Disambiguating and Specifying Social Actors in Big Data:

Using Wikipedia as Source for Demographic Information

Philipp Poschmann & Dr. Jan Goldenstein

Friedrich-Schiller-Universität Jena
Lehrstuhl für ABWL/Organisation, Führung und Human Resource Management

Focus: Social actors

Studies benefit if they, for instance, investigate the

- presence, communication, and role of economic or cultural elites in society (Bourdieu 1984; Davis 2003),
- role of economic leaders' national or educational background (Carpenter 2002; Nielsen and Nielsen 2013),
- origin of offenders and victims of crime (Franzosi et al. 2012),
- kind of organizations participating in public discourse about contested issues (Meyer and Höllerer 2010),
- competition across organizations in emerging markets (Kennedy 2008),
- inter-organizational linkages (Taylor and Helfat 2009),
- role of organizations in public scandals (Campbell 2007; Franzosi et al. 2012).

on a large scale.

The richness of big data cannot be fully exploited in studies that focus on the investigation of social actors

- Ongoing trend in the **social sciences** (e.g., Evans and Aceves 2016; George et al. 2016) → Computational Social Science, Digital Humanities
- **Social actors:** Named entity recognition (van Atteveldt, Kleinnijenhuis, and Ruigrok 2008; Evans and Aceves 2016; Mohr et al. 2013)
  - *Detection: What social actors appear in texts?*
  - But: Coarse-grained categories: 'Person' and 'Organization'
- **Challenge:**
  - *Disambiguation:* Who is a specific 'Person' or 'Organization'?
  - *Specification:* Demographic information are often indispensable to examine differential pattern of agency in social settings (Franzosi 1990; Sudhahar et al. 2013; Cornelissen & Werner, 2014; McCormick et al., 2015)
    - Age
    - Gender
    - Type of organization (NGO, corporation, political party)
    - Industry
    - Alma Mater
    - Job
    - …

Illustrating the challenge

Many brokers would like to see Robert J. McCann, an amiable executive who heads the brokerage division, take on a broader responsibility. And while Gregory J. Fleming, a co-president, has been hurt somewhat by his role in the Wachovia merger approach, his stature as one of the firm's top investment bankers and his close relationship with Boston Consulting, the leading contender for the job, make him likely to stay.

(New York Times)

Illustrating the challenge: Named entity recognition (*detection*)

Many brokers would like to see **Robert J. McCann**, an amiable executive who heads the brokerage division, take on a broader responsibility. And while **Gregory J. Fleming**, a co-president, has been hurt somewhat by his role in the **Wachovia** merger approach, his stature as one of the firm's top investment bankers and his close relationship with **Boston Consulting**, the leading contender for the job, make him likely to stay.

Person name

Organization name

Organization name

Person name

Now what?

Illustrating the challenge: Named entity recognition (*detection*)

Robert J. McCann

Chairman Americas of UBS Group AG

In office
January 2016 – present

Personal details

Born	Robert J. McCann March 15, 1958 (age 59) Pittsburgh, Pennsylvania USA
Nationality	American and Irish citizen
Alma mater	Bethany College Texas Christian University
Occupation	Financial Services

WIKIPEDIA
Die freie Enzyklopädie

Wachovia

WACHOVIA

Former type	Financial Services
Fate	Acquired by Wells Fargo
Founded	June 16, 1879; 137 years ago
Defunct	2008 (as an independent corporation) 2011 (as a brand)
Headquarters	Charlotte, North Carolina
Products	Banking, Investments
Owner	Wells Fargo
Website	Archived official website at the Wayback Machine (archive index)

Short description: Wikipedia as Data Source for the Social Sciences

- Especially the online encyclopedia Wikipedia serves as a global place of **collective memory** (Ferron and Massa 2014; Pentzold 2009)
- Unexploited access to a large amount of accurately prepared data (Badke 2008; Brown 2011; Fallis 2008; Javanmardi and Lopes 2010)
  - **demographic information about <u>persons</u> and <u>organizations</u>** that can be used to enhance the study of social agency (e.g., Cornelissen and Werner 2014; Franzosi, de Fazio, and Vicari 2012)
  - Social agency = What social actors (who?) do or what is done to them and how this is embedded (narrative, frame) into a given social setting

