Automated Grammatical Error Correction for Language Learners

Joel Tetreault
Yahoo Labs
111 W. 40th Street
New York, NY, 10018, USA
tetreaul@yahoo-inc.com

Claudia Leacock
McGraw-Hill Education CTB
22 Ryan Ranch Road
Monterey, CA, 93940
claudia.leacock@ctb.com

Tutorial Description

A fast growing area in Natural Language Processing is the use of automated tools for identifying and correcting grammatical errors made by language learners. This growth, in part, has been fueled by the needs of a large number of people in the world who are learning and using a second or foreign language. For example, it is estimated that there are currently over one billion people who are non-native writers of English. These numbers drive the demand for accurate tools that can help learners to write and speak proficiently in another language. Such demand also makes this an exciting time for those in the NLP community who are developing automated methods for grammatical error correction (GEC). Our motivation for the COLING tutorial is to make others more aware of this field and its particular set of challenges. For these reasons, we believe that the tutorial will potentially benefit a broad range of conference attendees.

In general, there has been a surge in interest in using NLP to address educational needs, which in turn, has spawned the recurring ACL/NAACL workshop "Innovative Use of Natural Language Processing for Building Educational Applications" that had its 9th edition at ACL 2014. The last three years, in particular, have been pivotal for GEC. Papers on the topic have become more commonplace at main conferences such as ACL, NAACL and EMNLP, as well as two editions of a Morgan Claypool Synthesis Series book on the topic (Leacock et al., 2010; Leacock et al., 2014). In 2011 and 2012, the first shared tasks in GEC (Dale and Kilgarriff, 2011; Dale et al., 2012) were created, and dozens of teams from all over the world participated. This was followed by two successful CoNLL Shared Tasks on the topic in 2013 and 2014 (Ng et al., 2013; Ng et al., 2014).

While there have been many exciting developments in GEC over the last few years, there is still considerable room for improvement as state-of-the-art performance in detecting and correcting several important error types is still inadequate for real world applications. We hope to engage researchers from other NLP fields to develop novel and effective approaches to these problems. Our tutorial is specifically designed to:

- Introduce an NLP audience to the challenges that language learners face and thus the challenges of designing NLP tools to assist in language acquisition
- Provide a history of GEC and the state-of-the-art approaches for different error types
- Show the need for multi-lingual error correction approaches and discuss novel methods for achieving this
- Discuss ways in which error correction techniques can have an impact on other NLP tasks

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts, pages 8–10, Dublin, Ireland, August 23-29 2014.

Outline

1. Introduction
2. Special Problems of Language Learners
   - Errors made by English Language Learners (ELLs)
   - Influence of L1
3. Heuristic and Data Driven Approaches to Error Correction
   (a) Early heuristic rule-based methods
   (b) Methods for detection and correction
   (c) Types of training data
   (d) Features
   (e) Web-based methods
4. Annotation and Evaluation
   (a) Annotation schemes
   (b) Proposals for efficient annotation
   (c) Evaluation Measures
   (d) Crowdsourcing for annotation and evaluation
5. Current Trends in Error Correction
   (a) Detection of ungrammatical sentences and Other error types
   (b) Shared tasks
   (c) Going beyond the classification methodology
   (d) Error correction in other languages
6. Conclusions

Organizers

Joel Tetreault is a Senior Research Scientist at Yahoo Labs in New York City. His research focus is Natural Language Processing with specific interests in anaphora, dialogue and discourse processing, machine learning, and applying these techniques to the analysis of English language learning and automated essay scoring. Previously he was Principal Manager of the Core Natural Language group at Nuance Communications, Inc. where he worked on the research and development of NLP tools and components for the next generation of intelligent dialogue systems. Prior to Nuance, he worked at Educational Testing Service for six years as a Managing Senior Research Scientist where he researched automated methods for detecting grammatical errors by non-native speakers, plagiarism detection, and content scoring. Tetreault received his B.A. in Computer Science from Harvard University (1998) and his M.S. and Ph.D. in Computer Science from the University of Rochester (2004). He was also a postdoctoral research scientist at the University of Pittsburgh's Learning Research and Development Center (2004-2007), where he worked on developing spoken dialogue tutoring systems. In addition he has coorganized the Building Educational Application workshop series for 7 years, the CoNLL 2013 Shared Task on Grammatical Error Correction, and is currently NAACL Treasurer.

Claudia Leacock is a Research Scientist at McGraw-Hill Education CTB who has been working on using NLP in educational applications for 20 years focusing on automated scoring and grammatical error detection. She was previously a consultant for Microsoft Research where she collaborated on the development of ESL Assistant: a web-based prototype tool for detecting and correcting grammatical errors of English language learners. As a Distinguished Member of Technical Staff at Pearson Knowledge

Technologies, and previously as a Principal Development Scientist at Educational Testing Service, she developed tools for automated assessment of short-response content-based questions and for grammatical error detection. As a member of the WordNet group at Princeton University's Cognitive Science Lab, her research focused on word sense disambiguation. Dr. Leacock received a B.A. in English from NYU, a Ph.D. in Linguistics from the City University of New York, Graduate Center and was a post-doctoral fellow at IBM, T.J. Watson Research Center.

References

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France, September. Association for Computational Linguistics.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada, June. Association for Computational Linguistics.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreaukt. 2010. *Automated grammatical error detection for language learners*. Morgan & Claypool Publishers.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreaukt. 2014. *Automated grammatical error detection for language learners, second edition*. Morgan & Claypool Publishers.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, and Joel Tetreault. 2013. The conll-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, June. Association for Computational Linguistics.