Using Word Semantics To Assist English as a Second Language Learners

Mahmoud Azab
University of Michigan
mazab@umich.edu

Chris Hokamp
Dublin City University
chokamp@computing.dcu.ie

Rada Mihalcea
University of Michigan
mihalcea@umich.edu

Abstract

We introduce an interactive interface that aims to help English as a Second Language (ESL) students overcome language related hindrances while reading a text. The interface allows the user to find supplementary information on selected difficult words. The interface is empowered by our lexical substitution engine that provides context-based synonyms for difficult words. We also provide a practical solution for a real-world usage scenario. We demonstrate using the lexical substitution engine – as a browser extension that can annotate and disambiguate difficult words on any webpage.

1 Introduction

According to the U.S. Department of Education, about 11% of all the students in public schools in the United States receive or have received English language learning services. The largest numbers of ESL students are in California (26% of all the students) and Texas (16%). About 70% of these students are Spanish speakers (Dep, 2004). Moreover, there is a large number of non-English speaking countries that have programs for learning English, with as many as 750 million English as a Foreign Language students in the world (Crystal, 1997).

The goal of a number of computer-based language learning tools developed to date is to provide assistance to those with limited language abilities, including students learning a second or a foreign language or people suffering from disabilities such as aphasia. These tools draw on research in education, which found that text adaptation can improve the reading comprehension skills for learners of English (Yano et al., 1994; Carlo et al., 2004). The language learning technology often consists of methods for text simplification and adaptation, which is performed either at syntactic (Carroll et al., 1999; Siddharthan et al., 2004) or lexical level (Carroll et al., 1998; Devlin et al., 2000; Canning and Tait, 1999; Burstein et al., 2007). Work has also been carried out on the prediction and simplification of difficult technical text (Elhadad, 2006a; Elhadad, 2006b) and on the use of syntactic constraints for translations in context (Grefenstette and Segond, 2003).

In this paper, we describe an interface developed with the goal of assisting ESL students in their English reading activities. The interface builds upon a lexical substitution system that we developed, which provides synonyms and definitions for target words in context. We first give a brief overview of the lexical substitution task, and then present our system SALSA (Sinha and Mihalcea, 2014) and (Sinha and Mihalcea, 2012). We then describe the functionality of the interface, and the interaction that a user can have with this interface.

2 Lexical Substitution

Lexical substitution, also known as contextual synonym expansion (McCarthy and Navigli, 2007), involves replacing a certain word in a given context with another, suitable word, such that the overall meaning of the word and the sentence are unchanged. As an example, see the four sentences in Table 1, drawn from the development data from the SEMEVAL-2007 lexical substitution task. In the first sentence, for instance, assuming we choose *bright* as

Proceedings of NAACL-HLT 2015, pages 116–120,
Denver, Colorado, May 31 – June 5, 2015.

the target word, a suitable substitute could be *brilliant*, which would both maintain the meaning of the target word and at the same time fit the context.

Sentence	Target	Synonym
The sun was **bright**.	bright	brilliant
He was **bright** and independent.	bright	intelligent
His feature **film** debut won awards.	film	movie
The market is **tight** right now.	tight	pressured

Table 1: Examples of synonym expansion in context

We perform contextual synonym expansion in two steps: *candidate synonym collection*, followed by *context-based synonym fitness scoring*.

Candidate synonym collection is the first step of our system, and refers to the sub task of collecting a set of potential synonym candidates for a given target word, starting with various resources. Note that this step does not disambiguate the meaning of the target word. Rather, all the possible synonyms are selected, and these synonyms can be further refined in the later step. For example, if we consider all the possible meanings of the word *bright*, it can be potentially replaced by *brilliant, smart, intelligent, vivid, luminous*. SALSA uses five lexical resources, as listed in Table 2, to ensure a good collection of candidate synonyms.

The second step is *context-based synonym fitness scoring*, which refers to picking the best candidates out of the several potential ones obtained as a result of the previous step. There are several ways in which fitness scoring can be performed, for example by accounting for the semantic similarity between the context and a candidate synonym, or for the substitutability of the synonym in the given context. We experimented with several unsupervised and supervised methods, and the method that was found to work best uses a set of features consisting of counts obtained from Google N-grams (Brants and Franz, 2006) for several N-grams centered around the candidate synonym when replaced in context.

The synonym selection process inside SALSA was evaluated under two different settings. The first evaluation setting consists of the lexical sample dataset made available during SEMEVAL 2007 (McCarthy and Navigli, 2007) - a set of 1,700 annotated examples for 170 open-class words. On this dataset, SALSA is able to find the synonym agreed upon by several human annotators as its best guess in 21.3% cases, and this synonym is in the top 10 candidates returned by our system in 64.7% cases. These results compare favorably with the best results reported during SEMEVAL 2007 task on Lexical Substitution.