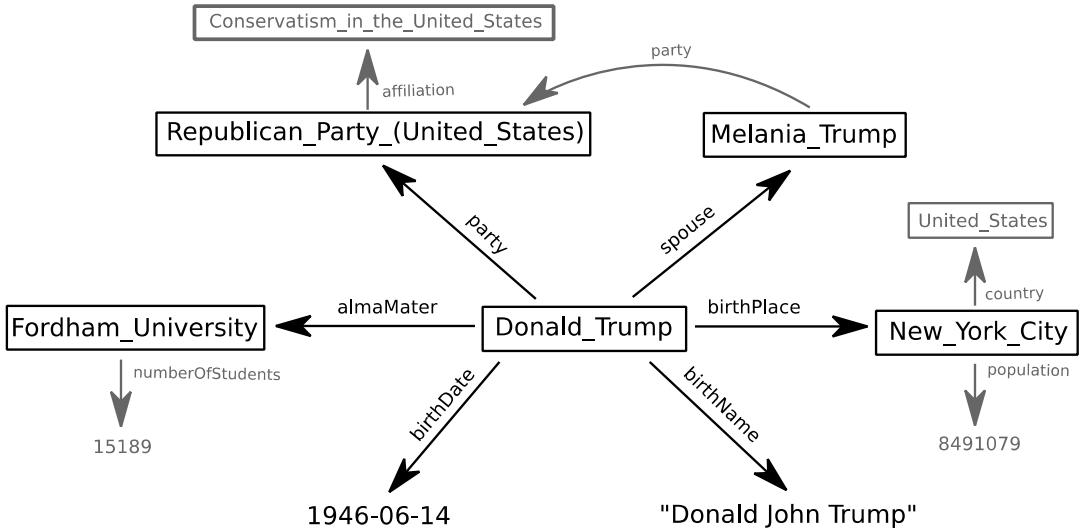
Case study:
Using Wikipedia to Disambiguate and Specify Social Actors in News Coverage about the US-Presidential Election 2016

- **US-newspaper articles** on the US presidential election 2016 (published between 01 January 2015 and 31 December 2016)
- Specification of social actors is particularly important when events are investigated in which many different actors are involved
- **Nexis database:**
  - ‘Major US-newspapers’ (articles published by US-newspapers that are ranked in the top 50 in Editor & Publisher Year Book regarding their circulation)
  - Topics: ‘US presidential election’ in combination with ‘United Sates’
  - Total: 36,117 newspaper articles

