Modeling Language Proficiency Using Implicit Feedback

Chris Hokamp[1], **Rada Mihalcea**[2], **Peter Schuelke**[3]
[1]Dublin City University, Dublin, Ireland
[2]University of Michigan, Ann Arbor (MI), United States
[3]University of Newcastle, Callghan (NSW) Australia
chokamp@computing.dcu.ie, mihalcea@umich.edu, peter.schuelke@uon.edu.au

Abstract

We describe the results of several experiments with interactive interfaces for native and L2 English students, designed to collect implicit feedback from students as they complete a reading activity. In this study, *implicit* means that all data is obtained without asking the user for feedback. To test the value of implicit feedback for assessing student proficiency, we collect features of user behavior and interaction, which are then used to train classification models. Based upon the feedback collected during these experiments, a student's performance on a quiz and proficiency relative to other students can be accurately predicted, which is a step on the path to our goal of providing automatic feedback and unintrusive evaluation in interactive learning environments.

Keywords: Educational Interfaces, Implicit Feedback, EFL

1. Introduction

The goal of this work is to evaluate the viability of assesing proficiency in a non-invasive and helpful manner. Despite the ongoing revolution in e-Learning and new media education, the test/quiz format still prevails as the standard means of assessing student performance. However, we believe that students naturally reveal their areas of strength and weakness as they work through educational materials. By looking at a word for an extended period, or requesting feedback on an item, a student shares information which can be used to tailor the feedback given by educational tools and inform instructors. This approach allows students to focus on absorbing and retaining new information, instead of rote memorization, or anticipation of test items.

Ideally, feedback should provide information that is targeted at the student's proficiency level relative to the task at hand, and that is designed to aid comprehension, with minimal distraction from the primary assignment.

In this paper, we describe the design of two interfaces that can collect implicit feedback from students while they work through a reading activity: one targeted at L2 adult English learners (the **EFL Assistant**), and one targeted at middle-school native speakers (the **English Assistant**). The interfaces allow readers to click on key vocabulary words in order to view supplementary information in a side panel. The supplementary information includes in-context synonyms and example sentences designed to make the meaning of the target word clear. Both interfaces are implemented as single-page applications designed for viewing in a modern web browser.

After finishing the reading activities, the students complete a short quiz, where they answer questions that test their understanding of the keywords from the reading. Although our long-term objective is to integrate learning and assesment materials to make assessment more transparent, we use the quiz format in these experiments to evaluate our methodologies in comparison to traditional methods, and train classifiers to predict students' performace on quizzes given features collected during the reading session.

A key aspect of the experimental design is that viewing feedback is optional. Readers must deliberately choose to view information about a certain word, and are free to complete exercises without viewing any feedback if they wish. Thus, interaction with a feedback item is likely to indicate that a student is unsure of its meaning, or at least curious to obtain more information about the word. Our results also show that interaction with feedback is correlated with better performance on the evaluation materials.

As students progress through the exercises, we collect information on the feedback items that are of interest to each user, aggregating a log containing features of user interaction for each session. Session logs are mapped into feature vectors, which can then be used to train classification algorithms to predict performance on a post-quiz, or, in the case of EFL students, their level as evaluated by their language program.

2. Related Work

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, students who lag behind in their reading skills, or people suffering from disabilities such as aphasia. These tools draw on research in education, which has indicated that text adaptation and augmentation can improve reading comprehension skills for learners of English (Yano et al., 1994; Carlo et al., 2004).

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).

One branch of work on mining implict feedback from computer-based language learning tools focuses on learning sequences of actions which can be used to identify clus-

Word
emerging
Original Sentence
An estimated 30,000 people are infected with Lyme disease every year, but other illnesses are emerging that can be deadly.
Synonyms
appearing, developing
Example Sentence
Some nice results emerged from the study.

Table 1: Example of the feedback shown for the word **emerging**

ters of learners. These interfaces generally require users to complete a problem in a sequence of steps.

Kardan and Conati (2011) create a framework for modeling and classifying users of a learning interface based upon logs of their actions. They collect features similar to ours, with the goal of classing users into Low-Learner (LL) or High-Learner (HL) groups. The goal of their work is to provide useful feedback to users engaging in sub-optimal problem solving strategies.

