Comparing Topic Model Stability across Language and Size

Simon Hengchen*
Université libre de Bruxelles
Information Science Department
Avenue Roosevelt 50 – CP 123
1050 Brussels
Belgium
shengche@ulb.ac.be

Alexander O'Connor
ADAPT Centre
School of Computing
Dublin City University
Glasnevin, Dublin 9
Ireland
Alexander.OConnor@dcu.ie

Gary Munnelly
ADAPT Centre
Trinity College Dublin
College Green, Dublin 2
Ireland
munnellg@tcd.ie

Jennifer Edmond
Long Room Hub
Trinity College Dublin
College Green, Dublin 2
Ireland
Jennifer.Edmond@tcd.ie

April 2016

Abstract

The rapid evolution of technology has freed the written word from the physical page. In the current era, it can be argued that the primary means of access to text is digitally mediated. This has given unprecedented reach to any individual with access to the Internet. However, the rate at which a human can absorb such information remains relatively unchanged, in particular in the case of linguistically and/or culturally complex data. Results in computer science continue to advance in areas of linguistic analysis and natural language processing, facilitating more complex numerical inquiries of language. This commoditisation of analytical tools has led to widespread experimentation with digital tools within the humanities: recent initiatives such as DARIAH[1], CENDARI[2] or TIC-Belgium[3] try to foster the use of computational methods and the reuse of digital data by and between researchers and practitioners alike.

A key question emerges: to what extent do these digital tools reveal signal, and to what extent are they merely responding to noise? This is a question of particular import to humanities researchers, for whom the difference between signal and noise may shift from project to project and from interpreter to interpreter, not to mention from linguistic context to linguistic context. Scholars currently must resort to a vehicular language (in Europe and North America, generally English) in order to find patterns between cultural and linguistic contexts. This approach is not wholly satisfying, however, where the sensitivities surrounding the object of

*Corresponding author

[1] http://dariah.eu/

[2] http://cendari.eu/

[3] http://tic.ugent.be/

study are high, meaning that speakers would choose specific words and phrases with great care, aware of the resonances of the choices.

Discourse regarding cultural traumas, such as war, occupation, economic collapse, environmental disaster, or other major disruption to national identity and social cohesion, present a clear example of this kind of issue: culturally specific, and yet present at some level or other in nearly every cultural narrative. The international SPECTRESS network[4] had hoped to provide a new approach to fostering cross-cultural dialogue regarding the impact of and responses to cultural trauma by topic modelling discourse around traumas, and seeking similar clustering effect across language- and event-specific contexts. The challenge with this approach was that appropriate corpora were generally too small to produce reliable models and results. However, initial experiments were not able to answer one key question of interest to both the computer scientists and the humanists in the project team: how small is too small?

We focus on the study of language and the semi-automatic discovery of topics in textual data. In order to extract meaning we use two algorithms, both often referred to as "topic modelling techniques": Latent Semantic Analysis (LSA) (Landauer et al., 1998) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Both algorithms construct matrices to try to determine topics within a set of texts by clustering similar words. These approaches both encode key assumptions about the statistical properties of the language, with statistical and stochastic aspects included. Whilst LDA is the most widely used algorithm in the literature these past years, we believe that a benchmarking study should include more than one take at the data, which is why we are comparing LDA and LSA. Both models also need a certain amount of input data to produce viable results – a notion that still has to be determined, as pointed out by Greene *et al* (Greene et al., 2014). Unfortunately, it is unclear how much data is enough. This lack of clear understanding of minimal functional corpus size poses a serious threat to topic modelling's viability as humanistic methodology. Topic modelling is currently an approach humanists are very aware of and see potential uses for (following the work of Jockers (Jockers, 2013; Jockers and Mimno, 2013) and others), but as many humanistic corpora are on the small side, the threshold for the utility of topic modelling across DH projects is as yet highly unclear. Unstable topics may lead to research being based on incorrect foundational assumptions regarding the presence or clustering of conceptual fields on a body of work or source material. Stable topics, however, indicate that the random component in the process has been minimised and the topics given do possess a coherence worthy of further investigation by a trained human, as advocated by Chang *et al* (Chang et al., 2009).

Building on previous work by Munnelly *et al* (Munnelly et al., 2015), we propose a methodology to try to determine how large a corpus must be to establish a stable model, with an added twist: whilst topic modelling techniques are language-independent, i.e. "use[] no manually-constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like."(Landauer et al., 1998), the morphology of the language processed can influence the size of the corpus required to build a stable set of topics. In order to do so, we compare French and English topic models from a bilingual corpus of articles.

Methodology We use the DBpedia (Auer et al., 2007) interlanguage links for the English language (`interlanguage-links_en.nt`) to search for every DBpedia URI existing in French and in English[5].

With all DBpedia URIs having a match – and linked via the `owl:sameAs` predicate – in both languages, we then parse both `long_abstracts_en.ttl` and `long_abstracts_fr.ttl` files to extract their respective long abstracts.

This process carried through, we decompose the resulting files in a number of smaller files: one for every DBpedia entity, each containing its abstract. With both corpus segments constituted, it is possible to apply LSA and LDA. The resulting models are stored and measured. The

[4] https://spectressnetwork.wordpress.com/

[5] The files are freely available for download at http://wiki.dbpedia.org/Downloads2015-10.

corpora are reduced in size, LDA and LSA re-applied, models stored, and corpora re-reduced, iteratively, each time recording the topic results.

Topic models are compared manually between languages at each stage, and programmatically between stages, using the Jaccard Index (Real and Vargas, 1996), for both languages.

A large deviation between stages indicates a loss of representativeness between models.

Perspectives By applying our methodology on parallel corpora, we try to determine whether the minimum sample size for a representative topic model is consistent across the two languages studied, i.e. French and English. Using the built-in multilingualism of DBpedia, it becomes possible to reapply the methodology on most written languages.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data.* Springer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Greene, D., O'Callaghan, D., and Cunningham, P. (2014). How many topics? Stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513. Springer.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history.* University of Illinois Press.

Jockers, M. L. and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Munnelly, G., O'Connor, A., Edmond, J., and Lawless, S. (2015). Finding meaning in the chaos.

Real, R. and Vargas, J. M. (1996). The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385.