EXPLORING THE INSTRUCTIONAL IMPLICATIONS OF AN AUTOMATED COURSE-SPECIFIC VOCABULARY IDENTIFICATION PROGRAM

Variawa, C., and Kinnear, P.

Department of Mechanical and Industrial Engineering, University of Toronto, Ontario Canada
variawa@mie.utoronto.ca; penny.kinnear@mie.utoronto.ca

Keywords: computational linguistics, vocabulary identification, vocabulary instruction, EAP, ESP.

Based on the principles of universal instructional design [1], a research study is being performed at the University of Toronto to explore whether an automated technique can be used to identify course-specific vocabulary. The motivation behind this is to create clearer course objectives and highlight the importance of developing a robust professional vocabulary, while promoting a more accessible engineering education. Additionally, the vocabulary used in engineering courses often contain vernacular that is neither technical nor course-specific, but is used to help contextualize the course content - the automated technique can also be used to identify such language.

The Term Frequency-Inverse Document Frequency (TF-IDF) algorithm is one approach that can help identify keywords that are specific to documents, when compared to relevant comparator sets. In this study, final exams from engineering courses are compared to a corpus of all electronically-available engineering final exams to develop wordlists for many courses. The algorithm works by multiplying the term frequency of each word on a specific exam being studied to the logarithm of how often that word appears in the group. The resulting data is tabulated to form a wordlist, with words characteristic of the input document having higher TF-IDF values. This approach has modified so that the wordlists are generated twice - once by comparing across exams from the same discipline, and another by calculating with all engineering disciplines - to increase the reliability of the wordlists. So far, the data generated shows that words that appear to be course-specific are assigned higher TF-IDF scores, and preliminary research is being conducted to understand the effectiveness of this sample data for a chemical engineering course. The theoretical effectiveness of distributing such wordlists as part of required course syllabi and course material is examined as part of this paper.

Vocabulary is critical to the academic and professional success of engineering students. All the students must learn to understand and use discipline specific professional language to practice engineering. For multilingual students this is a daunting task. While estimates vary, a student needs to know 98-99% of the lexical items to understand written discourse.

It is slightly lower for spoken discourse, around 80% for good comprehension (e.g. lectures, discussions). That translates to over 7000 word families [2]. Two questions

confront students entering an English-medium university and their instructors. The first question asks which vocabulary students need to learn and the second question queries what it means to “know” a word. These questions are fraught. Even among discipline experts, key vocabulary is contested. Attempts to deal with the first question include Coxhead’s Academic Word List (AWL) [3] and Xue and Nations’ University Word List [3]. However, these lists do not address the discipline specific vocabulary students will encounter in Engineering. English for Specific Purposes (ESP) teaching has attempted to target discipline specific technical vocabulary in its instructions. Mudraya [4], however, cites research, along with her own corpus-based research that learning the technical vocabulary is not the biggest challenge, rather the subtechnical, words that exist between the technical and the general, everyday vocabulary. Deciding where to focus student efforts is not a simple decision, however we can explore the use of automated course-specific vocabulary identification techniques to address this problem.

The problem we face currently is how to more effectively ensure that our multilingual students have access to the lexical resources they need to successfully develop and use professional engineering language. Evidence indicates that this incidental learning enhances knowledge of vocabulary the student has already seen more than the acquisition of new vocabulary [2]. The “word list” approach is equally inefficient as it relies on memorization of single forms and meanings. Students new to a discipline may well benefit from knowing which words carry discipline specific meanings, even if they may not yet understand the conceptual meanings and associations of the words, thus presenting students with a well-chosen word list with definitions at this point could be effective to introduce words. At other times a focus on the collocations and constraints may be most useful. Instruction that concentrates on the polysemous nature of the subtechnical vocabulary that students encounter in engineering communication documents and journal articles also has a place. Having a well-defined word list derived in a principled way from the discipline corpus provides a solid foundation from which to develop various instructional and self-study strategies to support student lexicogrammatical development in their professional language.

The development of a course-specific wordlist is a starting point for further research in the area of instructional support for professional language development in engineering education.

References

- [1] F. Bowe, *Universal Design in Education: Teaching Nontraditional Students*. Westpoint, CT: Bergin & Garvey, 2000.
- [2] N. Schmitt, "Instructed second language vocabulary learning," *Language Teaching Research*, vol. 12, pp. 329-363, 2008.
- [3] N. Hancioglu, S. Neufeld, and J. Eldridge, "Through the looking glass and into the land of lexico-grammar," *English for Specific Purposes*, vol. 27, pp. 459-479, 2008.
- [4] O. Mudraya, "Engineering English: A lexical frequency instructional model," *English for Specific Purposes*, vol. 25, pp. 235-256, 2006.