A large body of work in Educational Data Mining (EDM) utilizes data mining techniques and machine learning to extract knowledge from a variety of educational data. (Romero and Ventura, 2010) provides a survey of current work in EDM. (Merceron and Yacef, 2003) find correlations between classes of mistakes made while students practice creating proofs using propositional logic. The aim of this work is to aid the instructor in targeting areas of confusion, or areas requiring more detailed explanation.

3. Contextual Feedback

During evaluation sessions for both interfaces, instructors inform students that they will be quizzed on the meaning of vocabulary words after completing a reading activity. In the reading activities, some keywords are highlighted, and clicking a highlighted word opens an information column on the right side of the page which contains feedback on the word's meaning in the context.

We deliberately selected highly ambiguous words for the vocabulary quizzes, to ensure that students grasped the meaning of the word in the particular context. Table 1 gives some samples of the feedback available to the EFL group.

3.1. EFL Assistant

For the EFL assistant, the feedback for each word is manually created, with the goal of providing very close lexical substitutions and accurate examples that help students to understand the meaning from context.

We utilize two readings from the BBC's Learning English website,[1] and consulted with experienced EFL teachers to design fair and unambiguous vocabulary feedback and quiz questions.

[1]http://bbc.co.uk/worldservice/learningenglish/

3.2. English Assistant

For the English interface, we provide WordNet synonyms and usage examples for each of the feedback items via a server providing data from WordNet. The English interface was tested on Middle School students whose primary language is English, so we focused on providing feedback for difficult vocabulary words from "The Outsiders" by S.E. Hinton.[2] Students read short excerpts from the book, and have the opportunity to click on one word per excerpt. Some of the feedback words are then tested in the post-reading quiz, where students are presented with a series of sentences with missing words, and asked to drag the correct word into the corresponding blank.

These experiments test a real-world application scenario realistically, because manually creating feedback for every potential word of interest in a language is not feasible. Therefore, we provide feedback by leveraging existing resources via an interface that could be easily integrated with other data sources.

The detailed feedback obtained from a resource like WordNet requires students to perform disambiguation based on the word's context, so it is similar to a dictionary or 'automatic-lookup' functionality alongside the text. We tested this interface with Middle-school students, expecting that their English proficiency would be sufficient to make use of this information.

4. Features of User Interaction

To model the interaction between the student and the reading activity, as performed through our interface, we collect several features. All these features represent *implicit* feedback collected from the students, i.e., features that do not require any explicit actions other than the use of the interface.

First, the word difficulty metric ($meanDifficulty$) is a combination of word length (in syllables) and frequency, as measured by the Inverse Document Frequency (IDF). We take the mean of this feature over all feedback items that a user clicks. The difficulty function for feedback item i is:

$$Difficulty_i = idf_i * NumberOfSyllables_i$$

The intuition behind this metric is that words that contain fewer syllables should be easier to learn, and that rare words should be more difficult. We use CMUDict [3] to obtain the number of syllables in a word, and an IDF index built from the English version of Wikipedia to provide the IDF score of each token. This metric is simple, but effective, as illustrated in Figure 1.

We also record the time that each user spent on the reading activity ($totalTime$), the words they clicked ($numClicked$), and the percentage of the total available feedback that each user interacted with ($percentClicked$).

5. Results

We use the interaction features to predict the user's performance, using the Structured Perceptron implementation

[2]This book is part of the reading curriculum in the school where the study was conducted

[3]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

from the WEKA toolkit (Hall et al., 2009). The choice of the Structured Perceptron learning algorithm (Collins, 2002) is motivated both by the low feature count, and the need to learn a function to predict a numerical value (the users' quiz performance). We obtained a total of 44 test sessions for the English Assistant, and 42 sessions for the EFL assistant. All of the following results were obtained using the Structured Perceptron with 10-fold cross-validation over these instances.

5.1. Experiments With Middle School Students

For the middle school students, the optimal parameters for the Structured Perceptron achieve a Pearson Correlation of 0.61 with the students' performance on the quiz, indicating that our features have good predictive power. The feature $percentClicked$ has the most predictive power, followed by $averageDifficulty$.

