Automatic Detection of Semantic Fields for Political Science Research

Beata Beigman Klebanov, Daniel Diermeier, Eyal Beigman
Kellogg School of Management, Northwestern University

Abstract: One of the ways in which a given text makes sense to fellow humans is by drawing on shared domains of knowledge and experience, termed *semantic fields* in linguistic literature. We discuss three approaches to automatically capture the semantic fields that a given text draws upon. For a given author, we identify recurrent and characteristic semantic fields, as well as fields that are singular to the particular text at hand. We exemplify the various approaches using Margaret Thatcher's speeches.

The notion of semantic fields draws on early 20th century linguistic theories, and was succinctly defined by Kittay as "a content domain articulated by a lexical field" (Kittay, 1987). Content domains can be perceptual (such as colors), cultural (kinship), conceptual (a certain scientific theory), experiental (life cycle), or indeed can be tied to any identifiable activity. For a content domain to be articulated by a lexical field means that various entities, events, and relations within the domain can be systematically expressed by certain sets of words. In a text, a semantic field materializes as a group of words from the field in sufficient number and proximity to each other.

Semantic fields are a promising approach to formalize one of the key concepts in political communication: the notion of *framing* (Callaghan & Schnell, 2001; Chong & Druckman, 2008; Druckman & Nelson, 2003; Entman, 2003; Tversky & Kahneman, 1981). An issue is framed when words used to discuss it are drawn from a particular part of the relevant semantic field, at the expense of other parts. For example, stressing the risky aspect of a proposed policy at the expense of its potential benefits creates a frame that encourages risk and uncertainty aversion which may lead to a status quo bias and the rejection of the proposed policy. Alternatively, framing can occur when lexica from a different semantic field is consistently used when discussing an issue in question (Lakoff, 2002), such as using military terms (*war on drugs*, *enemy*, *battle*) when discussing drug use, rather than public health terminology (*treatment*, *epidemic*). Such framing is important because it may make some policy solutions (e.g. restrictions on civil rights) more plausible than others (e.g. a methadone program) because they make sense in the context of a war, but not as a public health response.

We discuss three types of approaches to automatic identification of semantic fields in the context of political speech: unsupervised clustering, dictionary-based methods, and methods based on experimental data.

Unsupervised clustering, represented by Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), is well suited for detecting recurrent issues in a corpus, reflecting faithfully the way they are construed by the author(s) of the corpus. Thus, in a corpus of over 6000 public statements by Thatcher, the discussion of nuclear weapons is tied to the Soviet Union, and that of Argentina – to military invasion and Falklands. We also found that Northern Ireland is consistently framed from the perspective of the bloody conflict there, evidenced by cluster members such as *violence* and *regiment*.

Dictionary-based methods use pre-existing sets of categories designed by expert lexicographers to perform semantic analysis; we use WMatrix (Rayson, 2003) as a representative. Such methods are most useful in a comparative perspective, detecting semantic domains that are characteristic, or discriminative, for the given author as opposed to another comparable individual. Comparing 11 Thatcher's speeches to the Conservative Party Conference to Blair's 13 speeches to Labour's Party Conference, we found a number of categories that clearly separate between Tories and Labour. For Thatcher, examples include industry and defense, while for Blair we have education and medical care. The same observation applies to the unemployment vs. poverty rhetoric. Furthermore, the *New* aspect of *New Labour* clearly shows in the extensive lexica of the New and Young and Change categories, as opposed to the continuity language of Thatcher.

Approaches of the third type, exemplified by lexical cohesion analyzer (Beigman Klebanov, 2007), draw on reader based experiments for detection of semantically related pairs of words in a text, subsequently using supervised machine learning to match the human annotations. Analyzing Thatcher's 1977 speech to the party, we detected some recurrent and characteristic themes, partially seconding LDA and WMatrix findings, as well as additional semantic fields that are not sufficiently recurrent, and hence are missed by LDA, and fall outside of the WMatrix inventory of categories, yet represent important rhetorical tools used by Thatcher in this speech to attain specific political ends. It is for detection of this last type of semantic domains – the singular rather than the typical in the text at hand – that this type of analysis holds most promise.

References

Beigman Klebanov, B. (2007). *Experimental and computational investigation of lexical cohesion in English texts.* Unpublished doctoral dissertation, The Hebrew University of Jerusalem, Israel.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Resarch*, *3*, 993-1022.

Callaghan, K., & Schnell, F. (2001). Assessing the democratic debate: How the news media frame elite policy discourse. *Political Communication*, *18*(2), 183-213.

Chong, D., & Druckman, J. N. (2008). *Framing public opinion in competitive democracies.* (Forthcoming in *American Political Science Review*)

Druckman, J. N., & Nelson, K. R. (2003). Framing and deliberation: How citizens' conversations limit elite influence. *American Journal of Political Science*, *47*, 729-745.

Entman, R. M. (2003). Cascading activation: Contesting the White House's frame after 9/11. *Political Communication*, *20*, 415-432.

Kittay, E. (1987). *Metaphor: Its cognitive force and linguistic structure.* Oxford University Press.

Lakoff, G. (2002). *Moral politics: How Liberals and Conservatives think* (2nd ed.). The University of Chicago Press.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison.* Unpublished doctoral dissertation, Lancaster University, UK, Retrieved March 8, 2008, from http://www.comp.lancs.ac.uk/computing/users/paul/phd/phd2003.pdf.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453-458.