou
Department of Linguistics and Translation
City University of Hong Kong
chak.yeung@my.cityu.edu.hk, jsylee@cityu.edu.hk, rlbtsou@cityu.edu.hk

***Abstract*—This paper presents the first data-driven model for selecting carrier sentences with word and context embeddings. In computer-assisted language learning systems, fill-in-the-blank items help users review or learn new vocabulary. A crucial step in automatic generation of fill-in-the-blank items is the selection of carrier sentences that illustrate the usage and meaning of the target word. Previous approaches for carrier sentence selection have mostly relied on features related to sentence length, vocabulary difficulty and word association strength. We train a statistical classifier on a large-scale, automatically constructed corpus of sample carrier sentences for learning Chinese as a foreign language, and use it to predict the suitability of a candidate carrier sentence for a target word. Human evaluation shows that our approach leads to substantial improvement over a word co-occurrence heuristic, and that context embeddings further enhance selection performance.**

***Keywords*-computer-assisted language learning; carrier sentence; word embeddings; context embeddings;**

I. Introduction

Learners of a foreign language often wish to practice and reinforce their linguistic knowledge with exercises. Traditional textbooks, however, can provide only a limited quantity of exercises. In the field of intelligent computer-assisted language learning (ICALL), the task of automatic exercise generation can address this need [1]. Automatic generation can potentially provide users not only with an increased quantity of exercises, but also exercises that fit their needs better in terms of pedagogical level and personal interest.

Many ICALL systems offer fill-in-the-blank (FIB) items, also known as cloze or gap-fill items, as exercises. FIBs are constructed on the basis of *carrier sentences* [2], which are also referred to as seed sentences [1]. As shown in Table I, the system blanks out one word — called the *target word* — in the carrier sentence, and asks the learner to fill in the blank, typically by selecting from multiple choices that include the target word and several distractors.

FIB items may be used for multiple purposes. When supporting language assessment, they should be able to discriminate among students at different proficiency levels. For this purpose, it would be appropriate to select challenging carrier sentences, for example those with complex syntactic structure, difficult vocabulary, or unusual usage of the target word.

In the context of language learning, which forms the focus of this paper, FIB items are intended to help users review the target word and understand it better. Therefore, relatively short, simple carrier sentences that illustrate a common usage of the target word are generally considered the most effective [3], [4].

Fill-in-the-blank item	Component
He stayed home because of the wet ___.	Carrier sentence
A. air	Distractor
B. climate	Distractor
C. weather	Target word
D. whether	Distractor

Table I
Components of an example fill-in-the-blank item with target word "weather"

Carrier sentences are traditionally composed by hand. In order to reduce the cost of manual composition and to take advantage of authentic material for language learning, there has been increasing interest in automatic selection of carrier sentences from large corpora [2], [5], [6]. This paper investigates data-driven methods to assess the suitability of candidate carrier sentences for language learning.

II. Carrier Sentence Selection

The growing size of available digital corpora means that one can find sample sentences for almost any target word. It does not suffice, however, to randomly select a sentence to serve as carrier sentence. Consider the candidate carrier sentences in Table II for the word "weather". Sentence (a) is probably too complex for the user to understand, given that the user is learning the word "weather" and therefore unlikely to be an advanced speaker of English. Sentence (b) is easy to understand, but is not a good carrier sentence since it does not provide a good illustration of the target word. Sentence (c), a straightforward sentence with the collocation "weather" and "wet", is the most appropriate carrier sentence among the three.

One line of research is to adopt readability assessment algorithms to the sentence level [7]. Automatic readability assessment assigns each paragraph or sentence a grade, or score, that indicates its linguistic complexity [8], [9], [10]. This task is related to carrier sentence selection in that it can help filter out sentences with difficult vocabulary or structure, e.g., sentence (a) in Table II. Readability assessment, however, cannot identify simple sentences that do not provide a good illustration of the meaning of the target word. Sentence (b) in Table II is one such example. Proper assessment for carrier sentences, similar to the selection of example sentences for dictionaries, requires

978-1-7281-5014-7/19/$31.00

Suitability	Candidate carrier sentence
×	(a) Humid ___ is imminent as the low-pressure system approaches.
×	(b) He likes to talk about the ___.
√	(c) He stayed home because of the wet ___.

