Interactive Second Language Learning from News Websites

Tao Chen[1] **Naijia Zheng**[1] **Yue Zhao**[1]
Muthu Kumar Chandrasekaran[1] **Min-Yen Kan**[1,2*]
[1]School of Computing, National University of Singapore
[2]NUS Interactive and Digital Media Institute, Singapore
{taochen,muthu.chandra,kanmy}@comp.nus.edu.sg
{znj472982642,immortalzhaoyue}@gmail.com

Abstract

We propose WordNews, a web browser extension that allows readers to learn a second language vocabulary while reading news online. Injected tooltips allow readers to look up selected vocabulary and take simple interactive tests.

We discover that two key system components needed improvement, both which stem from the need to model context. These two issues are real-world word sense disambiguation (WSD) to aid translation quality and constructing interactive tests. For the first, we start with Microsoft's Bing translation API but employ additional dictionary-based heuristics that significantly improve translation in both coverage and accuracy. For the second, we propose techniques for generating appropriate distractors for multiple-choice word mastery tests. Our preliminary user survey confirms the need and viability of such a language learning platform.

1 Introduction

Learning a new language from language learning websites is time consuming. Research shows that regular practice, guessing, memorization (Rubin, 1975) as well as immersion into real scenarios (Naiman, 1978) hastens the language learning process. To make second language learning attractive and efficient, we interleave language learning with the daily activity of online news reading.

Most existing language learning software are either instruction-driven or user-driven. Duolingo[1] is a popular instruction-driven system that teaches through structured lessons. Instruction driven systems demand dedicated learner time on a daily basis and are limited by learning materials as lesson curation is often labor-intensive.

In contrast, many people informally use Google Translate[2] or Amazon Kindle's Vocabulary Builder[3] to learn vocabulary, making these prominent examples of user-driven systems. These systems, however, lack the rigor of a learning platform as they omit tests to allow learners to demonstrate mastery. In our work, we merge learning and assessment within the single activity of news reading. Our system also adapts to the learner's skill during assessment.

We propose a system to enable online news readers to efficiently learn a new language's vocabulary. Our prototype targets Chinese language learning while reading English language news. Learners are provided translations of open-domain words for learning from an English news page. In the same environment – for words that the system deems mastered by the learner – learners are assessed by replacing the original English text in the article with their Chinese translations and asked to translate them back given a choice of possible translations. The system, **WordNews**, deployed as a Chrome web browser extension, is triggered when readers visit a preconfigured list of news websites (*e.g.*, CNN, BBC).

A key design property of our WordNews web browser extension is that it is only active on certain news websites. This is important as news articles typically are classified with respect to a news

* This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

[1] https://www.duolingo.com/
[2] https://translate.google.com/
[3] http://www.amazon.com/gp/help/customer/display.html?nodeId=201733850

category, such as *finance*, *world news*, and *sports*. If we know which category of news the learner is viewing, we can leverage this contextual knowledge to improve the learning experience.

In the development of the system, we discovered two key components that can be affected by this context modeling. We report on these developments here. In specific, we propose improved algorithms for two components: **(i)** for translating English words to Chinese from news articles, **(ii)** for generating distractors for learner assessment.

2 The WordNews Chrome Extension

Our method to directly enhance the web browser is inspired by earlier work in the computer-aided language learning community that also uses the web browser as the delivery vehicle for language learning. WERTi (Metcalf and Meurers, 2006; Meurers et al., 2010) was a monolingual, user-driven system that modified web pages in the target language to highlight or remove certain words from specific syntactic patterns to teach difficult-to-learn English grammar.

Our focus is to help build Chinese vocabulary for second language learners fluent in English. We give a running scenario to illustrate the use of WordNews. When a learner browses to an English webpage on a news website, our extension either selectively replaces certain original English words with their Chinese translation or underlines the English words, based on user configuration (Figure 1, middle). While the meaning of the Chinese word is often apparent in context, the learner can choose to learn more about the replaced/underlined word, by mousing over the word to reveal a definition tooltip (Figure 1, left) to aid mastery of the Chinese word. Once the learner has encountered the target word a few times, WordNews assesses learner's mastery by generating a multiple choice translation test on the target word (Figure 1, right). Our learning platform thus can be viewed as three logical use cases: *translating*, *learning* and *testing*.

