Cognition 135 (2015) 4–9

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Analyzing the history of *Cognition* using Topic Models

Uriel Cohen Priva *, Joseph L. Austerweil

Brown University, Department of Cognitive, Linguistic, and Psychological Sciences, Providence, RI 02912, United States

ARTICLE INFO

Article history:
Available online 8 December 2014

Keywords:
History of science
Topic Models
Framing topics

ABSTRACT

Very few articles have analyzed how cognitive science as a field has changed over the last six decades. We explore how *Cognition* changed over the last four decades using Topic Models. Topic Models assume that every word in every document is generated by one of a limited number of topics. Words that are likely to co-occur are likely to be generated by a single topic. We find a number of significant historical trends: the rise of moral cognition, eyetracking methods, and action, the fall of sentence processing, and the stability of development. We introduce the notion of framing topics, which frame content, rather than present the content itself. These framing topics suggest that over time Cognition turned from abstract theorizing to more experimental approaches.

1. Introduction

Many researchers are familiar with discussions during post-conference dinners speculating about which research areas are "hot" and lamenting that their own research area has fallen out of favor. Despite these all too common parlor debates, there is little to no empirical work analyzing how cognitive science as a field has changed over the last six decades (but see Leydesdorff & Goldstone, 2014, who analyze the interaction between cognitive science and related fields via citation analysis). Given that many articles have been digitized to be easily accessible electronically and that the abstract and title of most of these articles are free to download, it is now feasible to perform large, data-driven analyses of scientific trends. In this article, we describe one method for analyzing trends in scientific fields, focusing on the journal *Cognition* as part of this special issue.

What is a rigorous, data-driven method for analyzing trends in science? Perhaps the simplest approach would be to count how often particular words and phrases associated with a particular research area are used each year (e.g., "moral"), and analyze any resulting trends. Recently, Behrens, Fox, Laird, and Smith (2013) took this approach for analyzing publication patterns in cognitive neuroscience, and found that particular brain areas were positively correlated with high-impact journals (e.g., the fusiform gyrus). However, the word-based approach has several limitations. First, the choice of word trends is biased by the investigator's hypothesis, rather than the data. Second, there is a risk that changes in language use do not reflect academic interest itself, e.g. artificial intelligence still generates a lot of interest, but it is no longer referred to as artificial intelligence.[1] Finally, keyword-based approaches would fail to distinguish between two or more senses of a single word. For example, *movement* can refer to eye movement in eyetracking paradigms or to body movement in the study of action.

Rather than use words, Hall, Jurafsky, and Manning (2008) demonstrated that *Topic Models* (Blei, Ng, & Jordan, 2003; Griffiths, Steyvers, & Tenenbaum, 2007) present an appealing approach to track the rise and fall of specific research interests in general. Topic Models allow us to

* Corresponding author.
E-mail addresses: Uriel_Cohen_Priva@Brown.edu (U. Cohen Priva), Joseph_Austerweil@Brown.edu (J.L. Austerweil).

[1] As observed in the Google Books Ngram Viewer http://goo.gl/PUMKQP.