The second evaluation setting is a dataset consisting of 550 open-class words in running text. On this set of words, SALSA finds the best manually assigned synonym in 29.9% of the cases, and this synonym is in our top ten candidates in 73.7% of the cases.

Overall, we believe SALSA is able to identify good candidate synonyms for a target word in context, and therefore can form the basis for an interface to assist English learners.

3 An Interface for English as a Second Language Learners

Our goal is to leverage lexical substitution techniques in an interface that can provide support to ESL and EFL students in their reading activities. It is often the case that students who are not proficient in English have difficulty with understanding certain words. This in turn has implications for their comprehension of the text, and consequently can negatively impact their learning and knowledge acquisition process. By having inline access to an explanation of the words they have difficulty with, we believe these students will have easier access to the knowledge in the texts that they read.

In order to support various devices and platforms, we implemented the prototype interface as a web application. Given a text, the interface allows readers to click on selected vocabulary words, and view supplementary information in a side panel. This supplementary information includes a list of in-context synonyms, as provided by our system. In addition, we also include example sentences obtained from WordNet, corresponding to the target word meaning dictated by the top synonym selected by SALSA.

The interface also includes the possibility for the user to provide feedback by upvoting or downvoting supplementary information. The goal of this component is to allow the user to indicate whether they found the information provided useful or not. In addition to providing direct feedback on the quality of the interface, this user input will also indirectly con-

Table 2: Subsets of the candidates provided by different lexical resources for the adjective *bright*

Resource	Candidates
Roget (RG)	ablaze aglow alight argent auroral beaming blazing brilliant
WordNet (WN)	burnished sunny shiny lustrous undimmed sunshiny brilliant
TransGraph (TG)	nimble ringing fine aglow keen glad light picturesque
Lin (LN)	red yellow orange pink blue brilliant green white dark
Encarta (EN)	clear optimistic smart vivid dazzling brainy lively

tribute to the construction of a "gold standard" that we can use to further improve the tool.

We evaluated an earlier static version of this interface with ESL students who read two articles from the BBC's English learning website. We manually selected difficult words from the text, and for these words provided a list of in-context synonyms and clear examples. After each reading, the students took a post-reading quiz to evaluate their reading comprehension. We then evaluated the extent to which we could predict a student's performance on the post-quiz using features of their interaction with the tool.

We also used this interface with English middle school students whose primary language is English. The students had to read short excerpts of a book that was a part of their curriculum. Students were allowed to click on only one highlighted word per excerpt. In this experiment, supplementary information was provided from WordNet. There was a post-reading quiz to evaluate the students understanding of the words. By training a regression model on the interaction features collected during the reading exercises, we were able to accurately predict students' performance on the post-quiz (Hokamp et al., 2014).

We have now enabled the SALSA interface to provide feedback on arbitrary English content from the web. By implementing the tool as a browser extension, we are able to show inline additional information about text on any web page, even when the content is dynamically generated.

The interface also collects both explicit and implicit feedback. The explicit feedback is collected via upvotes and downvotes on feedback items. The implicit feedback is based on the user interactions with the system while they are reading. Currently, we collect several kinds of interactions. These interactions include the clicked words, counts of user clicks on a given word, the difficulty of the word as measured by the inverse document frequency, and the number of syllables it contains. In the future, this data will help us to adapt the tool to individual users.

4 Demonstration

During the demonstration, we will present the use of the interface. We will allow participants to freely browse the web with our tool enabled, to view feedback on lexical items, and to provide their own feedback on the quality of the results. The system will automatically identify and highlight the difficult words during browsing, and users can then click these highlighted words to receive supplementary information, consisting of synonyms and definitions, which should assist them in reading and comprehending the content.

By hovering or clicking on an annotated word, users can access a small popup window that includes supplementary information. This supplementary information includes a list of in-context synonyms, as provided by our system, and a clear example of the word in-context. Figure 1 shows an example of the current extension interface when a user hovers over the word **film**.

Although the reading activity + quiz format described above is necessary for the empirical evaluation of our tool, it does not demonstrate a real-world usage scenario. Therefore, we designed a browser extension to show a realistic use case for the lexical substitution engine as the backend for a flexible graphical component that can add additional information to any content. We anticipate that the extension will prove useful to English language learners as they navigate the Web, especially when they encounter difficult English content.