Case study:
Accessing Wikipedia through DBpedia

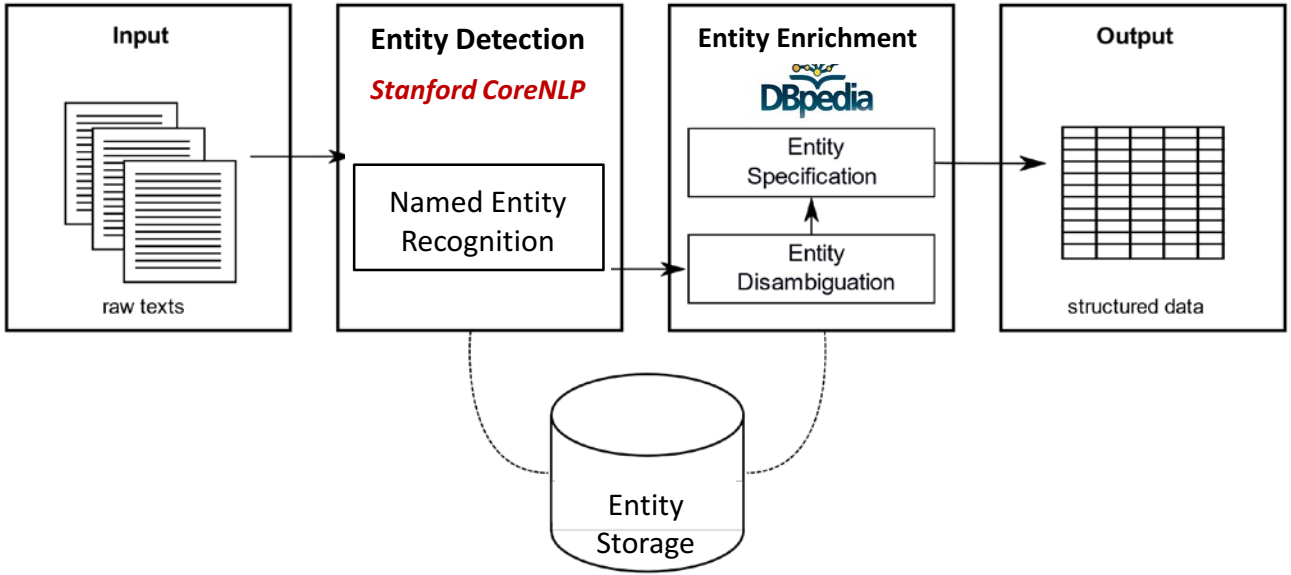
DBpedia

http://dbpedia.org/resource/Donald_Trump

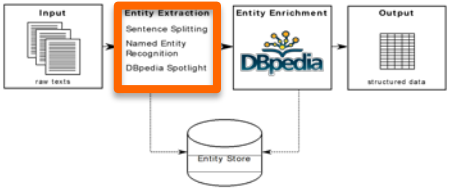


Our Architecture

Main purpose: (1) **detecting**, (2) **disambiguating** and (3) **specifying** social actors



Entity Detection



Named Entity Recognition (*Detection*)

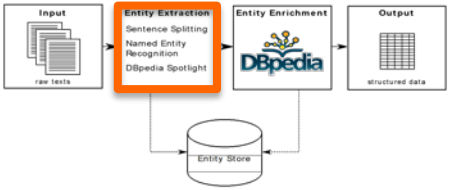
"**Donald Trump** is leading **Hillary Clinton** by four points in the battleground state of **Florida**, according to **Siena College** poll released Sunday."

"**Obama** will attend **Democratic Senatorial Campaign Committee** and **Democratic National Committee** events in **Miami** on **Friday**."

Donald Trump
Hillary Clinton
Florida
Siena College

Democratic Senatorial Campaign Committee
Democratic National Committee
Miami
Obama
Friday

Entity Detection



DBpedia Spotlight (*Disambiguation*)

"**Donald Trump** is leading **Hillary Clinton** by four points in the battleground state of **Florida**, according to **Siena College** poll released Sunday."

"**Obama** will attend **Democratic Senatorial Campaign Committee** and **Democratic National Committee** events in **Miami** on **Friday**."

DBpedia Spotlight

Donald Trump:
http://dbpedia.org/resource/Donald_Trump

Hillary Clinton:
http://dbpedia.org/resource/Hillary_Rodham_Clinton

Siena College:
http://dbpedia.org/resource/Siena_College

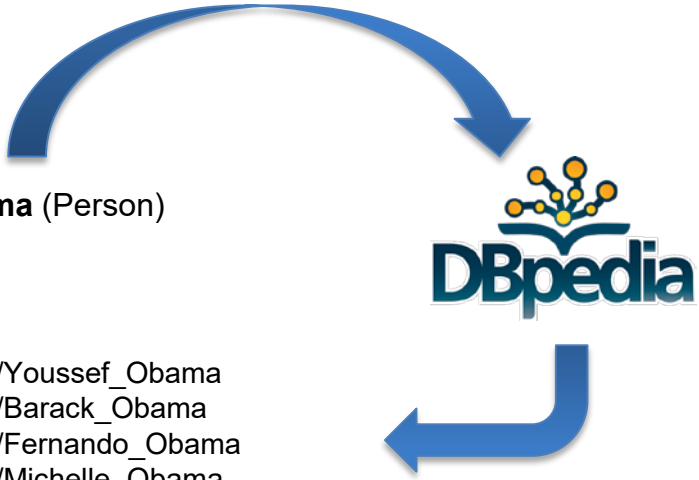
Democratic Senatorial Campaign Committee:
http://dbpedia.org/resource/Democratic_Senatorial_Campaign_Committee

Democratic National Committee:
http://dbpedia.org/resource/Democratic_National_Committee

Obama?

