Learning to Link Grammar and Encyclopedic Information to Assist ESL Learners

Jhih-Jie Chen[1], **Ching-Yu Yang**[1], **Pei-Chen Ho**[2], **Ming Chiao Tsai**[1],
Chia-Fang Ho[2], Kai-Wen Tuan[2], Chung-Ting Tsai[2], Wen-Bin Han[1], Jason S. Chang[1]
[1]Department of Computer Science
National Tsing Hua University
[2]Institute of Information Systems and Applications
National Tsing Hua University
{jjc, chingyu, patina, jason}@nlplab.cc

Abstract

We introduce a method for learning to extract vocabulary and encyclopedic information to assist second language (L2) learners acquiring deep knowledge of target vocabulary. In our approach, grammar patterns, collocations, representative examples are extracted, aimed at providing rich lexical information for any target words. The method involves word sense disambiguation on target words, automatically parsing the sentences in a large-scale corpus, automatically generating grammar patterns, collocations, examples, and quizzes for every target word, and automatically linking named entities to corresponding Wikipedia information. We present a prototype vocabulary learning system, *Linggle Booster*, that applies the method to corpora and web pages. Evaluation on a set of target words shows that the method has reasonably good performance in terms of generating useful and correct information for vocabulary learning.

1 Introduction

Many English learners read articles and watch videos on the Web everyday to improve their language skills, and an increasing number of services uses Web-based content to assist learning languages. For example, *VOA Learning English*[1] provides level-appropriate articles with a vocabulary list. Websites, such as *VoiceTube*[2], allow learners to watch English videos and read English subtitles with on-demand Chinese translations of vocabulary. *WordBooster*[3], highlights target words in user submitted articles, and provides vocabulary quizzes for users to learn and self-assess vocabulary and reading comprehension skills. These web services, however, do not support easy customization for different users' English proficiency level, nor do they provide other lexical information than definitions and examples. The lack of grammar patterns and collocations makes it inefficient for learners to acquire rich vocabulary knowledge.

To facilitate a more efficient learning process, we develop a prototype interactive system, *Linggle Booster*[4]. At run-time, *Linggle Booster* starts with an URL or text submitted by user, and then generates a reformatted, reader-friendly content in the left column of our system. In the column, vocabulary that fit user's English proficiency level is underlined and words linkable to Wikipedia information are shown in blue (see Figure 1). By clicking on an underlined word, the system will provide the Chinese definition of the target word. The most appropriate definition (e.g., 決賽 in Figure 1) is presented in the first line under the target word, along with other senses appended under the definition (e.g., 期中考試 in Figure 1). Additionally, we offer grammar patterns, collocations and examples of the target word with native language support (i.e., translation in learners' native language). Additionally, *Linggle Booster* also identifies and displays relevant Encyclopedic information (e.g., Wikipedia) to provide another level of information to users. Furthermore, a quiz is generated based on vocabulary in input text for self-assessment (see Figure 2).

2 Related Work

Learning English as a Second Language (ESL) has been an area of active research. For example, many researches have done on autonomous language learning (e.g., Kormos and Csizer (2014)) and on ESL learning strategy on the part of teachers (e.g., Richards and Renandya (2002)). In the field closely related to our work, the Common European Framework of Reference (CEFR)

[1] learningenglish.voanews.com
[2] tw.voicetube.com
[3] wordbooster.com
[4] https://read.linggle.com/

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 213–218
Florence, Italy, July 28 - August 2, 2019.

