Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research

Gregory Grefenstette
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
gregory.grefenstette@inria.fr

Abstract

Comparable Corpora have been used to improve statistical machine translation, for augmenting linked open data, for finding terminology equivalents, and to create other linguistic resources for natural language processing and language learning applications. Recently, continuous vector space models, creating and exploiting word embeddings, have been gaining in popularity in more powerful solutions to creating, and sometimes replacing, these resources. Both classical comparable corpora solutions and vector space models require the presence of a large quantity of multilingual content. In this talk, we will discuss the breadth of this content on the internet to provide some type of intuition in how successful comparable corpus approaches will be in achieving its goals of providing multilingual and cross lingual resources. We examine current estimates of language presence and growth on the web, and of the availability of the type of resources needed to continue and extend comparable corpus research. .

Keywords: web mining, under-resourced languages, comparable corpora, language resources

Bibliographical References

Barbaresi, A. (2015). *Ad Hoc And General-Purpose Corpus Construction From Web Sources* (Doctoral dissertation, ENS Lyon).

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language resources and evaluation, 43(3), 209-226.

Dimitrova, L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H. J., & Tufis, D. (1998, August). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central And Eastern European Languages. In *Proceedings Of The 17th International Conference On Computational Linguistics,* ACL, Volume 1 pp. 315-319.

Gatto, M. (2011). The 'Body' and The 'Web': The Web As Corpus Ten Years On. *ICAME J.,* 35, 35-58.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC* , pp. 759-765.

Grefenstette, G., & Nioche, J. (2000). Estimation of English and non-English Language Use on the WWW. In *Content-Based Multimedia Information Access-Volume 1,* RIAO 2000, pp. 237-246.

Hale, S. A. (2012). Net Increase? Cross-Lingual Linking in the Blogosphere. Journal of Computer-Mediated Communication, 17(2), 135-151.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the Special Issue On The Web As Corpus. *Computational Linguistics*, 29(3), 333-347.

Pimienta, D., Prado, D., & Blanco, Á. (2009). Twelve Years Of Measuring Linguistic Diversity In *The Internet: Balance And Perspectives.* Paris: United Nations Educational, Scientific and Cultural Organization.

Rehm, G., & Uszkoreit, H. (2011). Multilingual Europe: A Challenge For Language Tech. *MultiLingual*, 22(3), pp. 51-52.

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links That Speak: The Global Language Network And Its Association With Global Fame. *Proceedings of the National Academy of Sciences*, 111(52), E5616-E5622.

Scannell, K. P. (2007). The Crúbadán Project: Corpus Building for Under-resourced Languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, Vol. 4, pp. 5-15.

Soria, C., Calzolari, N., Monachini, M., Quochi, V., Bel, N., Choukri, K., Mariani, J., Odijk, J. and Piperidis, S., (2014). The Language Resource Strategic Agenda: the FLaReNet Synthesis Of Community Recommendations. *Language Resources and Evaluation,* 48(4), pp.753-775.

Van der Veken, A., & De Schryver, G. M. (2003). Les Langues Africaines Sur La Toile: Étude Des Cas Haoussa, Somali, Lingala Et Isixhosa [The African Languages on the Internet: Case Studies for Hausa, Somali, Lingala and isiXhosa]. *Cahiers du RIFAL*, 23, 33-45.