Recency is good: expanding with fresh news improves event detection in Twitter

Saša Petrović Miles Osborne Victor Lavrenko
University of Edinburgh

ABSTRACT

Twitter is a popular microblogging site that is a good source of real-time information. Detecting events in Twitter is an ongoing research effort and a fundamental task is clustering tweets according to which (news) event they describe. Document expansion can improve this clustering, especially for Twitter, given that tweets are short. While document expansion using external corpora has been around for years [1], all previous work treats the external corpus as temporally static. We are the first to treat the external corpus (newswire articles in this case) as a time-synchronous stream, expanding tweets with words found in similar, temporally aligned newswire articles. Tweets are expanded with terms from the most similar newswire document, where the terms are weighted by the cosine similarity between the tweet and the newswire document [2]. Using the tweet corpus compiled by [3], and newswire data from the same time period, coming from eight major newswire sources (Reuters, CNN, BBC, New York Times, Google News, Guardian, Wired, The Register), we find that using timely newswire for expansion material improves event detection for Twitter more than using older newswire for the same purpose.

BODY

Expanding tweets with fresh and with stale newswire improves event detection by 21% and 17% respectively, compared to not using expansion.

REFERENCES

[1] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, 2006.

[2] S. Petrović. *Real-time event detection in massive streams.* PhD thesis, University of Edinburgh, 2012.

[3] S. Petrović, M. Osborne, and V. Lavrenko. Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–346, 2012.

Volume 2 of Tiny Transactions on Computer Science