Translating. We pass the main content of the webpage from the extension client to our server for candidate selection and translation. As certain words are polysemous, the server must select the most appropriate translation among all possible meanings. Our initial selection method replaces any instance of words stored in our dictionary. For translation, we check the word's stored meanings against the machine translation of each sentence obtained from the Microsoft Bing Translation API[4] (hereafter, "Bing"). Matches are deemed as correct translations and are pushed back to the Chrome client for rendering.

Learning. Hovering the mouse over the replacement Chinese word causes a tooltip to appear, which gives the translation, pronunciation, and simplified written form, and a `More` link that loads additional contextual example sentences (that were previously translated by the backend) for the learner to study. The `More` link must be clicked for activation, as we find this two-click architecture helps to minimize latency and the loading of unnecessary data. The server keeps record of the learning tooltip activations, logging the enclosing webpage URL, the target word and the user identity.

Testing. After the learner encounters the same word a pre-defined number $t = 3$ times, WordNews generates a multiple choice question (MCQ) test to assess mastery. When the learner hovers over the replaced word, the test is shown for the learner to select the correct answer. When an option is clicked, the server logs the selection and updates the user's test history, and the client reveals the correct answer.

2.1 News Categories

As our learning platform is active only on certain news websites, we can model the news category (for individual words and whole webpages) as additional evidence to help with tasks. Of particular importance to WordNews is the association of words to a news category, which is used downstream in both word sense disambiguation (Section 3) and the generation of distractors in the interactive tests (Section 4). Here, our goal is to automatically find highly relevant words to a particular news category – *e.g.*, "what are typical *finance* words?"

We first obtain a large sample of categorized English news webpages, by creating custom crawlers for specific news websites (*e.g.* CNN). We use a seed list of words that are matched

[4] `https://www.bing.com/translator/`

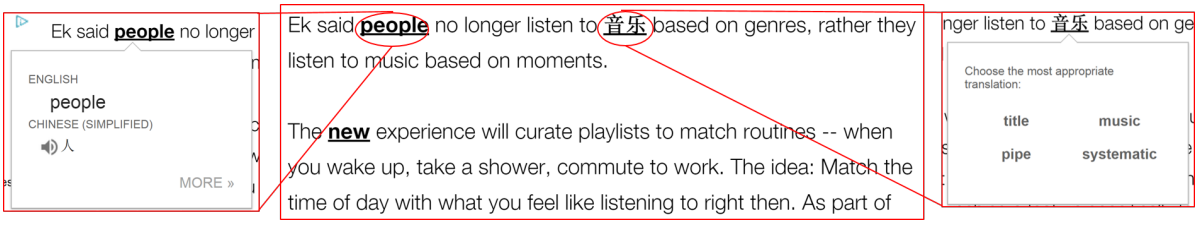

Figure 1: Merged screenshots of our Chrome extension on the CNN English article *Spotify wants to be the soundtrack of your life*. **Underlined components are clickable to yield tooltips of two different forms: (left) a definition for learning, (right) a multiple-choice interactive test.**

Table 1: News category alignment between English and Chinese.

English Category	Chinese Category	Example Words
1. Entertainment	Entertainment	"superstar", "明星"
2. World	Military, International, Social	"attacks", "军事"
3. Finance	Finance	"investment", "财富"
4. Sports	Sports	"score", "比赛"
5. Fashion	Beauty & Health	"jewelry", "时髦"
6. Technology	Technology	"cyber", "互联网"
7. Travel		"natural"

against a target webpage's URL. If any match, the webpage is deemed to be of that category. For example, a webpage that has the seed word "football" in its URL is deemed of category "Sports". Since the news category is also required for Chinese words for word sense disambiguation, we must perform a similar procedure to crawl Chinese news (*e.g.*, BaiduNews[5]) However, Chinese news sites follow a different categorization scheme, so we first manually align the categories based on observation (see Table 1), creating seven bilingual categories: namely, "World", "Technology", "Sports", "Entertainment", "Finance", "Fashion" and "Travel".