Table II
EXAMPLE CANDIDATE CARRIER SENTENCES FOR THE TARGET WORD "WEATHER"

Type	Suitable	Unsuitable
# sentences	867,838	867,838
# target words	19,845	19,845
# sentences / target word	43.73	43.73
Sentence length	12.03	26.20

Table III
STATISTICS ON THE SUITABLE AND UNSUITABLE CARRIER SENTENCES IN THE TRAINING SET.

taking into account the relation between the target word and the rest of the sentence. To identify sentence (c) in Table II as a promising candidate, for example, the system needs to recognize that the target word is associated with one of its properties ("wet") and a consequence ("stayed home").

Because of the lack of large-scale, annotated datasets for carrier sentences, previous approaches either relied on heuristics [11], [12], [13] or applied machine learning to mostly surface features such as sentence length and vocabulary difficulty. This paper presents and evaluates the first data-driven model for selecting carrier sentence that exploits word and context embeddings. We compile a large-scale corpus, of close to 868K carrier sentences in Chinese harvested from online material, to train a statistical classifier. Experimental results show that combining context embeddings and word-occurrence statistics leads to optimal performance, outperforming baseline classifiers as well as the best heuristic reported in previous work.

The rest of the paper is organized as follows. The next section summarizes previous work. Section IV describes our dataset. Section V presents our classification approach. Section VI reports cross-validation results and motivates our feature selection. Section VII reports results on a human-annotated evaluation dataset. Finally, Section VIII concludes and outlines future work.

III. PREVIOUS WORK

This section first summarizes heuristic and rule-based approaches on carrier sentence selection (Section III-A), and then describes more recent work on machine learning approaches (Section III-B).

A. Heuristic approaches

Volodina et al. [11] proposed an algorithm that uses weighted heuristics to score each candidate sentence in Swedish. These heuristics specify the range of acceptable sentence lengths; penalize the presence of words below a word frequency threshold; and the preferred location of the target word. In manual evaluation, 56.6% of the sentences were considered "acceptable". The fourth requires the presence of finite verbs. A subsequent study considered additional heuristics involving context independence and sentence structure, as well as machine learning for L2 complexity. In an evaluation, 80% of the selected sentences were satisfactory in terms of context independence, and 64% adhered to the expected CEFR level [14].

A more recent study further investigated the use of lexical similarity and word co-occurrence measures [13]. The algorithm first filtered out unsuitable sentences with heuristics similar to those proposed by [11]. Two criteria were then evaluated for sentence selection. The first optimizes lexical similarity, as approximated by word2vec scores between the target word and the other words in the sentence. The second focuses on word co-occurrence, as measured by pointwise mutual information (PMI) between the target word and other words in the sentence. Human evaluation shows that the PMI criterion identified carrier sentence of higher quality.

B. Machine learning approaches

Recent research has also begun to investigate machine learning approaches. A Croatian corpus of 1094 sentences was annotated on a four-point scale from "Very good" to "Very bad". A Random Forest regressor was trained on this corpus to predict the quality of example sentences, using 23 variables corresponding to sentence and word length, frequency and other properties, proportions of words in certain parts-of-speech (POS) such as pronouns, proper nouns and conjunctions, as well as syntactic complexity, as measured by parse tree depth. The best model achieved 89.3% precision in selecting the top three sentences [15].

For German, a supervised machine learning approach was developed to refine the results of a heuristics-based approach for selecting example sentences for dictionary headwords [6]. An SVM classifier, trained on a corpus of 13,000 example sentences annotated as "Good" or "Bad", used features based on lemma and part-of-speech subsequences, as well as sub-tree patterns from constituent trees. Experimental results show that the classifier achieved 68% accuracy [16].

IV. DATA

We constructed a training set consisting of an equal number of suitable carrier sentences and unsuitable ones (Section IV-A), and a human-annotated test set (Section IV-B). We will use the training set to select features for the classifier (Section VI), and then test the classifier on the human dataset (Section VII).

A. Training set

Since there is no large-scale corpus of sentences annotated with their suitability as carrier sentences, we constructed our training set automatically from online resources. Detailed statistics of this set are shown in Table III.

1) Suitable carrier sentences: Many websites offer exercises for learners of Chinese as a foreign language, usually in the form of flashcard-style questions and fill-in-the-blank items.[1] However, these resources offer only limited amounts of example sentences. To collect a large dataset, we used sentences on the *Zaoju Cidian* website[2], which provides example sentences for Chinese word usage, as samples of suitable carrier sentences.

To identify candidate target words for which language learners are likely to request fill-in-the-blank exercises, we retrieved the 40,000 most frequent words in Chinese Wikipedia, excluding words on the stopword list.[3] We retained only those target words for which the *Zaoju Cidian* website provides at least 30 example sentences. This process harvested a total of there 19,845 target words, with an average of 43.73 example sentences each.