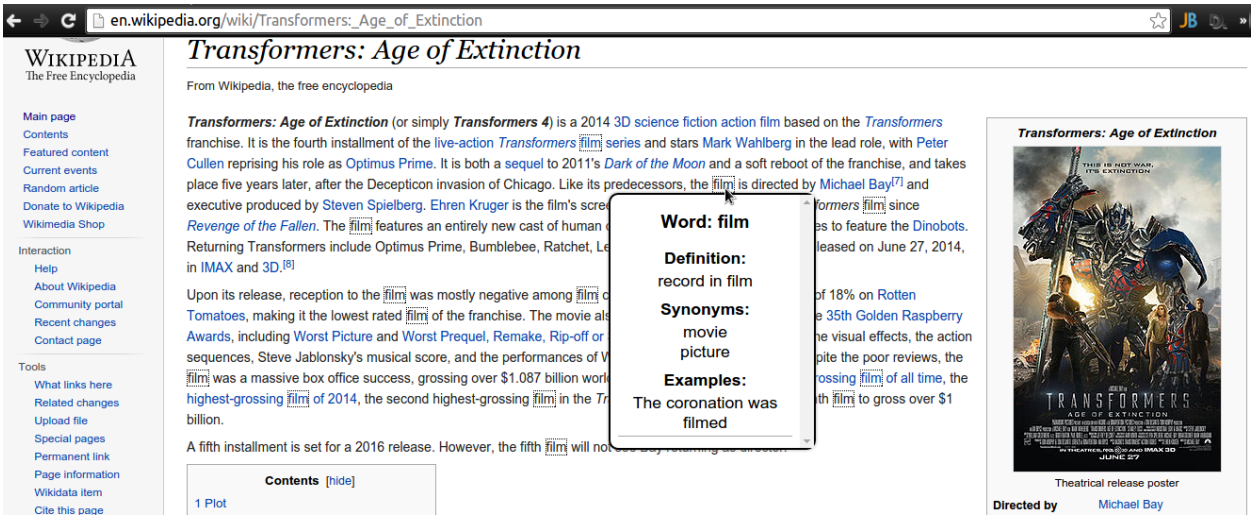

Figure 1: Example of supplementary information that the extension provides the user with when a user hovers over the word *film*.

Acknowledgments

This work was partially funded by the National Science Foundation (CAREER award #1361274) and by DARPA (DEFT grant #12475008). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, DARPA, or the other sources of support.

References

T. Brants and A. Franz. 2006. Web 1T 5-gram version 1. Linguistic Data Consortium.

J. Burstein, J. Shore, J. Sabatini, and Y. Lee. 2007. Developing a reading support tool for English language learners. In *Demo proceedings of the the annual conference of the North American chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, Rochester, NY.

Y. Canning and J. Tait. 1999. Syntactic simplification of newspaper text for aphasic readers. In *Proceedings of the ACM SIGIR'99 Workshop on Customised Information Delivery*, Berkeley, California.

M.S. Carlo, D. August, B. McLaughlin, C.E. Snow, C. Dressler, D. Lippman, T. Lively, and C. White. 2004. Closing the gap: Addressing the vocabulary needs of english language learners in bilingual and mainstream classrooms. *Reading Research Quarterly*, 39(2).

J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Madison, Wisconsin.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the Conference of the European Chapter of the ACL (EACL 1999)*, Bergen, Norway.

D. Crystal. 1997. *English as a global language*. Cambridge University Press.

2004. http://nces.ed.gov/fastfacts/display.asp?id=96.

S. Devlin, J. Tait, J. Carroll, G. Minnen, and D. Pearce. 2000. Making accessible international communication for people with language comprehension difficulties. In *Proceedings of the Conference of Computers Helping People with Special Needs*.

N. Elhadad. 2006a. Comprehending technical texts: Predicting and defining unfamiliar terms. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, Washington.

N. Elhadad. 2006b. *User-Sensitive Text Summarization: Application to the Medical Domain*. Ph.D. thesis, Columbia University.

G. Grefenstette and F. Segond, 2003. *Multilingual On-Line Natural Language Processing*, chapter 38.

C. Hokamp, R. Mihalcea, and P. Schuelke. 2014. Modeling language proficiency using implicit feedback. In *Proceedings of the Conference on Language Resources and Evaluations (LREC 2014)*, Reykjavik, Iceland, May.

D. McCarthy and R. Navigli. 2007. The semeval English lexical substitution task. In *Proceedings of the ACL Semeval workshop*.

A. Siddharthan, A. Nenkova, and K. McKeown. 2004. Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th international conference on Computational Linguistics*.

R. Sinha and R. Mihalcea. 2012. Explorations in lexical-sample and all-words lexical substitution. *Journal of Natural Language Engineering*.

Ravi Som Sinha and Rada Mihalcea. 2014. Explorations in lexical sample and all-words lexical substitution. *Natural Language Engineering*, 20(1):99–129.

Y. Yano, M. Long, and S. Ross. 1994. The effects of simplified and elaborated texts on foreign language tool's utility and effectiveness in terms of students' reading comprehension. *Language Learning*, 44.