Creating dictionaries for argument identification by reference data

Andreas Niederle¹
aniederle@infomatik.uni-leipzig.de
Gregor Wiedemann³
gregor.wiedemann@uni-hamburg.de
Sebastian Dummi³
sebastian.dummi@hsuhh.de
Gerhard Heyer³
heyer@infomatik.uni-leipzig.de
¹Universität Leipzig, Germany; ²Helmut-Schmidt-Universität / Universität der Bundeswehr

Introduction

The creation of dictionaries is an important task to conceptualize and operationalize research questions in content analysis (Neuendorf, 2002). One can define concepts for coding operationalized variables in the form of mutually exclusive categories or decide if the content of documents is relevant for coding within the research task by the formalization of meaning through dictionaries (Krippendorff, 2004). Dictionaries are often defined on the basis of a “theory of meaning that reflects a research question or the vocabulary of an academic discipline” (Krippendorff, 2004). Thus, we can think of dictionaries as operationalized representations of historical, sociological, cultural or political theories that are investigated within humanities research. Procedures to create dictionaries automatically allow for more methodical and reproducible research designs especially when dealing with large corpora.

Approach

In contrast to manual dictionary creation from a small set of selected sample documents we present an approach to automatically extract dictionaries from arbitrary size. Within the “ePol-project” the goal of identifying arguments for a political science research task is approached by creation of one semantic dictionary on the utilization of topic models (Blei/Paisley/Jordan, 2003; Teh/Jordan, 2010) to identify thematically relevant documents; and one rather syntactical dictionary based on paradigmatically similar markers to identify a high density of argument structures. This poster presents ideas, results and an example extraction of dictionaries for referencing ranking of retrieved results in large document collections.

Semantic dictionaries

Domain experts easily can compile a small reference corpus of paradigmatic documents containing contents of their interest. On this basis we refer to our reference corpus we call a topic model based on the Pitman-Yor Process ( Teh, 2006). It employs Poisson instead of Dirichlet distributions which better approximate distributions of natural language (Teh et al., 2006). It employs Poisson instead of Dirichlet distributions which better approximate distributions of natural language (Teh et al., 2006). Domain experts easily can compile a small reference corpus of paradigmatic documents containing contents of their interest. On this basis we refer to our reference corpus we call a topic model based on the Pitman-Yor Process ( Teh, 2006). It employs Poisson instead of Dirichlet distributions which better approximate distributions of natural language (Teh et al., 2006).

In contrast to manual dictionary creation from a small set of selected sample documents we present an approach to automatically extract dictionaries from arbitrary size. Within the “ePol-project” the goal of identifying arguments for a political science research task is approached by creation of one semantic dictionary on the utilization of topic models (Blei/Paisley/Jordan, 2003; Teh/Jordan, 2010) to identify thematically relevant documents; and one rather syntactical dictionary based on paradigmatically similar markers to identify a high density of argument structures. This poster presents ideas, results and an example extraction of dictionaries for referencing ranking of retrieved results in large document collections.

Semantic dictionaries

Domain experts easily can compile a small reference corpus of paradigmatic documents containing contents of their interest. On this basis we refer to our reference corpus we call a topic model based on the Pitman-Yor Process ( Teh, 2006). It employs Poisson instead of Dirichlet distributions which better approximate distributions of natural language (Teh et al., 2006). Domain experts easily can compile a small reference corpus of paradigmatic documents containing contents of their interest. On this basis we refer to our reference corpus we call a topic model based on the Pitman-Yor Process ( Teh, 2006). It employs Poisson instead of Dirichlet distributions which better approximate distributions of natural language (Teh et al., 2006).

Synthetic dictionaries via term similarities

Additionally to our dictionary containing semantic information related to theoretical aspects of the political science research task we created a second dictionary of linguistic markers, called Argumentmarker, which can be employed to identify argumentative structures (Dummer/Tamke, 2013). We took a list of 46 German argument markers from another research project on causality and textual coherence (Breidler/Walther, 2009) as a starting point. This list was incrementally extended up to 127 terms by automatically computed synonyms of the markers retrieved from the database of the Project ‘Deutscher Wortschatz’ (Quasthoff/Eckert, 2009), a representative corpus of German language.

Application

We applied these dictionaries for retrieval of documents in a large collection of newspaper articles to identify argumentative tests with a certain没错looking framing. First, a subcorpus of 5 000 thematically relevant articles is retrieved with a specialized retrieval process (Wiedemann/Niklas) 2014) out of a corpus of 3.5 million articles using the semantic dictionary as query. These 10 000 documents then are ranked with the synthetic dictionary of Argumentmarker to retrieve documents with potentially high density of arguments. The best ranked N tests then are subject to a close reading process by political scientists who also utilize the dictionaries for qualitative coding schemes.

References