2) Unsuitable sentences: For each target word, we obtained an equal number of sentences from Chinese Wikipedia to serve as samples of unsuitable carrier sentences. We recognize that some Wikipedia sentences can be suitable for this purpose; indeed, some of the sentences that were deemed suitable by human raters in the test set (Section IV-B) were taken from Wikipedia.[4] While the noise could have been reduced by deliberately choosing low-quality sentences, that would have hurt the classifier's ability to distinguish between sentences that are merely grammatical and fluent, and those truly suitable as language learning material. We therefore accepted the noise to enable automatic compilation of a large-scale, open-source corpus. Our evaluation results show the effectiveness of this corpus despite the noise.

A significant difference between the sentences from *Zaoju Cidian* and Wikipedia is their length. The average length among the former, with only 12.0 words, is much shorter. Sentence length is an attribute that can be easily controlled and does not require machine learning. We therefore did not use sentence length as a feature (Section V). Instead, we sampled Wikipedia sentences whose length is closest to the average length of the *Zaoju Cidian* sentences.

B. Test set

We used the human-annotated dataset in [13] as test set. This dataset contains 100 target words, each with four carrier sentences: one taken from a textbook, and the other three drawn from Chinese Wikipedia by three different heuristics. To each sentence, two human judges assigned a Word Score ("good", "fair", or "unacceptable") that assesses how well it illustrates a typical usage of the target word, and a Sentence Score on the same scale to assess the grammaticality and fluency of the sentence.

[1]E.g., Clavis Sinica (clavisinica.com) and Du Chinese (duchinese.net).

[2]Accessed at www.ichacha.net/zaoju/ in November 2018.

[3]We scraped sentences from a dump downloaded in April 2016. The stopword list was taken from https://github.com/stopwords-iso/stopwords-zh/blob/master/stopwords-zh.txt.

[4]The training set excludes sentences in the human-annotated test set.

C. Evaluation metric

We converted the two scores in the test set (Section IV-B) into a gold label — either "suitable" or "unsuitable" — for evaluating our classification approach. Sentences annotated as "good" by both judges for both the Word Score and the Sentence Score were labelled as "suitable"; those annotated as "unacceptable" by both judges for *either* the Word Score *or* the Sentence Score were labelled as "unsuitable".

By this metric, the test set contains 226 "suitable" carrier sentences and 72 "unsuitable" ones, covering all 100 target words.

V. Approach

Given an input sentence S that contains the word t, our task is to predict whether S is "suitable" or "unsuitable" as carrier sentence for the target word t. Let W represent the set of words in S, excluding t and those on a stopword list[5]. We investigated two main classes of features.

A. Features on vocabulary difficulty

The following features characterize the level of difficulty of the words in the sentence.

1) Vocabulary difficulty: To ensure the learner can understand the sentence, carrier sentences with easier vocabulary items are preferred. Word frequency is often used as a proxy for its difficulty level. We split the 40,000 most frequent words in Chinese Wikipedia into 200 buckets, each with 200 words, then calculate the proportions of words in W that fall into each bucket.

2) Relative vocabulary difficulty: Vocabulary difficulty is relative to the user's proficiency. Rather than attempting to assess the user's vocabulary proficiency, we determine its upper limit with the following assumption: since the user requested an exercise for word t, he is likely to find that word difficult, and therefore he can be expected to have difficulty understanding other words at the same difficulty level or above. For example, a carrier sentence designed to teach the word "weather" should not assume the learner to know more advanced words, such as "approach".

A simple rule-based approach would be to reject a sentence if any word $w \in W$ has a lower word frequency than the target word t [13]. This heuristic can lead to unnecessary rejection of good candidate sentences. Instead, our feature allows the classifier to learn the acceptable proportion of words in W that are harder than t from training data.

B. Features on sentential context

The following features characterize the relation between W and the target word t, following the intuition that in a good carrier sentence, one or several words $w \in W$ should form a construction with t that illustrates a typical usage.

[5]We used the list of 748 Chinese function words from https://gist.github.com/dreampuf/5548203.

1) Word co-occurrence: We measure co-occurrence with pointwise mutual information (PMI). For each word $w \in W$, we calculated $pmi(w, t)$, the PMI score between w and t, as estimated on sentences in Chinese Wikipedia. This feature then takes the highest PMI score.