We tokenize and part-of-speech tag the main body text of the categorized articles, discarding punctuation and stopwords. For Chinese, we segment words using the Stanford Chinese word segmenter (Chang et al., 2008). The remaining words are classified to a news category based on document frequency. A word w is classified to a category c if it appears more often (a tunable threshold δ[6]) than its average category document frequency. Note that a word can be categorized to multiple categories under this scheme.

3 Word Sense Disambiguation (WSD) Component

Our extension needs to show the most appropriate translation sense based on the context. Such a translation selection task – cross-lingual word sense disambiguation – is a common problem in machine translation. In this section, we describe how we improved WordNews' WSD capabilities through a series of six approaches.

The context evidence that we leverage for WSD comes in two forms: the news category of the target word and its enclosing sentence.

3.1 Bilingual Dictionary and Baseline

WordNews's server component includes a bilingual lexicon of English words with possible Chinese senses. The English words in our dictionary is based on the publicly-available College English Test (CET 4) list, which has a breadth of about 4,000 words. We augment the list to include the relative frequency among Chinese senses, with their part-of-speech, per English word.

Our baseline translation uses the most frequent sense: for an English word to be translated, it chooses the most frequent relative Chinese translation sense c from the possible set of senses C.

[5] `http://news.baidu.com`

[6] We empricially set δ to 10.

Table 2: Example translations from our approaches to WSD. Target words are italicized and correct translations are bolded.

English Sentence	Dictionary	Baseline	POS	Machine Translation		
				Substring	Relax	Align
(1) ... a very *close* friend of ...	verb: 关闭, 合, 关 ... adj: 密切, ... 亲密 ...	关闭	密切	**亲密**	**亲密**	**亲密**
(2) ... kids can't *stop* singing ...	verb: 停止, 站, 阻止, 停 ...	**停止**	阻止	**停止**	**停止**	**停止**
(3) ... about Elsa being happy and *free* ...	adj: 免费, 自由, 游离, 畅, 空闲的...	免费	免费	**自由**	**自由**	**自由**
(4) ... why Obama's *trip* to my homeland is meaningful ...	noun: 旅, 旅程 ... 旅游 ...	旅	旅	旅	**旅行**	**旅行**
(5) ... winning more points in the *match* ...	noun: 匹配, 比赛, 赛, 敌手, 对手, 火柴 ...	匹配	匹配	**比赛**	**比赛**	**比赛**
(6) ... *state* department spokeswoman Jen Psaki said that the allies ...	noun: 态, 国, 州, ... verb: 声明, 陈述, 述, 申明 ... 发言 ... adj: 国家的 ...	态	态	发言	发言人	**国家**

This method has complete coverage over the CET 4 list (as the word frequency rule always yields a prospective translation), but as it lacks any context model, it is the least accurate.

3.2 Approach 1: News Category

Topic information has been shown to be useful in WSD (Boyd-Graber et al., 2007). For example, consider the English word *interest*. In finance related articles, "interest" is more likely to carry the sense of "a share, right, or title in the ownership of property" ("利息" in Chinese), over other senses. Therefore, analysing the topic of the original article and selecting the translation with the same topic label might help disambiguate the word sense. For a target English word e, for each prospective Chinese sense $c \in C$, choose the first (in terms of relative frequency) sense that has the same news category as the containing webpage.