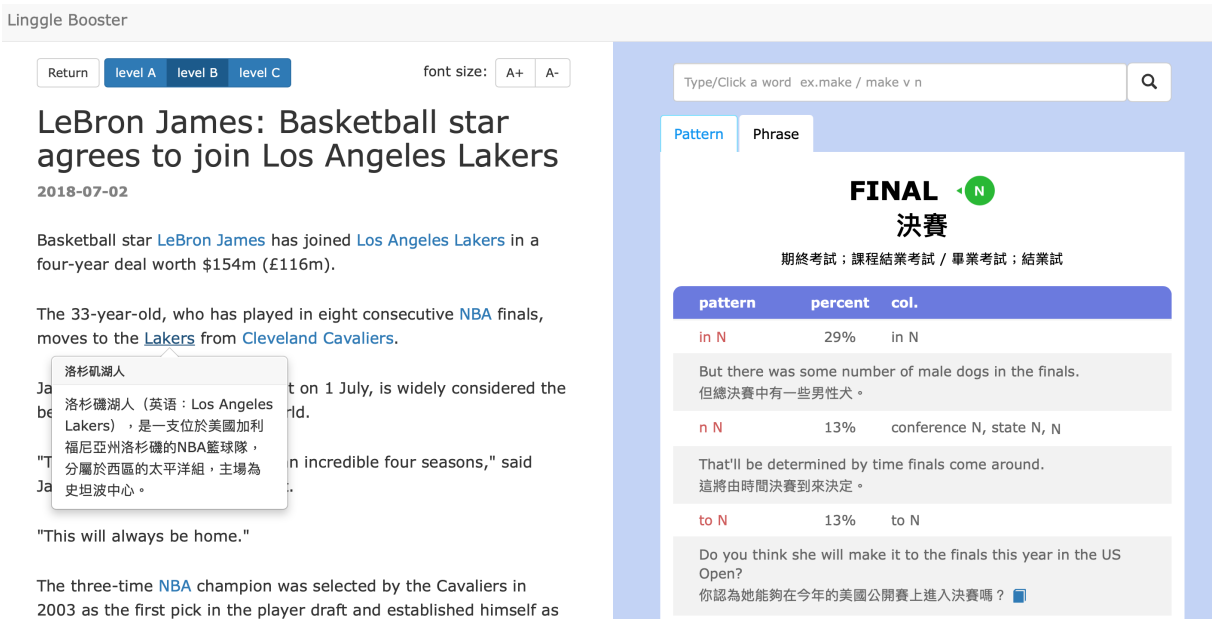

Figure 1: An example *Linggle Booster* session for the user-selected web page [5], presenting the reformatted article in the left column, where Wikipedia information shown in a pop-up, and we provide the following vocabulary information for the highlighted word, *finals*: Chinese translations of the word sense, grammar patterns, collocations, and examples in the right column.

describes what language learners can do at six language stages (i.e., A1, A2, B1, B2, C1, and C2), which has a major effect on language exams and course material design. Stemmed from CERF, Cambridge University Press compiles the English Vocabulary Profile which classifies words and phrases by CERF levels. In our system, we perform word sense disambiguation on user-submitted content and label words with simplified CERF level (i.e., A, B, C) offered by Cambridge online Dictionary and English Vocabulary Profile.

In the field of computer-assisted English learning, there have been an increasing interest in helping second language learners acquire the grammatical usage of a target word. Hunston et al. (1996) and Francis et al. (1998) manually mapped out lexical grammar patterns for common verbs, nouns, and adjectives, using the Collins COBUILD corpus. To explore the feasibility of identifying grammar patterns computationally, Mason (2004) conducted a limited experiment of automatic parsing based on COBUILD grammar patterns with reasonable success. More recently, Yen et al. (2015) introduced a method for inducing grammar patterns to use in an interactive writing environment aimed at assisting language learners in writing.

Identifying the intended word sense relevant to the context has long been an active topic of word sense disambiguation (WSD) research. In general, WSD systems typically use supervised learning approach with a sense inventory such as WordNet WSD systems based on dictionary-based sense inventory (e.g., WordNet) and a sense-annotated corpus (e.g., Semcor Miller et al. (1994)). In our work, we adopt BERT introduced by Devlin et al. (2018) to disambiguate words in user-submitted contents to provide correct word definition and appropriate quizzes.

Wikification of educational materials has been touted as a novel approach to facilitate reading and learning. In this work, we use the existing method proposed by Kolitsas et al. (2018), to identify potentially ambiguous mentions of key phrases in a document and link them to relevant Wikipedia articles.