Jiang and Lee [13] selected sentences that maximized $pmi(w, t)$, and reported that this heuristic yielded the best performance. We will refer to this baseline as the "**word co-occurrence heuristic**".

2) Example-based word co-occurrence: Sentence S is likely to be a good carrier sentence if it contains word occurrences that are also attested among the suitable sentences in our dataset with target word t. Specifically, for each word $w \in W$, we calculate the proportion of suitable sentences that contain w. This feature takes the maximum proportion.

3) Word embeddings: Word embeddings have been shown to be effectivess in measuring word similarity and relatedness in a large range of NLP tasks. For each word $w \in W$, we calculate the cosine similarity between the word embeddings of w and t, taking the maximum score as the feature. We used word embeddings trained by Skip-gram with negative sampling on Baidu Encyclopedia [17].

4) Context embeddings: Context embeddings have also been shown to be useful in predicting word similarity and relatedness [18]. For each word $w \in W$, we calculate the cosine similarity between the context embeddings of t and the word embeddings of w. This asymmetric measure predicts how likely t or similar words can be found in the context of w. The word embeddings and context embeddings were both trained by Skipgram with negative sampling on Baidu Encyclopedia [17].

VI. Experiment

To assess the usefulness of the features proposed in Section V, we perform cross-validation on our training set (Section IV-A) to measure their performance in classifying candidate carrier sentences as suitable or unsuitable.

A. Set-up

To ensure our evaluation data represent target words at various levels of difficulty, we split the 19,845 target words in our dataset (Section IV) into four difficulty levels:

- "Level 1", the easiest, consists of the 5000 words with highest frequencies in Chinese Wikipedia;
- "Level 2" consists of the next 5,000 most frequent words;
- "Level 3" consists of the next 5,000;
- "Level 4" consists of the remaining 4,845 words, and therefore is the most difficult.

We then randomly selected 1,000 words from each of the four difficulty levels for use in this experiment. We trained a logistic regression model with scikit-learn and performed word segmentation with the Stanford Chinese parser [19].

B. Results

Table V shows the overall classification accuracy on 10-fold cross validation, as well as the breakdown into the four difficulty levels. The figures represent the average accuracy for each of the 4,000 target words, i.e. each target word has equal weight.

Vocabulary difficulty achieved 67.35% accuracy when used alone, with each additional feature significantly improving the classification accuracy[6] and producing the best performance at 83.02% accuracy. Word co-occurrence is shown to be already a strong predictor, yielding an accuracy of 81.13%. The addition of the example-based word co-occurrence statistics from our dataset further improved the accuracy to 82.77%, indicating that our corpus contains word co-occurrence information about specific target words and their carrier sentences that is not covered by PMI scores. Word embeddings features increased the accuracy to 82.91% and the use of context embeddings further improved the accuracy to 83.02%. This matches the previous observation that context embeddings are useful in predicting asymmetric association between words [18].

C. Impact of target word difficulty

Accuracy in carrier sentence quality assessment depends on the difficulty of the target word. When only vocabulary difficulty features were used, the more difficult the target word, the easier it is to predict whether a carrier sentence is suitable for it. In our dataset, most words in a suitable carrier sentence are at a lower difficulty level than the target word. This requirement can be more easily realized when the target words are more advanced. In contrast, easier target words have a smaller pool of carrier sentence candidates that can fulfill this requirement.

The pattern was reversed when word co-occurrence and embeddings features were added in. This could be because more advanced target words tend to have more characteristic word usage, so it is more likely for the carrier sentences and the Wikipedia sentences to have similar contexts, and thus lessening the use of co-occurrence and embeddings features. In contrast, easier target words can be found in a larger pool of different contexts, and so co-occurrence and embeddings features can play a bigger role in distinguishing whether a carrier sentence is suitable.

VII. Evaluation on human data

We now evaluate our classifier on human-annotated data. We trained multiple logistic regression models with scikit-learn [20] using features listed in Section V. Table V shows classification results on the test set, contrasting the results obtained when using single features and feature sets; when using the word-occurrence heuristic, which yielded the best result in [13]; and hand-crafted textbook materials.

[6] At $p \leq 0.001$ by McNemar's test for all cases.