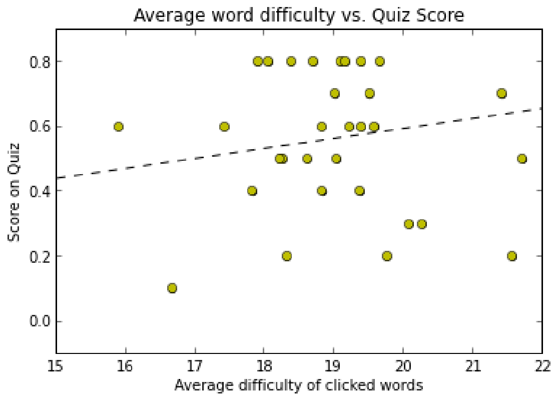

Figure 1: Middle-school students' quiz performance vs. average difficulty of feedback items viewed

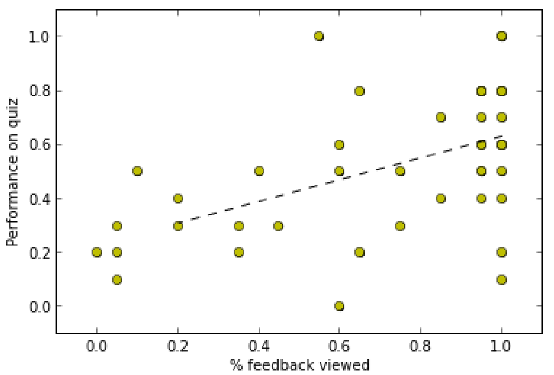

Figure 2: Middle-school students' quiz performance vs. % of feedback items viewed

5.2. Experiments with EFL Students

The Structured Perceptron did not perform well in the experiments with the EFL interface, achieving a Pearson Correlation of -0.22. We believe that the low accuracy of the classifiers on the L2 English learners' data stems from the difficulty of the feedback. Although most students did interact with feedback, they may have had difficulty understanding the contextual synonyms, especially given the range of proficiency levels in the test group. It is also possible that the quiz questions were too difficult for most students.

However, on this dataset, since we had additional information obtained directly from the instructor, consisting of the student level, we also run an SVM classifier with two classes: Beginner and Intermediate, based upon the course levels of the students in the study.

We assumed that the EFL students' level (1-6) would be strongly correlated with their performance on the quizzes. Surprisingly, the institutionally-designated class turned out to be only a weak predictor of students' performance, suggesting that the level system may need review. However, by mapping the levels 1-2 and 3-4 into two classes (beginner and intermediate), we achieve an accuracy of 62% using only the four features described above with the SMO implementation from the WEKA toolkit. Although this is a significant improvement over the 51% baseline, we believe that features collected from an interface with activities designed specifically for beginner and intermediate L2 English learners would be more accurate, especially if feedback is also tailored to each proficiency level, and we intend to address this in future work.

6. Conclusion

This paper describes several experiments designed to test the value of implicit feedback for predicting the performance of students on comprehension quizzes. Given a computer-based language learning interface, we address the question of whether we can use the interaction of a student with the interface in order to predict the student's language skills.

It is not surprising that students who sought feedback were more successful on the vocabulary quizzes. One important takeaway is that certain behaviors lead to success, namely the tendency to seek out feedback or help when it is available. A critical aspect of the experimental design is that students must volitionally click on items to receive feedback, and no extra information is provided by default.

This work also has interesting implications for the placement criteria used in most EFL programs. Our work indicates that the level of a student may be better predicted by their ability to understand and use context to bootstrap their learning, as opposed to their performance on a placement test. However, this is not certain, as our quizzes may have simply been too difficult for the L2 English learners.

We believe that the capacity to learn from and adapt to a user's proficiency level is desirable in all educational tools. Implicit feedback can preclude the need for the traditional activity/quiz format, instead focusing on the end goal: to help the individual to improve.

Acknowledgements

We are grateful to Anca Morcovescu for her help with running some of the evaluations. This material is based in part upon work supported by the National Science Foundation CAREER award #0747340 and IIS award #1018613.

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.

7. References

Burstein, J., Shore, J., Sabatini, J., and Lee, Y. (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.

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

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

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (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.

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

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. pages 1–8.

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

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

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

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

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

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

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Kardan, S. and Conati, C. (2011). A framework for capturing distinguishing user interaction behaviors in novel interfaces. In Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C., and Stamper, J. C., editors, *EDM*, pages 159–168. www.educationaldatamining.org.

Merceron, A. and Yacef, K. (2003). A web-based tutoring tool with mining facilities to improve learning and teaching. In *in Proceedings of the 11th International Conference on Artificial Intelligence in Education*, pages 201–208. IOS Press.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *Trans. Sys. Man Cyber Part C*, 40(6):601–618, November.

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

Yano, Y., Long, M., and Ross, S. (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.