ANALYSING ENTITY CONTEXT IN MULTILINGUAL WIKIPEDIA TO SUPPORT ENTITY-CENTRIC RETRIEVAL APPLICATIONS

Yiwei Zhou, Elena Demidova and Alexandra I. Cristea

September 9, 2015

University of Warwick, Coventry, UK
L3S Research Center and Leibniz Universität Hannover, Germany

LANGUAGE-SPECIFIC REPRESENTATIONS OF FAMOUS ENTITY

Various representations of the same entity under various language cultures — language-specific entity aspects

Angela Merkel related aspects in

- English context: Barack Obama, David Cameron, Greek financial situation ...
- German context: domestic political topics, featuring discussions of political parties in Germany, scandals arising around German politicians, local elections ...

OVERVIEW OF THIS PAPER

Objective

To obtain a comprehensive overview over the language-specific entity aspects and their representations in different languages.

Knowledge Base

Multilingual Wikipedia: comprehensive entities' representations, useful manually-defined linking structure

Pipeline

Context Definition, Context Extraction, Similarity Analysis

CONTEXT DEFINITION

CONTEXT DEFINITION

Context Definition: The context $C(e, L_i)$ of the entity e in the language L_i is represented through the set of aspects $\{a_1, \ldots, a_n\}$ of e in L_i, weighted to reflect the relevance of the aspects in the context: $C(e, L_i) = (w_1 * a_1, \ldots, w_n * a_n)$.

Aspects: noun phrases that co-occur with the entity in a given language.

Weights: $w(a_k, e, L_i) = af(a_k, e, L_i) \cdot \log \frac{N}{af(a_k, e, L)}$
af: language-specific aspect co-occurrence frequency.

CONTEXT EXTRACTION

BASELINE: ARTICLE-BASED CONTEXT EXTRACTION

Sources of context: All sentences from an article representing the entity in a language edition.

Drawbacks: Incompleteness.
e.g. “Economic Council Germany” page: “Although the organisation is both financially and ideologically independent it has traditionally had close ties to the free-market liberal wing of the conservative Christian Democratic Union (CDU) of Chancellor Angela Merkel.”.
“The nightmare (painting)” page: “On 7 November 2011 Steve Bell produced a cartoon with Angela Merkel as the sleeper and Silvio Berlusconi as the monster.”

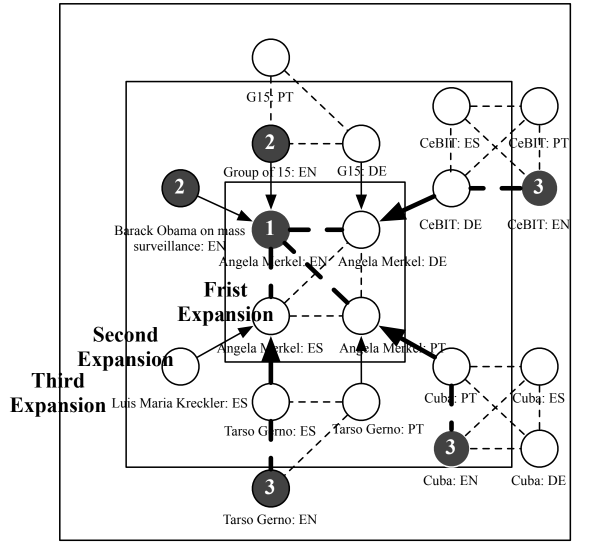
GRAPH-BASED CONTEXT EXTRACTION

Sources of context: “The whole Wikipedia.”

Basic Idea:

- More comprehensive: Graph Creation.
  Use the in-links to the main Wikipedia article describing the entity and the language-links of these articles to efficiently collect the articles that are probable to mention the target entity in different language editions;
- More precise: Context Construction.
  Extract the sentences mentioning the target entity using named entity disambiguation tool (DBpedia Spotlight).

GRAPH-BASED CONTEXT EXTRACTION



SIMILARITY ANALYSIS

SIMILARITY ANALYSIS

Similarity Measure

$\mathrm{Sim}(C(e, L_i), C(e, L_j)) = \frac{C(e,L_i) \cdot C(e,L_j)}{|C(e,L_i)| \times |C(e,L_j)|}$
$C(e, L_i)$: context of entity e in language L_i

Dataset

80 entities with world-wide influence evenly come from four categories: politicians, international corporations, celebrities, sport stars.
Five European languages: English, German, Spanish, Portuguese and Dutch. Depend on the performance of Google Translate.
Article-based: 50 sentences per entity per language.
Graph-based: 1000.

SIMILARITY ANALYSIS

Table: Article-based cross-lingual similarity

Entity	EN-DE	EN-ES	EN-PT	EN-NL	DE-ES	DE-NL	ES-PT
GlaxoSmithKline	0.43	0.34	0.29	0.29	0.31	0.22	0.26
Angela Merkel	0.68	0.66	0.84	0.54	0.60	0.59	0.66
Shakira	0.71	0.58	0.84	0.75	0.48	0.64	0.58
Lionel Messi	0.71	0.86	0.81	0.89	0.71	0.68	0.82
Average of 80	0.50	0.47	0.46	0.43	0.38	0.36	0.39
Stdev of 80	0.16	0.20	0.23	0.22	0.18	0.19	0.22

SIMILARITY ANALYSIS

Table: Graph-based cross-lingual similarity

Entity	EN-DE	EN-ES	EN-PT	EN-NL	DE-ES	DE-NL	ES-PT
GlaxoSmithKline	0.72	0.73	0.59	0.61	0.63	0.62	0.55
Angela Merkel	0.64	0.62	0.42	0.60	0.75	0.82	0.51
Shakira	0.91	0.94	0.90	0.88	0.94	0.91	0.94
Lionel Messi	0.63	0.76	0.77	0.68	0.70	0.62	0.76
Average of 80	0.53	0.60	0.56	0.52	0.53	0.48	0.61
Stdev of 80	0.25	0.22	0.21	0.24	0.24	0.25	0.20

SIMILARITY ANALYSIS

Table: Top-30 highly weighted aspects of “Angela Merkel” (graph-based)

English	angela merkel, battle, berlin, cdu, chancellor, chancellor angela merkel, church, edit, election, emperor, empire, england, france, george, german, german chancellor angela merkel, germany, government, jesus, john, kingdom, merkel, minister, party, president, talk, union, university, utc, war
German	academy, angela merkel, article, berlin, cdu, cet, chancellor, chancellor angela merkel, csu, election, example, german, german chancellor angela merkel, german children, germany, government, kasner, merkel, minister, november, october, office, party, president, propaganda, ribbon, september, speech, time, utc
Portuguese	ali, angela merkel, bank, cdu, ceo, chairman, chancellor, chancellor angela merkel, china, co-founder, coalition, csu, dilma rousseff, german chancellor angela merkel, germany, government, government merkel, koch, leader, merkel, minister, november, october, party, petroleum, president, saudi arabia, state, union, york

CONCLUSION

CONCLUSION

- The editors of different Wikipedia language editions describe some common entity aspects, they can have different focus with respect to the aspects of interest.
- The graph-based method is a promising approach to obtain a comprehensive overview of the language-specific entity representation.
- The language-specific entity representation could be used in targeted retrieval of entity-centric information in a specific language context.

Thank you & Questions?