Feature set	All	Level 1	Level 2	Level 3	Level 4
Vocabulary difficulty	67.35%	66.18%	67.72%	67.72%	**67.86%**
+ Relative vocabulary difficulty	68.09%	64.87%	67.89%	69.76%	**69.98%**
+ Word co-occurrence	81.13%	**82.93%**	81.87%	80.98%	78.79%
+ Example-based word co-occurrence	82.77%	**84.82%**	83.70%	82.80%	79.82%
+ Word embeddings	82.91%	**85.13%**	83.93%	82.94%	79.63%
+ Context embeddings	83.02%	**85.23%**	84.16%	83.11%	79.63%

Table IV
CARRIER SENTENCE CLASSIFICATION ACCURACY WITH BREAKDOWN TO DIFFERENT DIFFICULTY LEVELS

Model	Feature(s)	Precision	Recall	F-score
Single-feature	Vocabulary difficulty	0.83	0.49	0.61
	Relative vocabulary difficulty	0.84	0.40	0.54
	Word co-occurrence	0.76	0.92	**0.83**
	Example-based word co-occurrence	0.83	0.08	0.15
	Word embeddings	0.71	0.53	0.60
	Context embeddings	0.68	0.46	0.55
Feature sets	Word co-occurrence + Relative vocabulary difficulty	0.78	0.85	0.81
	Word co-occurrence + Example-based word co-occurrence	0.76	0.88	0.81
	Word co-occurrence + Word embeddings	0.77	0.93	0.84
	Word co-occurrence + Context embeddings	0.77	0.94	**0.85**
Heuristic	Word co-occurrence heuristic	0.50	n/a	n/a
Human	Textbook	0.70	n/a	n/a

Table V
CLASSIFIER PERFORMANCE ON IDENTIFYING SUITABLE CARRIER SENTENCES IN THE TEST SET

A. Baseline and human performance

Using our metric (Section IV-B), the word co-occurrence heuristic (Section V-B) performs at 0.50 precision. The human-crafted sentences, drawn from textbooks, were considered "suitable" 70% of the time. Note that the recall for these two approaches cannot be estimated.

B. Word occurrence only

Among single-feature models, word co-occurrence achieved the best F-score (0.83), at 0.76 precision and 0.92 recall. This model, significantly better than any other individual features[7], is consistent with previous observation of the effectiveness of the use of PMI [13]. However, this model achieves substantially higher precision (0.76) than the word co-occurrence heuristic (0.50) since it was able to learn the lower-bound of suitable PMI scores from the training data in order to prevent false positives. Direct comparison with the heuristic, however, is not possible since its recall is not known.

The addition of the Relative vocabulary difficulty feature degraded the performance of the word co-occurrence model to an F-score of 0.81.

C. Enhancement with context embeddings

Word embeddings were useful in supplementing the word co-occurrence feature, increasing the F-score to 0.84. It was Context embeddings, however, that combined with word co-occurrence to achieve the best results (0.85 F-score), at 0.77 precision and 0.94 recall. The improvement over "Word-occurrence" model is however not statistically significant.[8], which likely reflected the limited size of the test set.

[7]At $p \leq 0.001$ by McNemar's test for all cases.
[8]At $p \leq 0.23$ by McNemar's test.

Since context embeddings predict asymmetric associations between words [18], they indicate how likely the target words can be found in the context of the words in the carrier sentences. They thus in effect predict if there are cue words in the sentence to elicit the correct answer for the blank (i.e., the target word), serving as a good proxy for the ability of the sentence to illustrate the usage of the target word.

VIII. CONCLUSIONS

We have presented the first data-driven model on carrier sentence selection that exploits word and context embeddings. Using a statistical classifier trained on a large corpus of automatically annotated sample sentence, our best model achieved 85% F-score in predicting the suitability of a sentence for a target word. Our experimental results show that context embeddings improve upon word co-occurrence statistics in this task, outperforming the previous best heuristic based on pointwise mutual information statistics.

We have applied this model to select carrier sentences from the LIVAC (Linguistic Variation in Chinese Speech Communities) synchronous corpus [21], as part of our project to build a computer-assisted language learning system for Chinese as a foreign language. In future work, we wish to pursue two directions. First, we intend to investigate carrier sentence selection methods that distinguish between multiple senses of a word [22]. Second, we plan to investigate other features to further improve the results and to evaluate the model on other languages.

ACKNOWLEDGMENT

This work was supported by the Innovation and Technology Fund (Ref: ITS/389/17) of the Innovation and

Technology Commission, the Government of the Hong Kong Special Administrative Region; and by a Strategic Research Grant (#7004941) from City University of Hong Kong. We thank Ka Po Chow for his assistance with this research.

REFERENCES

[1] E. Sumita, F. Sugaya, and S. Yamamoto, "Measuring Non-native Speakers Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions," in *Proc. 2nd Workshop on Building Educational Applications using NLP*, 2005.