Entity Disambiguation

Step 1: Search for missing entity URIs



http://dbpedia.org/resource/Youssef_Obama
http://dbpedia.org/resource/Barack_Obama
http://dbpedia.org/resource/Fernando_Obama
http://dbpedia.org/resource/Michelle_Obama

Compare persons‘ abstract with context…

“Obama to visit Miami. **President** Obama plans to take a fundraising swing through Miami on Friday, the White House said. *Obama will attend* ***Democratic*** *Senatorial Campaign Committee and* ***Democratic National*** *Committee events in Miami on Friday.* He'll remain in Florida through Sunday and then return to **Washington, D.C.** This will not have been his last **election campaign** trip across the **United States**.”

http://dbpedia.org/resource/Youssef_Obama

"Youssef **Obama** (Arabic: يوسف اوباما; born 26 May 1994) is an Egyptian footballer who plays for Zamalek SC. He is an Attacking Midfielder. Obama has a great skill in dribbling, scoring passing and shooting with both feet."

http://dbpedia.org/resource/Fernando_Obama

"Fernando **Obama** Galindo (born 7 April 1985 in Murcia, Spain) is an Equatoguinean footballer. Mainly a central defender, he can also operate on the flanks."

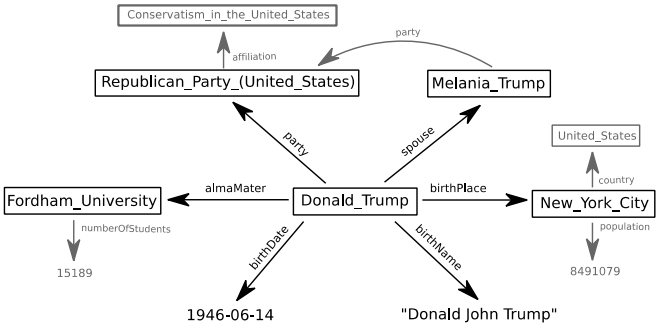
http://dbpedia.org/resource/Barack_Obama

„Barack Hussein **Obama** II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/; born August 4, 1961) is an American politician serving as the 44th **President** of the **United States**, the first African American to hold the office. [...] He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, and ran unsuccessfully in the **Democratic** primary for the **United States** House of Representatives in 2000 against incumbent Bobby Rush. In 2004, Obama received **national** attention during his **campaign** to represent Illinois in the **United States** Senate with his victory in the March **Democratic** Party primary, his keynote address at the **Democratic National** Convention in July, and his **election** to the Senate in November.[...]. He currently resides in **Washington, D.C.** “

Entity Enrichment

Step 2: Get Data for Entity URLs (*Specification*)

http://dbpedia.org/resource/Donald_Trump



DBpedia

Output: Bring order into data!

Case Study:
Demographic information about actors in public discourse

Persons					
Gender	Male Female	82.13 % 17.87 %	Educational Background	Harvard University University of California Princeton University	3.93 % 1.33 % 1.21 %
Age	31-50 51-60 61-80	13.69 % 21.81 % 45.92 %	Religion	Catholic Church Episcopal Church Judaism	24.55 % 7.18 % 6.53 %
Party	Democratic Party Republican Party Independent politician	35.86 % 31.72 % 1.83 %	Position	politician office holder congressman senator Governor mayor	20.68 % 67.17 % 5.17 % 2.57 % 2.46 % 0.87 %
Birth state	New York California Pennsylvania	6.82 % 3.26 % 3.21 %			