Much of previous work shows that one of the most efficient way to learn a second language is through extensive reading, using engaging extracurricular articles, news or books (e.g., Coady (1997), Pigada and Schmitt (2006)). Inspired by their insights, we present *Linggle Booster*, an interactive environment which provides helpful information related to input article, to help learners acquiring deeper knowledge while reading.

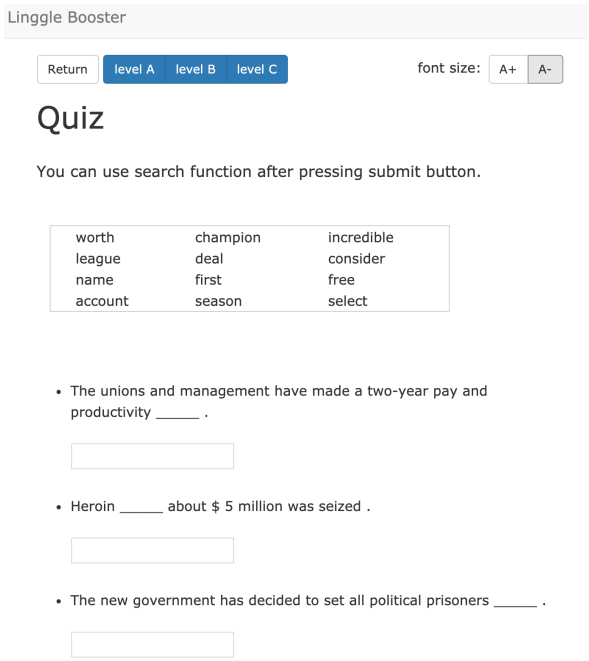

Figure 2: An example of auto-generated test items for the user-selected web page [6]

3 Method

Our system is composed of four main components: (i) extracting grammar patterns, collocations, and example sentences; (ii) generating words or phrases linked to Wikipedia information; (iii) training language representations for WSD; (iv) generating vocabulary quizzes.

3.1 Extracting grammar patterns, collocations and example sentences

We extract grammar patterns, collocations and example sentences from Corpus of Contemporary American English (COCA) and data from Cambridge online dictionary (CAM)[7]. We first parse sentences in the two datasets using spaCy toolkit. From the result of dependency parsing, we extract grammar patterns of content words (i.e., verbs, nouns and adjectives) based on handcrafted templates. For each target content word, we only keep words which are its children and labeled by specific dependency relations. For example, the extracted grammar pattern of the verb *chew* in the sentence *"She is chewing at her nail"* is **V at n**.

Then, to cope with noise caused by parser errors, we discard extracted grammar patterns not listed in extended Collins COBUILD Grammar Patterns (COBUILD) (Hunston et al., 1996). We also extract collocations to accompany grammar patterns. For example, the grammar pattern of **role** is extended from **N in n** to **v N in n** by adding the verb collocation **play** (school **play** an important **role in** society).

After that, for each target word, we calculate patterns and collocations and filter out those less frequent than the mean by 1.0 standard deviation.

Finally, we select examples of each pattern from COCA and Cambridge online dictionary (with Chinese translations) using the GDEX method (Kilgarriff et al., 2008).

3.2 Link Words or Phrases to Wikipedia Information

To link words and phrases in user-submitted contents to correct Wikipedia entries, we perform Mention Detection and Entity Disambiguation on user-submitted contents, using the End-to-End Neural Entity Linking method (Kolitsas et al., 2018). We generate possible spans from unigram to trigram, and each span selects some Wikipedia entry candidates with an empirical probabilistic entity map (Ganea and Hofmann, 2017) from Wikipedia hyperlinks, Crosswikis and YAGO dictionaries. Each mention candidate produces a local contextual similarity scores. Accordingly, we provide correct Wikipedia knowledge for words to assist ESL better understanding the contents and world knowledge.