[2] S. Smith, P. V. S. Avinesh, and A. Kilgarriff, "Gap-fill Tests for Language Learners: Corpus-Driven Item Generation," in *Proc. 8th International Conference on Natural Language Processing (ICON)*, 2010.

[3] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment," *Applied Measurement in Education*, vol. 15, no. 3, pp. 309–333, 2002.

[4] W. Xu, "A Research on Blanked Cloze Exercises in Intermediate TCSL Comprehensive Textbooks Taking Four Textbooks as Examples [in Chinese]," in *Ł (Proc. 5th Forum of CFL Graduate Students)*. Beijing, China: School of Chinese as a Second Language, Peking University, 2012.

[5] A. Kilgarriff, M. Husák, K. McAdam, M. Rundell, and P. Rychlý, "GDEX: Automatically Finding Good Dictionary Examples in a Corpus," in *Proc. EURALEX*, 2008.

[6] J. Didakowski, L. Lemnitzer, and A. Geyken, "Automatic Example Sentence Extraction for a Contemporary German Dictionary," in *Proc. EURALEX*, 2012.

[7] I. Pilán, E. Volodina, and R. Johansson, "Rule-based and Machine Learning Approaches for Second Language Sentence-level Readability," in *Proc. 9th Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[8] E. Pitler and A. Nenkova, "Revisiting readability: a unified framework for predicting text quality," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[9] R. J. Kate, X. Luo, S. Patwardhan, M. Franz, R. Florian, R. J. Mooney, S. Roukos, and C. Welty, "Learning to Predict Readability using Diverse Linguistic Features," in *Proc. 23rd International Conference on Computational Linguistics (COLING)*, 2010, pp. 546–554.

[10] Y.-T. Sung, J.-L. Chen, J.-H. Cha, H.-C. Tseng, T.-H. Chang, and K.-E. Chang, "Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning," *Behavior Research Methods*, vol. 47, pp. 340–354, 2015.

[11] E. Volodina, R. Johansson, and S. J. Kokkinakis, "Semi-automatic Selection of Best Corpus Examples for Swedish: Initial Algorithm Evaluation," in *Proc. Workshop on NLP in Computer-Assisted Language Learning*, 2012.

[12] I. Pilán, E. Volodina, and R. Johansson, "Automatic Selection of Suitable Sentences for Language Learning Exercises," in *Proc. EUROCALL*, 2013.

[13] S. Jiang and J. Lee, "Carrier Sentence Selection for Fill-in-the-blank Items," in *Proc. 4th Workshop on Natural Language Processing Techniques for Educational Applications*, 2017, pp. 17–22.

[14] I. Pilán, E. Volodina, and L. Borin, "Candidate Sentence Selection for Language Learning Exercises: from a Comprehensive Framework to an Empirical Evaluation," *Traitement Automatique des Langues (TAL) Journal, Special issue on NLP for Learning and Teaching*, vol. 57, no. 3, pp. 67–91, 2017.

[15] N. Ljubešić and M. Peronja, "Predicting Corpus Example Quality via Supervised Machine Learning," in *Proc. Electronic Lexicography in the 21st Century Conference (eLex)*, 2015, pp. 477–485.

[16] L. Lemnitzer, C. Pölitz, J. Didakowski, and A. Geyken, "Combining a Rule-based Approach and Machine Learning in a Good Example Extraction Task for the Purpose of Lexicographic Work on Contemporary Standard German," in *Proc. Electronic Lexicography in the 21st Century Conference (eLex)*, 2015, pp. 21–31.

[17] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du, "Analogical Reasoning on Chinese Morphological and Semantic Relations," *arXiv preprint arXiv:1805.06504*, 2018.

[18] F. T. Asr, R. Zinkov, and M. Jones, "Querying Word Embeddings for Similarity and Relatedness," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 675–684.

[19] R. Levy and C. D. Manning, "Is it harder to parse Chinese, or the Chinese Treebank?" in *Proc. ACL*, 2003.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[21] B. K. Tsou and O. Y. Kwong, "LIVAC as a Monitoring Corpus for Tracking Trends beyond Linguistics," *Journal of Chinese Linguistics Monograph Series*, no. 25, pp. 447–472, 2015.

[22] A. Geyken, C. Pölitz, and T. Bartz, "Using a Maximum Entropy Classifier to Link "Good" Corpus Examples to Dictionary Senses," in *Proc. Electronic Lexicography in the 21st Century Conference*, 2015, pp. 304–314.