Introducing the Common Text Analysis Platform

February 13, 2017

Abstract

The Common Text Analysis Platform (CTAP) is a Web-based computational system for automatic extraction of linguistic features from language productions. It combines state-of-the-art Natural Language Processing (NLP) technologies and complexity research to provide language researchers and education practitioners a tool to effectively and efficiently analyze large amount of language data.

One use case of the CTAP platform is supporting modeling of text readability or language proficiency development from the complexity perspective (Housen et al., 2012), which is difficult, if not impossible, without the help of modern NLP technologies since it usually involves analysis of large quantities of texts. A number of automatic complexity analysis tools such as the Syntactic and Lexical Complexity Analyzers (Lu, 2010), CohMetrix (McNamara et al., 2014), and the Tool for the Automatic Analysis of Lexical Sophistication (Kyle and Crossley, 2015) have emerged in the past few years. Although these systems provide a valuable toolkit for analyzing language, they are geared more towards expert users of computers. Furthermore, a comprehensive analysis of large volumes of language data is only achievable by utilizing the individual tools separately, because each of the tool deals with one or a few aspects of the complexity construct. Consequently, a platform that allows easy and comprehensive acquisition of complexity measures from texts to support research on readability assessment, performance assessment and proficiency development is on demand.

The CTAP system is designed to meet these needs. It features 1) a consistent, easy-to-use, and friendly user interface, 2) modularized, reusable, and collaborative development of analysis components, and 3) flexible corpus and feature management. Four main user modules, namely the Corpus Manager, the Feature Selector, the Analysis Generator, and the Result Visualizer and a server module make up the system. The Corpus Manager helps users organize language materials to be analyzed into corpora, labeled groups, and corpus folders. The Feature Selector allows for selection of different complexity measures for different analysis needs. The selected features could then be applied to different corpora as an analysis with the Analysis Generator. The analysis results or

complexity feature values can be plotted with the Result Visualizer or be downloaded as Comma Separated Values files for further analysis with external statistical tools.

More than 170 complexity features including measures of lexical density, lexical variation, lexical sophistication and syntactic complexity have been included into the CTAP system and the feature list will keep growing with contributions from developers all over the world thanks to the open-source nature of the system. This is by far the most comprehensive and easy-to-use system freely available online for complexity analysis.

The CTAP system as both a running production Web application and an open-source project for collaboration is accessible at `http://ctapweb.com` and `https://github.com/ctapweb` respectively.

References

Housen, A., Kuiken, F., and Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In Housen, A., Kuiken, F., and Vedder, I., editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 1–20. John Benjamins.

Kyle, K. and Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press, Cambridge, M.A.