3.3 Approach 2: Part-of-Speech

Part-of-Speech (POS) tags are also useful for word sense disambiguation (Wilks and Stevenson, 1998) and machine translation (Toutanova et al., 2002; Ueffing and Ney, 2003). For example, the English word "book" can function as a verb or a noun, which gives rise to two different dominant senses: "reserve" ("预定" in Chinese) and "printed work" ("书"), respectively. As senses often correspond cross-lingually, knowledge of the English word's POS can assist disambiguation. We employ the Standford log-linear Part-of-Speech tagger (Toutanova et al., 2003) to obtain the POS tag for the English word, whereas the POS tag for target Chinese senses are provided in our dictionary. In cases where multiple candidate Chinese translations fit the same sense, we again break ties using relative frequency of the prospective candidates.

3.4 Approaches 3–5: Machine Translation

Neighbouring words provide the necessary context to perform WSD in many contexts. In our work, we consider the sentence in which the target word appears as our context. We then acquire its translation from Microsoft Bing Translator using its API. As we access the translation as a third party, the Chinese translation comes as-is, without the needed explicit word to locate the target English word to translate in the original input sentence. We need to perform alignment of the Chinese and English sentences in order to recover the target word's translation from the sentence translation.

Approach 3 – Substring Match. As potential Chinese translations are available in our dictionary, a straightforward use of substring matching recovers a Chinese translation; *i.e.*, check whether the candidate Bing translation is a substring of the Chinese translation. If more than one candidate matches, we use the longest string match heuristic and pick the one with the longest match as the final output. If none matches, the system does not output a translation for the word.

Approach 4 – Relaxed Match. The final rule in the substring match method unfortunately fires often, as the coverage of WordNews's lexicon is limited. As we wish to offer correct translations that are not limited by our lexicon, we relax our substring condition, allowing the Bing translation to be a superset of a candidate translation in our dictionary (see Example 4 in Table 2, where the Bing translation "旅行" is allowed to be relaxed to match the dictionary "旅"). To this end, we must know the extent of the words in the translation. We first segment the obtained Bing translation with the Stanford Chinese Word Segmenter, and then use string matching to find a Chinese translation c. If more than one candidate matches, we heuristically use the last matched candidate. This technique significantly augments the translation range of our extension beyond the reach of our lexicon.

Approach 5 – Word Alignment. The relaxed method runs into difficulties when the target English e's Chinese prospective translations which come from our lexicon generate several possible matches. Consider Example 6 in Table 2. The target English word "state" has corresponding Chinese entries "发言" and "国家的" in our dictionary. For this reason, both "国家" ("country", correct) and "发言人" ("spokeswoman", incorrect) are relaxed matches. As relaxed approach always picks up the last candidate, "发 言人" is the final output, which is incorrect.

To address this, we use the Bing Word Alignment API[7] to provide a possibly different prospective Chinese sense c. In this example, "state" matches "国家" ("country", correct) from word alignment, and the final algorithm chooses "国家" as the output.

[7] https://msdn.microsoft.com/en-us/library/dn198370.aspx

Table 3: WSD performance over our test set.

	Coverage	**Accuracy**
Baseline	**100%**	57.3%
News Category	2.0%	7.1%
POS	94.5%	55.2%
Bing – Substring	78.5%	79.8%
Bing – Relaxed	75.7%	80.9%
Bing – Align	76.9%	**97.4**%

3.5 Evaluation

To evaluate the effectiveness of our proposed methods, we randomly sampled 707 words and their sentences from recent CNN[8] news articles, manually annotating the ground truth translation for each target English word. We report both the **coverage** (*i.e.*, the ability of the system to return a translation) and **accuracy** (*i.e.*, whether the translation is contextually accurate).

Table 3 shows the experimental results for the six approaches. As expected, frequency-based baseline achieves 100% coverage, but a low accuracy (57.3%); POS also performs similarly . The category-based approach performs the worst, due to low coverage. This is because news category only provides a high-level context and many of the Chinese word senses do not have a strong topic tendency.

Of most promise is our use of web based translation related APIs. The three Bing methods iteratively improve the accuracy and have reasonable coverage. Among all the methods, the additional step of word alignment is the best in terms of accuracy (97.4%), significantly bettering the others. This validates previous work that sentence-level context is helpful in WSD.

