CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis

Xiaobin Chen,
xiaobin.chen@uni-tuebingen.de
Detmar Meurers,
dm@sfs.uni-tuebingen.de

Tübingen University

December 1, 2016

Linguistic analysis of texts

(Automatic) Linguistic analysis has been widely used for:

- assessing text readability
- modeling processing difficulty of sentences
- analyzing/scoring student writings
- comparing languages typologies and their historical development
- attributing authorship
- identifying native languages
- detecting plagiarism
- assessing answers to questions
- predicting diseases
- ...

Existing tools for for text analysis

A number of tools have been released in the past few years. e.g.

- Syntactic and Lexical Complexity Analyzers (Lu, 2010)
- Cohmetrix (McNamara et al., 2014)
- Suite of Linguistic Analysis Tools (Crossley et al., 2016a,b), also `http://www.kristopherkyle.com/tools.html`
- Computerized Propositional Idea Density Rater (Brown et al., 2008, CPIDR).
- ETS's TextEvaluator `https://texteval-pilot.ets.org/TextEvaluator/`
- Pearson's Reading Maturity Metric
- Text Analysis, Crawling, and interpretation Tool (Dehghani et al., 2016, TACIT)

Problems with existing tools

- Duplicated work: significant amount of feature overlap
- (Re)Usability of tools and analysis components
    - OS-dependent standalone deploy
    - Source code release hard to use for non-programmers
    - Unfriendly user interface
- Feature prolieration, e.g.
    - CohMetrix: 106 metrics
    - Housen (2015): >200 features for measuring L2 complexity
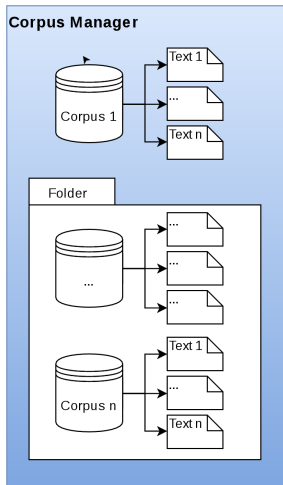   - Vajjala (2015): >200 features for readability assessment

System demands

A system that is:

- Web-based
- user-friendly
- modularized and reusable in terms of analysis components
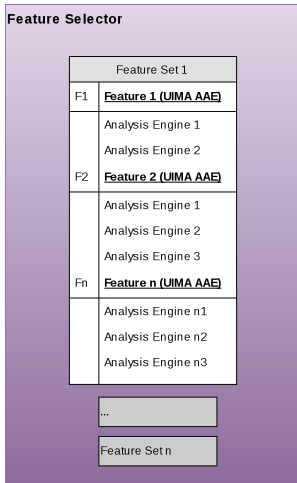- supports real-life use by ordinary users

CTAP System Architecture
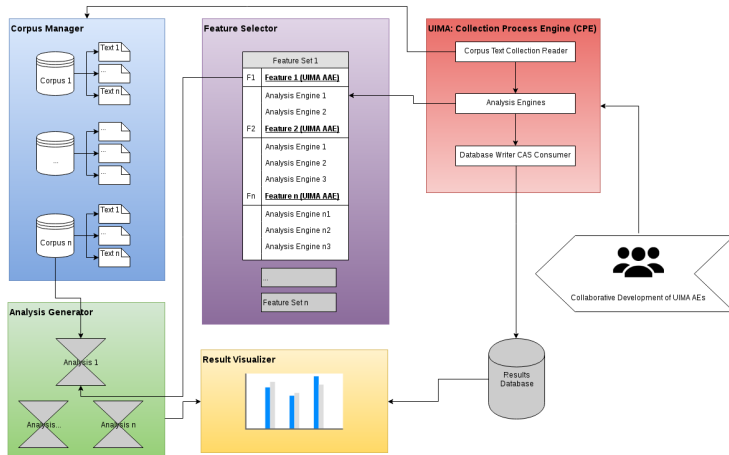
Corpus Manager

Helps users manage the language materials that need to be analyzed.

- use folders to groups corpora
- use corpora to hold texts
- use tags to label texts based on e.g. document genre, target reader levels, etc.
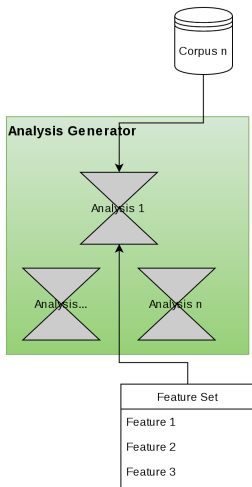
Feature Selector

The Feature Selector supports:

- creating feature set to hold selected features
- add/remove features from feature set

Developers are encouraged to participate in in feature development at `https://github.com/ctapweb`.
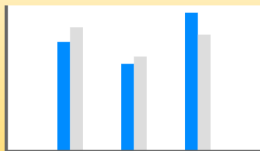
Analysis Generator

Each analysis extracts a set of features from the designated corpus. The analysis generator is used to:

- create new analyses
- run analyses and monitor their progress
- export analysis results in CSV format
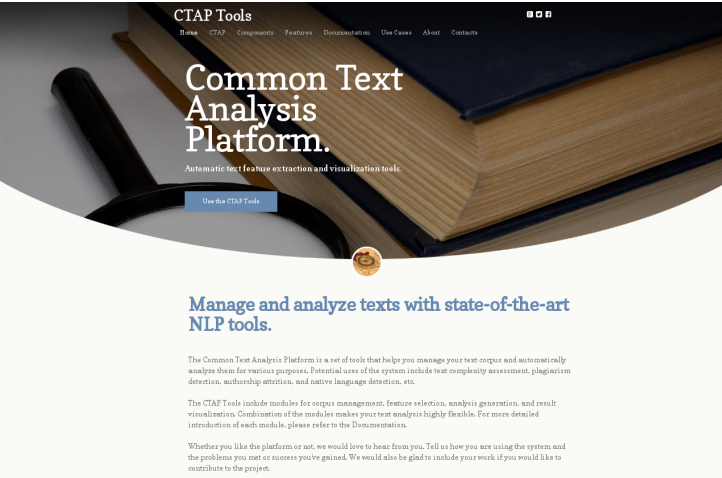
Result Visualizer

The Result Visualizer is a simple and intuitive module that plots analysis results for the user to visualize preliminary findings from the analysis.

Design features of CTAP

- Consistent, easy-to-use, friendly user interface
- Modularized, reusable, and collaborative development of analysis components
- Flexible corpus and feature management

System demo

http://ctapweb.com

Outlook

- Populating the system with more features
- Replicating studies that involved text analysis to validate the system and identify other function needs
- Model construction functionality (machine learning)

More details available in the paper:
Chen, X.B., Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of The Workshop on Computational Linguistics for Linguistic Complexity*. Osaka, Japan. The International Committee on Computational Linguistics.

References

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., and Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2):540–545.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016a). Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, pages 1–19.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., Singh, A., Shankar, Y., Pulickal, L., Rajkumar, A., and Parmar, N. J. (2016). Tacit: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, pages 1–10.

Housen, A. (2015). L2 complexity—a difficult(y) matter. Oral presentation given at the Measuring Linguistic Complexity: A Multidisciplinary Perspective workshop, Université catholique de Louvain, Louvain-la-Neuve.

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.

Vajjala, S. (2015). *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. PhD thesis, University of Tübingen.