3.3 Word Sense Disambiguation

We disambiguate polysemous words in user-submitted contents using a pre-trained language representation model, BERT (Devlin et al., 2018). We use word definitions in CAM as word sense labels. For a given word, CAM offers all possible word definitions, CERF levels and example sentences. We view example sentence as the feature of a word sense. Then, we use the last four hidden layers of BERT hidden state to compute the vector representations of each example sentence. Next, we use BERT again to compute word vector for words in user-submitted contents. Finally, we disambiguate the word sense by calculating the cosine similarity of the representations and each representation of word definition in CAM, and return the word definition of which examples contains the most similar representation. After word sense disambiguation, we provide appropriate word definitions and correspondent word level to learners.

[7] https://dictionary.cambridge.org/

3.4 Generating Quizzes

Fill in the blank questions (FBQ) are automatically generated after the reading session. We randomly select vocabulary from user-submitted content that matches the user-declared proficiency level. To form questions, we select representative examples containing the target word with the word sense in the user-submitted content from CAM. Then, the target word is replaced with a blank to help learners self-assess the acquisition of vocabulary. After users complete a test, *Linggle Booster* presents the scores and corrections to the users.

4 Run-Time Interactive Environment

Linggle Booster is implemented in Python based on Django Web framework. For faster retrieval, we save the added reference information (cf. Section 3.1) in JSON format using PostgreSQL and hash table. We choose to host *Linggle Booster* on Heroku, a cloud-platform-as-a-service site for uninterrupted service and scalability. The server of *Linggle Booster* with AJAX techniques receives users-submitted content (e.g., Web page URLs, URL of YouTube video with closed caption, or essay draft) from any popular browser (e.g., Chrome, Safari, or Firefox).

If users submit an URL, we use an existing tool [8] to parse the html of give URL and extract article content. We detect possible Named Entity and link to correct Wikipedia entries using the method in Section 3.2. At the same time, we parse the article content using spaCy toolkit and compute representations using BERT. After disambiguating the word sense of each word using the method in Section 3.3, we can access the Word Level of the word sense in CAM.

Then, we reformat the article content in a reader-friendly layout presented in the left column of *Linggle Booster*. Words with the level matched to the user-selected level are underlined, and keywords and phrases linked to Wikipedia information are presented in blue. For each word in the content, we retrieve five pieces of information, the definition of the word sense in Chinese, the grammar patterns of the word, the frequency, collocations, and example for each grammar pattern, and commonly used phrases if they exist. If a key word lacks grammar patterns, we present the vocabulary definitions and synonyms based on WordNet (Miller, 1998). We process rare words not in vocabulary by decomposing them into affixes and stems, and retrieving linguistic information accordingly. In the self-assessment session, users can access a vocabulary quiz with one click, along with scores and corrections after answering the quizzes.

[8] https://github.com/buriy/python-readability

	WSD	Pattern	Col.	Example
Level A	70 %	92 %	82 %	85 %
Level B	90 %	91 %	89 %	89 %
Level C	75 %	92 %	87 %	91 %

Table 1: Accuracy of human evaluations of *Linggle Booster* for CNN news article.

5 Evaluation

In this section, we report the results of preliminary evaluations on automated extraction of grammar patterns, collocations, and examples. The quality evaluation of Wikification and word sense disambiguation is also included in this section.

Vocabulary knowledge extraction is a king of information extraction (IE) tasks, which are traditionally evaluated based on the quality of accuracy or appropriateness of generated result. We selected a CNN news article[9] to assess *Linggle Booster's* performance. We examined the Chinese word sense, grammar patterns, collocations, and examples for first 20 unique vocabulary in each word level. We checked if *Linggle Booster* returns the correct word sense used in the article. For each vocabulary, we check if grammar patterns more frequent than 5% frequency are valid. We also examined the accuracy of collocations. Finally, we evaluated whether the example for each grammar pattern is actually a good representation of it's usage.

Across all three Word Levels, the results (shown in Table 1) indicates *Linggle Booster* provides good definition at least 70% of time and grammar patterns, collocations and examples are all close to 90% correct.