4 Distractor Generation Component

Assesing mastery over vocabulary is the other key functionality of our prototype learning platform. The generation of the multiple choice selection test requires the selection of alternative choices aside from the correct answer of the target word. In this section, we investigate a way to automatically generate such choices (called *distractors* in the literature) in English, given a target word.

[8] http://edition.cnn.com/

4.1 Related Work

Much research exists on distractor generation for Multiple Choice Questions (MCQ). However, distractor generation for factoid questions are different from those for vocabulary assessment. Semantic and syntactic properties of the target word need to be considered while generating their distractors. For instance, distractor generation for English prepositions (Lee and Seneff, 2007) used a corpus collected from non-native English speakers that captured their idiosyncratic incorrect usage. Recently, (Pho et al., 2014) studied homogeneity characteristics among distractors from multiple MCQ corpora, generating new distractors over a test corpus. They validated their hypothesis of syntactic and semantic homogeneity among distractors. (Susanti et al., 2015) generate distractors using WordNet and word sense disambiguation given a target word. Unlike our system that generates distractors for cloze questions for target words from a single news article, they retrieve separate passages from the web for each target word they choose to test on. While their approach serves in testing mastery, it does not provide the capability for learning new vocabulary in context. Finally, Lärka (Volodina et al., 2014) – a Swedish language learning system – generates vocabulary assessment exercises using a corpus. They also have different modes of exercise generation to allow learning and testing via the same interface.

4.2 Approach

WordNews postulates "a set of suitable distractors" as: 1) having the same form as the target word, 2) fitting the sentence's context, and 3) having proper difficulty level according to user's level of mastery. As input to the distractor generation algorithm, we provide the target word, its part-of-speech (obtained by tagging the input sentence first) and the enclosing webpage's news category. We restrict the algorithm to produce distractors matching the input POS, and which match the news category of the page.

We can design the test to be more difficult by choosing distractors that are more similar to the target word. By varying the semantic distance, we can generate tests at varying difficulty levels. We quantify similarity by using the Lin distance (Lin, 1998) between two input candidate concepts in WordNet (Miller, 1995):

$$sim(c1, c2) = \frac{2 * logP(lso(c1, c2))}{logP(c1) + logP(c2)} \quad (1)$$

where $P(c)$ denotes the probability of encountering concept c, and $lso(c1, c2)$ denotes the lowest common subsumer synset, which is the lowest node in the WordNet hierarchy that is a hypernym of both $c1$ and $c2$. This returns a score from 0 (completely dissimilar) to 1 (semantically equivalent).

If we use a target word e as the starting point, we can use WordNet to retrieve related words using WordNet relations (hypernyms/hyponyms, synonyms/antonyms) and determine their similarity using Lin distance.

We empirically set 0.1 as the similarity threshold – words that are deemed more similar than 0.1 are returned as possible distractors for our algorithm. We note that Lin distance often returns a score of 0 for many pairs and the threshold of 0.1 allows us to have a large set of distractors to choose from, while remaining fairly efficient in run-time distractor generation.

We discretize a learner's knowledge of the word based on their prior exposure to it. We then adopt a strategy to generate distractors for the input word based learners' knowledge level:

Easy: The learner has been exposed to the word at least $t = 3$ times. Two distractors are randomly selected from words that share the same news category as the target word e. The third distractor is generated using our algorithm.

Hard: The learner has passed the Easy level test $x = 6$ times. All three distractors are generated from the same news category, using our algorithm.

4.3 Evaluation

The WordGap system (Knoop and Wilske, 2013) represents the most related prior work on automated distractor generation, and forms our baseline. WordGap adopts a knowledge-based approach: selecting the synonyms of synonyms (also computed by WordNet) as distractors. They first select the most frequently used word, $w1$, from

the target word's synonym set, and then select the synonyms of $w1$, called $s1$. Finally, WordGap selects the three most frequently-used words from $s1$ as distractors.