http://dx.doi.org/10.1016/j.cognition.2014.11.006
0010-0277/

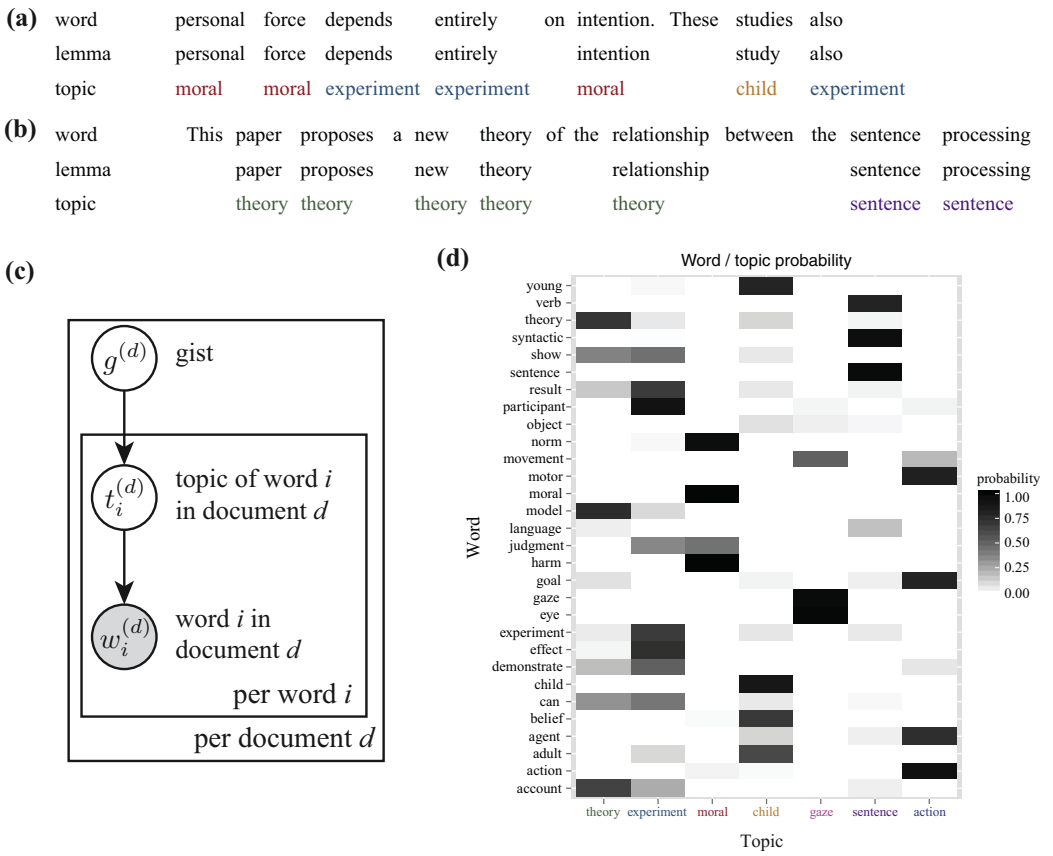

Fig. 1. Modeling word meaning using Topic Models. (a and b) Nine and twelve word excerpts from Greene et al. (2009) and Gibson (1998), respectively. The first row is the original text. The second row is the text after being processed through the lemmatizer. The third row is the topic assigned to each word by the model. (c) A graphical model representation of the Topic Model. Each node is a variable in the model and edges encode probabilistic dependencies between the two variables. Each rectangle is a called a "plate", where the variables and edges in the rectangle are copied multiple times (e.g., the inner rectangle represents each word in the current document and each word's topic). Shaded nodes denote that the variable was observed. (d) A word-topic matrix from our trained Topic Model limited to seven topics and their most probable words. Note how the gaze topic focuses on words related to eyetracking methods and that some words (e.g., movement) are probable in multiple topics (e.g., eye and action), capturing different senses of the same word.

see how topics, rather than words, trend over time. We followed their approach by analyzing the titles and abstracts of 34 years of *Cognition* articles to track trends in research topics. We discuss how topics such as *morality* surged, while others such as *developmental* remained stable. Additionally, we extend the work of Hall et al. (2008) in two ways. First, we propose a method for selecting which topics represent consistent trends, regardless of the number of topics used to build the model. Second, we show that interest may lie in what we label *framing topics*, topics that do not model any research domain, but contain words used to frame more contentful topics. Cognition's abstracts and titles support two framing topics: theory-centric and experimental.

2. Topic Models

Although there is no agreed upon representation of word meaning, Topic Models provide a relatively simple and practical method for exploring hypotheses about the meanings of words in documents. Topic Models assume that every word in every document is generated by one of a number of *topics* (see Fig. 1a and b for examples). Topics are (Dirichlet-distributed) mixtures of words (a topic specifies the probability of a word being produced by that topic), and documents are (Dirichlet-distributed) mixtures of topics. As illustrated in Fig. 1c, the model generates documents according to a hierarchical process. First, a mixture of topics (the *gist* of the document) is sampled from a Dirichlet distribution. Subsequently each word is sampled from the topics of the document. Documents are biased to be more likely to generate some topics rather than others, and topics are biased to be more likely to generate some words rather than others. Together, these biases lead the model when given a corpus of documents to converge[2] on solutions in which words that are likely to co-occur are generated by the same topic. For example, the "child"

[2] This is done using standard machine learning techniques, see Griffiths et al. (2007) for more details.