To evaluate the quality of linking words to Wikipedia information, we conduct experiments on public Entity Linking dataset AIDA Hoffart et al. (2011) using the Gerbil platform Usbeck et al. (2015). Micro and macro F1 scores are 0.83

[9] https://edition.cnn.com/2019/04/10/australia/australia-china-election-intl/index.html

Dataset	F1 Score
senseval2	66.8
senseval3	66.1
SemEval 2007	55.1
SemEval 2013	62.8
SemEval 2015	67.8

Table 2: WSD evaluation

and 0.84 respectively.

We also performed an experiment on word sense disambiguation based on method proposed in ELMO using SemCor 3.0 Miller et al. (1994) and OMSTI Taghipour and Ng (2015) as training data. After training, we take the average representations for each Wordnet sense. To test our WSD method using (Raganato et al., 2017), we use BERT again to compute word vectors for every target word and take the most similar sense from the training set. If lemma is not in training set, we use the first sense from Wordnet as our word sense. The result of this test is shown in Table 2.

6 Conclusion and Future Work

We have presented *Linggle Booster*, an interactive and customizable environment for reading to improve language skills, where ESL learners can submit self-selected engaging content and set an appropriate proficiency level of vocabulary. With *Linggle Booster*, second language learners should have a much better chance of acquiring deeper vocabulary knowledge (e.g. grammar patterns, collocations, examples and encyclopedic information). In addition, users can self-assess how well they have acquired the vocabulary. Our methodology supports adaptive, self-paced vocabulary learning, resulting in an effective and engaging system that combines the advantages of freedom in the selection of learning content and rich and rewarding learning experiences enhanced by technology.

Many avenues exist for improving *Linggle Booster*. We could improve the ability to download an URL and parse the content. Our system cannot extract part of or all of the content for some web pages due to the limit of the adopted tool Readability. One solution is to use different parsing tools (e.g., Mercury[10]). *L*inggle Booster attempts to disambiguate words in user-submitted content and provide users with correspondent Chinese definitions. We will take one step further to offer users with grammar patterns and collocations specific to a word sense. Besides, we could improve our results by expanding training corpus for WSD. Additionally, an interesting direction to explore is ranking grammar patterns to match the proficiency level of readers. Yet another direction of research would be using the same design to assist writing in English. Instead of providing supports for reading a user-selected article, the system could take the user's own writing as input and use grammar patterns and collocations to improve writing quality and correct grammatical errors.

[10] https://mercury.postlight.com/web-parser/

References

James Coady. 1997. 1 1 l2 vocabulary acquisition through extensive reading. *Second language vocabulary acquisition: A rationale for pedagogy*, page 225.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

G Francis, S Hunston, and E Manning. 1998. Cobuild grammar patterns 2: Nouns.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920*.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Susan Hunston, Gill Francis, and E Manning. 1996. Collins cobuild grammar patterns 1: verbs.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proc. Euralex*.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*.

Judit Kormos and Kata Csizer. 2014. The interaction of motivation, self-regulatory strategies, and autonomous learning behavior in different learner groups. *Tesol Quarterly*, 48(2):275–299.

Oliver Mason. 2004. Automatic processing of local grammar patterns. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, University of Birmingham*, pages 166–171. Citeseer.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, pages 240–243. Association for Computational Linguistics.

Maria Pigada and Norbert Schmitt. 2006. Vocabulary acquisition from extensive reading: A case study. *Reading in a foreign language*, 18(1):1–28.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Jack C Richards and Willy A Renandya. 2002. *Methodology in language teaching: An anthology of current practice*. Cambridge university press.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 338–344.

Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, et al. 2015. Gerbil: general entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1133–1143. International World Wide Web Conferences Steering Committee.

Tzu-Hsi Yen, Jian-Cheng Wu, Jim Chang, Joanne Boisson, and Jason Chang. 2015. Writeahead: Mining grammar patterns in corpora for assisted writing. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 139–144.