Case Study:
Demographic information about actors in public discourse

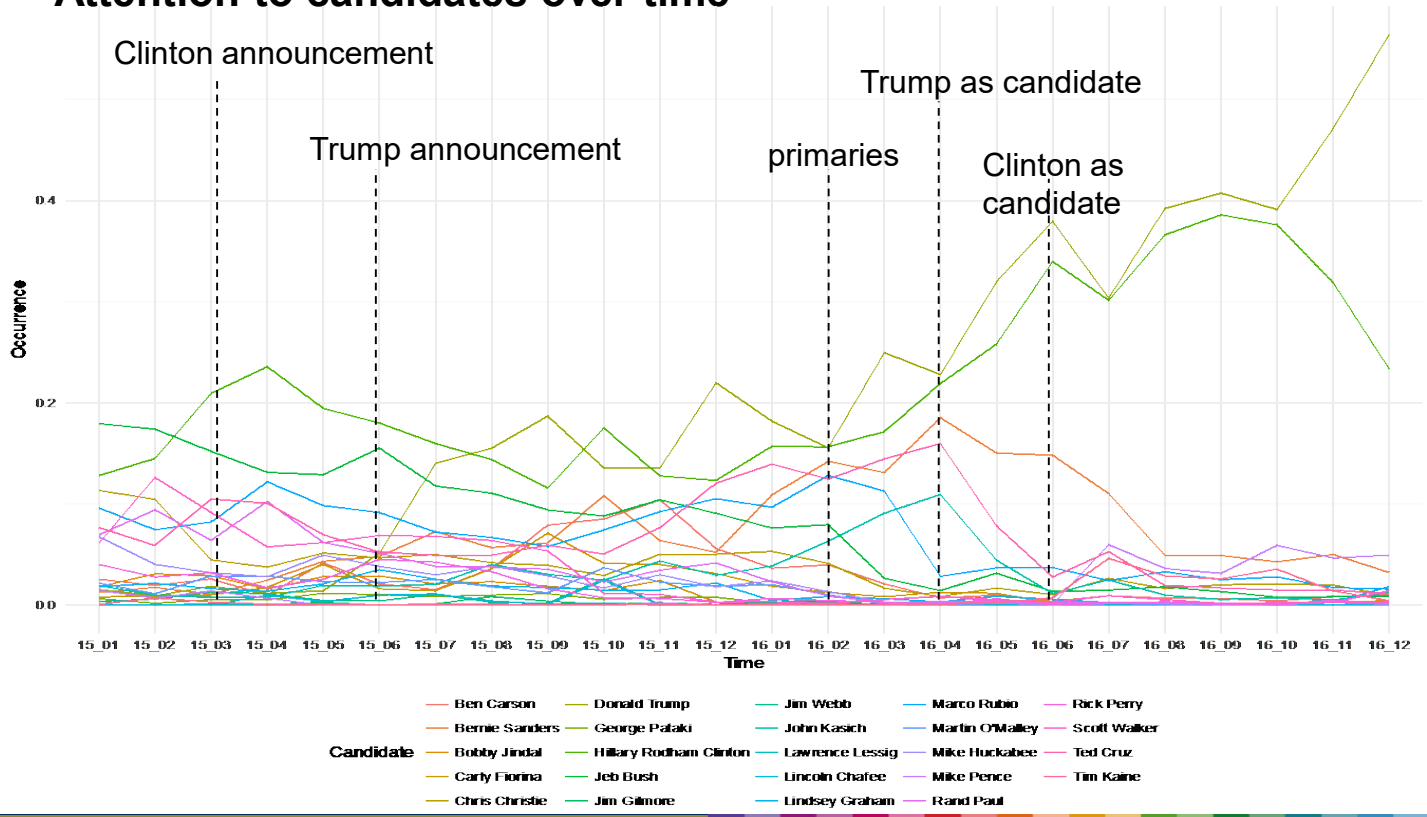
Companies					
Foundation	before 1900 before 1945 after 1945	17.37 % 22.90 % 59.74 %	Industries	Financial services Retail Aviation	5.67 % 3.17 % 3.12 %
Employees	< 1,000 1,001 – 10,000 > 10,000	25.86 % 48.28 % 25.86 %	Headquarter	New York California Florida	8.01 % 5.97 % 1.97 %
Educational Institutions					
Foundation	before 1900 before 1945 after 1945	19.55 % 24.58 % 55.87 %	State	Pennsylvania California New York	9.27 % 8.56 % 8.38 %
Students	< 10,000 10,000 – 50,000 > 50,000	68.01 % 27.45 % 4.54 %	Type	Public Private	42.67 % 57.33 %
Non-profit Organizations					
Foundation	before 1900 before 1945 after 1945	1.79 % 17.86 % 80.36 %	Location	New York City Washington, D.C. United Kingdom	11.34 % 9.28 % 4.12 %