We conducted a human subject evaluation of distractor generation to assess its fitness for use. The subjects were asked to rank the feasibility of a distractor (inclusive of the actual answer) from a given sentential context. The contexts were sentences retrieved from actual news webpages, identical to WordNews's use case.

We randomly selected 50 sentences from recent news articles, choosing a noun or adjective from the sentence as the target word. We show the original sentence (leaving the target word as blank) as the context, and display distractors as choices (see Figure 2). Subjects were required to read the sentence and rank the distractors by plausibility: 1 (the original answer), 2 (most plausible alternative) to 7 (least plausible alternative). We recruited 15 subjects from within our institution for the survey. All of them are fluent English speakers, and half are native speakers.

We evaluated two scenarios, for two different purposes. In both evaluations, we generate three distractors using each of the two systems, and add the original target word for validation (7 options in total, conforming to our ranking options of 1–7).

Since we have news category information, we wanted to check whether that information alone could improve distractor generation. Evaluation 1 tests the WordGap baseline system versus a **Random News** system that uses random word selection. It just uses the constraint that chosen distractors must conform to the news category (be classified to the news category of the target word).

In our Evaluation 2, we tested our **Hard** setup where our algorithm is used to generate all distractors against WordGap. This evaluation aims to assess the efficacy of our algorithm over the baseline.

4.3.1 Results and Analysis

Each question was answered by five different users. We compute the average ranking for each choice. A lower rating means a more plausible (harder) distractor. The rating for all the target words is low (1.1 on average) validating their truth and implying that the subjects answered the survey seriously, assuring the validity of the evaluation.

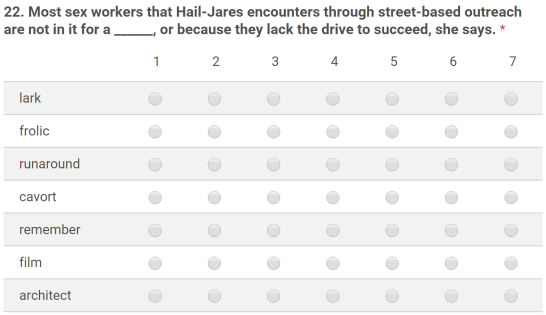

Figure 2: Sample distractor ranking question.

For each question, we deem an algorithm to be the winner if its three distractors as a whole (the sum of three average ratings) are assessed to be more plausible than the distractors by its competitor. We calculate the number of wins for each algorithm over the 50 questions in each evaluation.

Table 4: WordGap vs. Random News. Lower scores are better.

	# of wins	Avg. score
WordGap	27	3.84
Random News	23	4.10

Table 5: WordGap vs. WordNews Hard. Lower scores are better.

	# of wins	Avg. score
WordGap	21	4.16
WordNews Hard	29	3.49

We display the results of both evaluations in Table 4 and Table 5. We see that the WordGap baseline outperforms the random selection, constrained solely by news category, by 4 wins and a 0.26 lower average score. This shows that word news category alone is insufficient for generating good distractors. When a target word does not have a strong category tendency, *e.g.*, "venue" and "week", the random news method cannot select highly plausible distractors.

In the second table, our distractor algorithm significantly betters the baseline in both number of wins (8 more) and average score (0.67 lower).

Table 6: Distractors generated by WordGap and WordNews Hard for example question in Figure 2. The identified news category for the enclosing webpage was Entertainment.

System	Distractor	Lin Dist.	Avg. Rate
Target Word	lark		1.33
WordGap	frolic		3.33
	runaround		5.67
	cavort		4.17
WordNews Hard	art	0.154	1.67
	film	0.147	3.33
	actress	0.217	4.83

This further confirms that context and semantic information are complementary for distractor generation. As we mentioned before, a good distractor should fit the reading context and have a certain level of difficulty. Finally, in Table 6 we show the distractors generated for the target word "lark" in the example survey question (Figure 2).

