Vector based measure of semantic shifts across different cultural corpora as a proxy to comparative history of ideas

Marianne Reboul, Alexandre Gefen, Mark Andrew, David Mcclure, J. D. Porter, Marine Riguet

Vector based measure of semantic shifts across different cultural corpora as a proxy to comparative history of ideas

Alexandre Gefen (Université Paris Sorbonne), Mark Andrew Algee-Hewitt, David McClure (Stanford University), Frédéric Glorieux, Marianne Reboul (Université Paris Sorbonne), J.D. Porter (Stanford University), Marine Riguet (Université Paris Sorbonne)

Our joint research group, a collaboration between Paris Sorbonne and Stanford University, traces the parallel birth and evolution of key modern literary and aesthetic concepts in French and English using vector based semantic analysis methods. This inter-institutional, multilingual, collaboration aims not only to explore the development of “literature” as a distinct discursive and disciplinary field, but also to discover the mutual influence — or, perhaps, surprising independence — of these two closely related national literary traditions. By sharing our methods and results, we hope to foster the use of distributional semantics as a useful tool for investigating the comparative history of ideas in the literary field.

For this project, we have elected to compare two corpora of important and representative XIXth century magazines: in France, La Revue des Deux mondes (1829-1893), and in English, the Blackwood's Magazine (1817-1880), both influenced by the emergence of Romanticism. We started with a simple word, « littérature/literature » (9100 occurrences in the Blackwood corpus, 12400 occurrences in the Revues des deux mondes) — a word that we know has grown specialized during the century — in order to plot its semantic evolution and semantic space.

In this paper, we discuss the methods we have developed to:
Select, prepare and compare dissimilar corpora of old newspapers. After many discussions, we identified these two magazines as the most equivalent single-periodical literary corpora in our respective national literatures. To maintain this equivalence, we restricted the two corpora to the same time period, 1830-1880. We preprocessed both corpora with custom methods conversion to TEI encoding, and lemmatized them with Alix (Alix is a custom-designed lemmatizer and Part-Of-Speech tagger) in French, with the Stanford CoreNLP tagger in English.

Graph the comparative historical evolution of our key word « littérature/literature » in vector space using word2vec and Glove vectors. This builds upon our previous work with the HathiTrust corpus in which we identified the lexemes with the most distinctive histories in relation to literature. In this pilot attempt, we identified words that appeared frequently on the same page as “literature” but with significant variation over time -- words that shifted from low to high levels of correlation with “literature” across literary history, and vice versa. It is our contention that by moving from this initial method to a vector-based analysis of our selected periodical corpora, we will better be able to capture the nuances of the semantic field of “literature” over time.

Define a reliable way to measure polysemy. First, we build a set of chosen concept words with their equivalent in both languages in WordNet. Eg. literature:littérature ; poetry:poésie ; art:art ; etc. Next, we build a set of bigrams for each of the words of this set, with the ten most associated adjectives. We then compute the cosine distance between each bigram and apply the methods used in Köper and Schulte im Walde 2014: once the cosine distance is computed for each bigram vector, build a matrix with ranking proximity of those vectors, and compute their mean rank. We are then able to determine a score that would weight the computing of the clustering coefficient of bigrams. As a low-rank bigram should in our assumption represent a highly polysemic term, the lower the rank, the higher the weight on the clustering coefficient. We then determine the clustering coefficient of each term in diachronic sequences, to draw a dynamic representation in Gephi.

This method allows us to find possible semantic attractors in the evolution of key literary concepts and to plot the behaviour of specific semantic word pairs (for instance, littérature/poésie and literature/poetry). In our final cultural interpretation, we compare semantic matrices and historical evolutions in the two corpora/languages and confront the results with known existing hypotheses in the history of aesthetic concepts in order to confirm the supposed specialization of the field of literary studies. By focusing on concepts that undergo dramatic transitions during the period of study, our analysis brings together the study of diachronic conceptual change and synchronic polysemy, allowing us to probe how multiple senses of a word coexisting in tension can eventually give rise to changes in meaning.

Bibliography

[1] Boleda, Gemma, Sebastian Padó and Jason Utt. “Regular polysemy: A distributional model.” SemEval 12 (2012): 151-160.
[2] Hamilton, William, Jure Leskovec, J. and Dan Jurafsky. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.” arXiv:1605.09096 [cs.CL] (2016)
[3] Jorge-Botana, Guillermo, José A. León and Ricardo Olmos. “The representation of polysemy through vectors: some building blocks for constructing models and applications with LSA.” Int. J. Cont. Engineering 21 (2011): 328 – 342.
[4] Köper, Maximilian, and Sabine Schulte im Walde. "A Rank-based Distance Measure to Detect Polysemy and to Determine Salient Vector-Space Features for German Prepositions." LREC 9 (2014): 4459-4466.
[5] Mu, Jiaqi, Suma Bhat, Pramod Viswanath. “Geometry of Polysemy.” arXiv:1610.07569 [cs.CL] (2016).
[6] Ravin, Yael, and Claudia Leacock (editors). Polysemy: Theoretical and Computational Approaches. New York: Oxford UP, 2000.
[7] Springorum, Sylvia, Sabine Shulte im Walde and Jason Utt. “Detecting Polysemy in Hard and Soft Cluster Analyses of German Preposition Vector Spaces.” IJCNLP 6 (2013): 632-640.