Analysis of Word Semantic Change in Political Texts through Spatio-Temporal Word Embeddings

Aurélien Bellet, Clément Bellet, Pascal Denis, Caroline Le Pennec

October 4, 2017

Team and contact

- The internship will take place within the Magnet Team at INRIA (Lille, France): http://team.inria.fr/magnet
- The internship will be supervised by:
  Aurélien Bellet (INRIA, aurelien.bellet@inria.fr)
  Clément Bellet (INSEAD, clement.bellet@gmail.com)
  Pascal Denis (INRIA, pascal.denis@inria.fr)
  Caroline Le Pennec (UC Berkeley, clpennec@berkeley.edu)

Keywords

Natural Language Processing, Machine Learning, Word Embeddings, Political Texts.

Context

How to adequately represent the meaning of words is a long-standing and crucial problem in the fields of Natural Language Processing. Following the "distributional" hypothesis according to which the meaning of a word can be inferred from the context in which it is used, there has been a recent surge of research in learning vectorial word representations (or "word embeddings") from co-occurrence statistics extracted from massive text corpora. Existing algorithms typically rely on neural networks (this includes the popular `word2vec` approach [7, 8]) or some variants of matrix factorization [9, 6]. The main appeal of these low-dimensional word representations is two-fold: they can be derived directly from raw text data in an unsupervised or weakly-supervised manner, and their latent dimensions condense interesting distributional information about the words. For instance, word embeddings have been shown to convey some syntactic or semantic relationships between words (including word similarity or syntactic/semantic analogy tasks) [7].

Objectives

The goal of this internship is to study word semantic change in political texts across time and space using word embeddings. We will work on a large corpus of political manifestos from the French general elections for the period 1958–1993.[1] We aim to uncover significant changes in some words' usage and meaning across time and space. We expect some of these evolutions to be consistent with known historical and regional shifts in political language. We will contrast this evolution to that of the general language, and attempt to highlight correlations with factors such as important political events, levels of wealth inequality, etc.

[1] Note that the text corpus is in French. See https://archive.org/details/archiveselectoralesducevipof

From the machine learning and natural language processing perspective, a simple first approach to tackle these questions consists in learning word embeddings separately for each election year and/or for each spatial region (e.g., electoral district). However, the resulting word embeddings may be unstable and misleading due to data scarcity. We will thus propose some approaches to learn word embeddings jointly across time and space, building upon recent work on dynamic word embeddings [2, 5, 3] as well as in relation to political contexts [1, 4].

The tentative work-plan is as follows:

1. Review the relevant literature on word embeddings, extract and pre-process the text corpus.
2. Propose and apply methods to learn word embeddings which vary over time and space.
3. Use the resulting embeddings to analyze word semantic change in political language.

Skills

Basics in machine learning, algorithms and linear algebra. Familiarity with natural language processing and interest for political science are a plus.

References

[1] H. Azarbonyad, M. Dehghani, K. Beelen, A. Arkut, M. Marx, and J. Kamps. Words are Malleable: Computing Semantic Shifts in Political and Media Discourse. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, 2017.

[2] R. Bamler and S. Mandt. Dynamic Word Embeddings. In *Proceedings of the International Conference on Machine Learning*, 2017.

[3] H. Dubossarsky, E. Grossman, and D. Weinsha. Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[4] M. Gentzkow, J. M. Shapiro, and M. Taddy. Measuring polarization in high-dimensional data: Method and application to congressional speech. Technical Report w22423, National Bureau of Economic Research.

[5] W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

[6] R. Lebret and R. Collobert. Word Embeddings through Hellinger PCA. In *EACL*, 2014.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. Technical report, arXiv:1301.3781, 2013.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 2013.

[9] J. Pennington, R. Socher, and C. D. Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.