5 Platform Viability and Usability Survey

We have thus far described and evaluated two critical components that can benefit from capturing the learner's news article context. In the larger context, we also need to check the viability of second language learning intertwined with news reading. In a requirements survey prior to the prototype development, two-thirds of the respondents indicated that although they have interest in learning a second language, they only have only used language learning software infrequently (less than once per week) yet frequently read news, giving us motivation for our development.

Post-prototype, we conducted a summative survey to assess whether our prototype product satisfied the target niche, in terms of interest, usability and possible interference with normal reading activities. We gathered 16 respondents, 15 of which were between the ages of 18–24. 11 (the majority) also claimed native Chinese language proficiency.

The respondents felt that the extension platform was a viable language learning platform (3.4 of 5; on a scale of 1 "disagreement" to 5 "agreement") and that they would like to try it when available for their language pair (3 of 5).

In our original prototype, we replaced the original English word with the Chinese translation. While most felt that replacing the original English with the Chinese translation would not hamper their reading, they still felt a bit uncomfortable (3.7 of 5). This finding prompted us to change the default learning tooltip behavior to underlining to hint at the tooltip presence.

6 Conclusion

We described WordNews, a client extension and server backend that transforms the web browser into a second language learning platform. Leveraging web-based machine translation APIs and a static dictionary, it offers a viable user-driven language learning experience by pairing an improved, context-sensitive tooltip definition service with the generation of context-sensitive multiple choice questions.

WordNews is potentially not confined to use in news websites; one respondent noted that they would like to use it on arbitrary websites, but currently we feel usable word sense disambiguation is difficult enough even in the restricted news domain. We also note that respondents are more willing to use a mobile client for news reading, such that our future development work may be geared towards an independent mobile application, rather than a browser extension. We also plan to conduct a longitudinal study with a cohort of second language learners to better evaluate WordNews' real-world effectiveness.

References

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A Topic Model for Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'07, pages 1024–1033.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT'08, pages 224–232.

Susanne Knoop and Sabrina Wilske. 2013. WordGap-Automatic Generation of Gap-filling Vocabulary Exercises for Mobile Learning. In *Proceedings of Second Workshop NLP Computer-Assisted Language Learning*, pages 39–47.

John Lee and Stephanie Seneff. 2007. Automatic generation of cloze items for prepositions. In *INTERSPEECH*, pages 2173–2176.

Dekang Lin. 1998. An information-theoretic definition of similarity.

Vanessa Metcalf and Detmar Meurers. 2006. Generating web-based english preposition exercises from real-world texts. Presentation at EUROCALL, September. http://purl.org/dm/handouts/eurocall06-metcalf-meurers.pdf.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Neil Naiman. 1978. *The Good Language Learner*, volume 4. Multilingual Matters.

Van-Minh Pho, Thibault André, Anne-Laure Ligozat, B Grau, G Illouz, Thomas François, et al. 2014. Multiple choice question corpus analysis for distractor characterization. In *9th International Conference on Language Resources and Evaluation (LREC 2014)*.

Joan Rubin. 1975. What the "good language learner" can teach us. *TESOL quarterly*, pages 41–51.

Yuni Susanti, Ryu Iida, and Takenobu Tokunaga. 2015. Automatic generation of english vocabulary tests. In *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015)*, pages 77–78.

Kristina Toutanova, H Tolga Ilhan, and Christopher D Manning. 2002. Extensions to HMM-based Statistical Word Alignment Models. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, EMNLP EMNLP'02, pages 87–94.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 173–180.

Nicola Ueffing and Hermann Ney. 2003. Using POS Information for Statistical Machine Translation into Morphologically Rich Languages. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics*, EACL'03, pages 347–354.

Elena Volodina, Ildikó Pilán, Lars Borin, and Therese Lindström Tiedemann. 2014. A flexible language learning platform based on language resources and web services. *Proceedings of LREC 2014*.

Yorick Wilks and Mark Stevenson. 1998. The Grammar of Sense: Using Part-of-speech Tags As a First Step in Semantic Disambiguation. *Natural Language Engineering*, 4(2):135–143.