Text normalization for endangered languages: A shared task challenge

Patrick Littell
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh PA 15213
plittell@cs.cmu.edu

Shobhana Chelliah
University of North Texas
3940 North Elm, Suite B201
Denton, TX 76203
chelliah@unt.edu

Gina-Anne Levow
University of Washington
P.O. Box 352425
Seattle, WA 98195
levow@ut.edu

The development of text technologies for endangered languages often faces a resource bottleneck, in that the text that does exist is written by transcribers of very different skill levels, using a variety of orthographic conventions. This results in very heterogeneous text corpora, and normalization into an orthographically homogeneous form is made challenging by this very lack.

Meanwhile, the users of endangered text technology will likewise have a variety of skill levels and use a variety of orthographic conventions. For example, any given user entry in the Kwak'wala language of British Columbia might be an orthographically correct form (e.g., 1a), or might be from a variant but still systematic orthography (1b), or might be an unsystematic, majority-language-influenced rendering (1c), or might be somewhere in between these, like an attempt at an orthographic rendering by a student who cannot yet reliably distinguish all the necessary phonemic differences (1d).

(1) a. *Tłuma̲nu'x̲ pusk̲a x̲wa 'nala!*
very.1SG.EXCL hungry this day

“We're very hungry today!”

b. ƛumənuʔx̌ pusqax̌ʷa n̓ala!

c. Kloomenok pooskah hwanala!

d. Tłumenox puska xwa nala!

Any text technology for an endangered language community therefore requires a significant normalization step, both to assemble the backing text corpus in the first place and to respond appropriately to user-generated text.

To help identify the unique challenges that very-low-resource languages bring to text normalization, and to discover what techniques best address these challenges, we propose a “Shared Task Evaluation Challenge” (STEC) on orthographic regularization in several endangered languages. STECs have become an important driver of progress in NLP (Belz and Kilgarriff, 2006), and several shared tasks have concentrated particularly on text normalization, such as Dale and Kilgarriff (2011), Mohit et al. (2014), and Baldwin et al. (2015).

Expanding such tasks to endangered languages poses an interesting challenge for existing normalization systems, allowing the NLP community to test whether their techniques generalize beyond well-studied languages, and meanwhile providing a valuable service to the language communities in question. Many communities have collections of texts in heterogeneous orthographies, and writers have often been trained in different orthographies (and trained to varying degrees), so the possibility of normalizing texts (both old and new) to a consistent format can solve many practical problems communities face.

Acknowledgments

This project is funded by NSF grant #1500157.

References

Timothy Baldwin, Young-Bum Kim, Marie Catherine de Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP*, 126:2015.

Anja Belz and Adam Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of INLGC 4*, pages 133–135, Sydney, Australia. Association for Computational Linguistics.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the EWNLG 13*, ENLG '11, pages 242–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of EMNLP 2014 ANLP Workshop*, pages 39–47.