Case Study:
Reliability and Accuracy

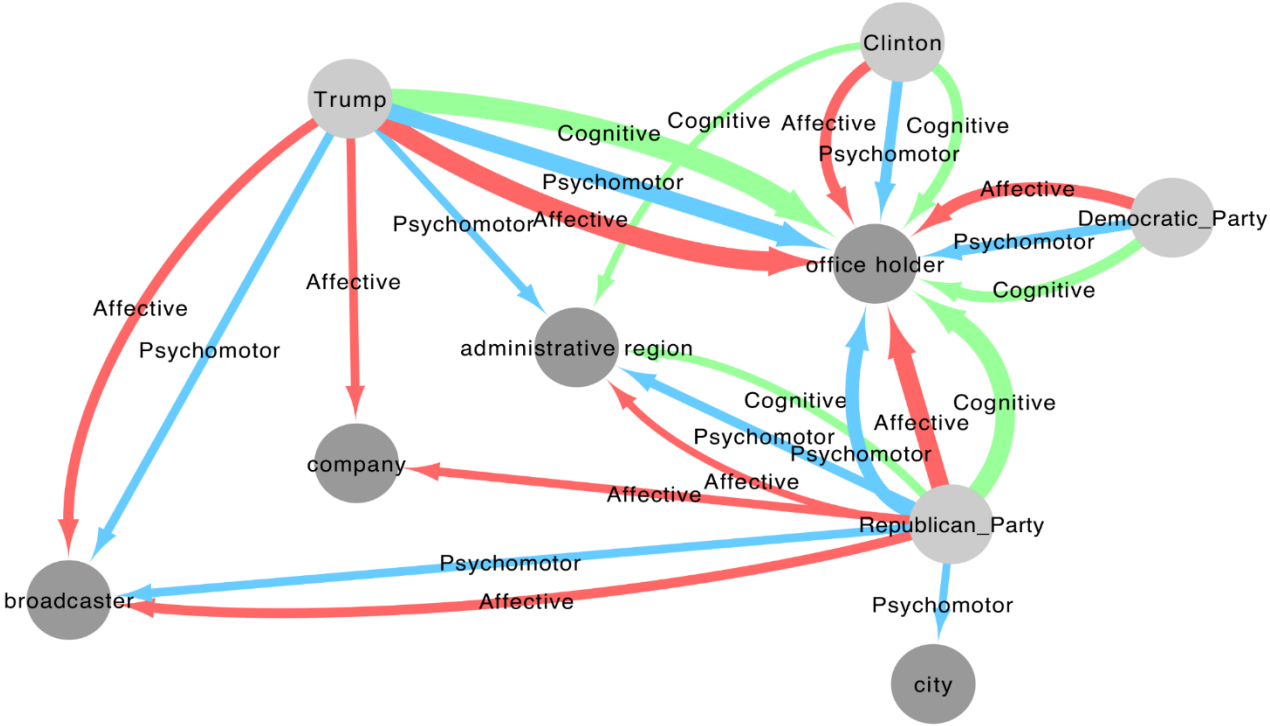
Entity detection			Entity disambiguation
Precision	Recall	F-measure	Accuracy
93.85 %	90.85 %	92.32 %	90.60 %

