CO-OCCURRENCE AS AN INDICATOR OF IDEAS AND SENTIMENT

KRISTEN SUMMERS, CACI

The prevalence of terms in online material serves as an indicator of the prominence of their corresponding topics. For example, “trending” hash tags in Twitter indicate topics, or categories of topics, that are gaining attention in the communications within this medium, which in turn indicates some degree of current attention to these topics in general, at least by the segment of the population that uses Twitter. The same type of indication can be found in word usage on social media, and this can be generalized to concepts indicated by a variety of words. How can we move from the prevalence of a topic itself to the prevalence of a point of view about the topic? One way is to look at the words and concepts that occur together with that topic. By comparing these co-occurrences over time and space, we can identify shifts and geographic differences in the way a topic is discussed, which can reflect differences in the way a topic is viewed, i.e., the dominant ideas in play about that topic. For example, if a common topic is “immigration”, and content from one area has heavy co-occurrences of “illegal”, “border”, “enforcement”, and similar words, this indicates the prevalence of a very different idea from an area with heavy co-occurrences of “entrepreneur”, “dream act”, etc.

In prior work at CACI, we have developed a Corpus Analysis tool called CorpusDOG. On a static corpus, this tool allows a user to enter terms and see the terms that co-occur with them heavily, based on the Normalized Google Distance [1]. This exploration alone can produce interesting results. For example, on a collection of Arabic language blog material, the masculine form of the word for suicide co-occurred with very different words than did the feminine form; perhaps unsurprisingly, the masculine form was associated with political words, while the feminine form was associated with domestic and family related terms. On a corpus with temporal information, CorpusDOG shows prominent words over time, and the user can interactively select words and see the other common words with which they co-occur strongly. For example, on March 19, 2011, examining Twitter data for Libya showed the prominent word “missiles” associated at 19:10 GMT with the words “U.S.”, “cruise”, “official” and a few others, and indirectly associated (via “official”) with “submarines”, which in turn was associated with “@breakingnews” and “preparing”. At 19:40 GMT, the word “missiles” was increasing in use and was associated strongly with “launched”, “launches”, “inside”, “fired”, “U.S.”, “cruise”, and “@breakingnews”. The shift in associated terms reflected the change in content away from the *potential* for missiles to be launched to the fact that they had been launched.

Expanding and generalizing these approaches can provide insight into the spread of conceptual associations that reflect points of view and sentiments. By expanding from words to concepts that may be indicated by a variety of words, the analysis can address themes and ideas rather than their specific expression in individual words. These concepts and the words to indicate them may come from established lexical resources such as WordNet [2] for English, or from more domain-specific taxonomies. In some cases it may also be valuable to compare concept-based results with pure word-driven results, since distinguishing between surface forms can provide value, as in the example of the different associations with forms of the word for suicide.

Combining the identification of terms and concepts of increasing prevalence with the analysis of their co-occurrence with other terms and concepts can enable a data-driven analysis of the emergence and changes in the focus and points of view of the groups or regions whose data is considered. The associated concepts may indicate implicit opinions, as in the immigration example above, or similar ideas. They may also indicate sentiment about the topics, whether purely positive or negative or more fine-grained (e.g., word use that signals mood states). Considering a variety of co-occurrence measures may reveal a more effective choice of measure, or different characteristics of the associations of ideas indicated by the different measures.

Applying this type of corpus linguistic analysis to data sources with known geographic and organizational associations can enable a variety of further insights and interpretations relating to the emergence and spread of ideas and sentiment over time, such as:

- Observe the spread of topics and related ideas and sentiment through geography over time. Do the main concepts and their associations emerge in new areas together? Does the topic appear in a new area first and the associations follow? Do the associations stay with the topic, or do topics move but rapidly acquire new associations in new areas?

- Observe the spread of topics and related ideas and sentiment through networks of organizational and social associations. Does this occur similarly to geographic movement, or does it have different characteristics?

- Observe the emergence and spread of topics and related ideas through social media, and compare this to known organizational networks. Does the path through social media mirror the known organizational associations, or does it indicate a different network that may be specific to the medium?

- Identify the “thought leaders” whose ideas and concepts tend to spread. Are these the same for general topics, associated concepts, and sentiment about the topics, or are different individuals and organizations influential with respect to these different aspects of ideas?

Addressing these kinds of questions will require tracking a large number of topics and their associated concepts and sentiment indicators consistently, in a variety of settings. This is an excellent application for the type of corpus linguistics described here, making use of both automated calculations and interactive display and visualization. For example, a useful approach may be to start with semi-automated analysis, refining the indicators and selecting relevant topics to track, followed by automated tracking of the topics and their associations throughout selected sources of social media with known characteristics of interest.

REFERENCES

[1] Cilibrasi, R. and Vitányi, P. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), pages 370—383, 2007.

[2] Fellbaum, C., ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.