Further applications

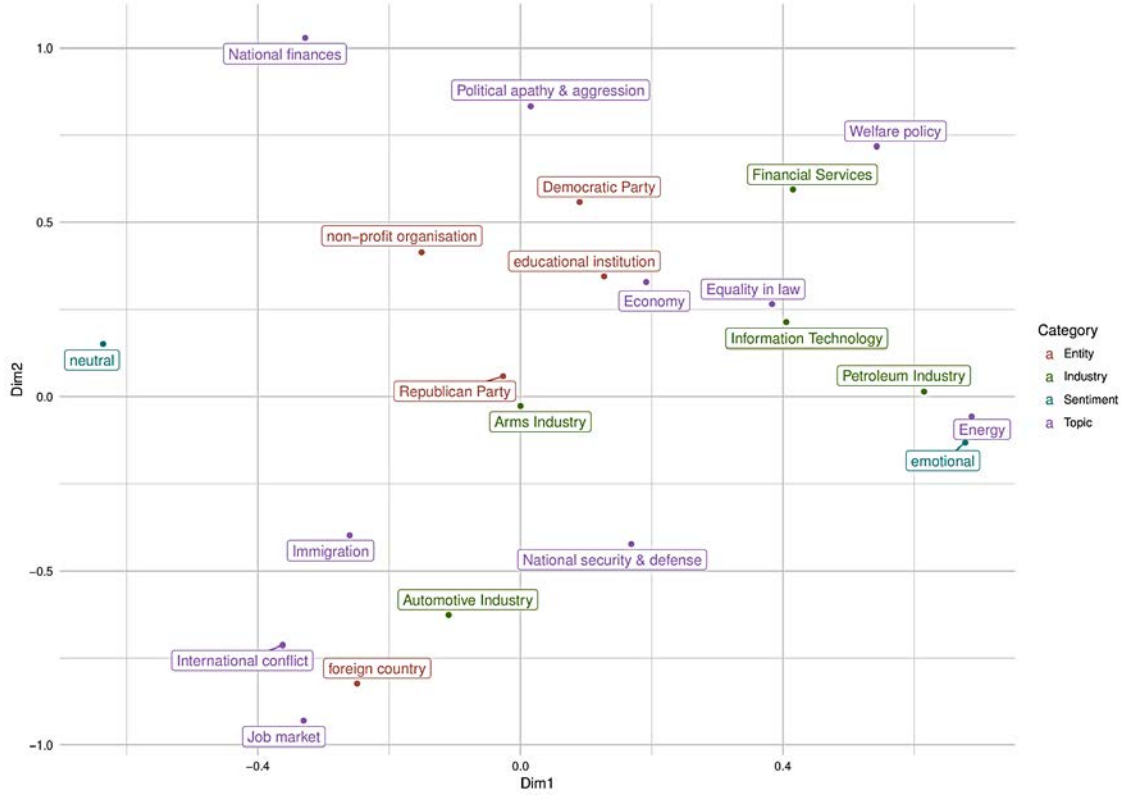
Visualization:
Attention to candidates over time



Visualization:
Who acts how? Actors-Actions-Objects



Visualization:
Topics, Sentiments, Entities



Discussion

- Adding demographic information automatically from Wikipedia opens the opportunity for social scientists to measure social agency on a larger scale in terms of analyzing **who is acting**.

Advances of Using Wikipedia as Source to Disambiguate and Specify Social Actors

- Availability: 1,517,816 Persons & 275,077 Organizations
- Our software application automatically provides demographic information that cannot be found in texts directly
- Structured demographic information: Semantic narrative analysis (Franzosi, 1989, 1990), social network analysis (Mohr et al., 2013), parsing approaches (Evans & Aceves, 2016).

Discussion

Limitations of Using Wikipedia as Source for Demographic Information

- Wikipedia only contains social actors for which verifiable knowledge exists
- English Wikipedia has a bias towards actors that are foremost known in the Anglo-American language region
- Only up-to-date information

Disambiguating and Specifying Social Actors in Big Data:

Using Wikipedia as Source for Demographic Information

Philipp Poschmann & Dr. Jan Goldenstein

Friedrich-Schiller-Universität Jena
Lehrstuhl für ABWL/Organisation, Führung und Human Resource Management