OPI

OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Web Search Result Clustering with BabelNet

Luxembourg 03.03.16

Introduction

- The goal is to verify how Babelnet/Babelfy can improve the quality of the search result clustering.
- We present well-known search result clustering method enriched with BabelNet information.
- Three level experiments:
  - Comparison of three search results clustering methods
  - Quantifying the impact of BabelNet/Babelfy on search result clustering algorithm
  - Verification of the idea of clustering snippets without the specialized algorithm, namely only with the use of BabelNet/Babelfy systems.
- Evaluation on the dataset AMBIENT using four distinct measures, namely: Rand Index (RI), Adjusted Rand Index (ARI), Jaccard Index (JI) and F1 measure

Related Work

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

- Search Results Clustering (SRC) is a specific area of documents clustering
- Contextual descriptions (snippets) of documents returned by a search engine are short, often incomplete, and highly biased toward the query, so establishing a notion of proximity between documents is a challenging task.
- Approaches to search result clustering can be classified as data-centric or description-centric
- Data-centric – Bisecting K-means, HAC
- Description-centric – STC, Lingo, KeySRC

Babel interfaces consumption

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

- Babelfy – text disambiguation
  - https://babelfy.io/v1/disambiguate
- BabelNet – get categories and glosses for the given synset
  - https://babelnet.io/v3/getSynset
- BabelNet – get hypernyms for the given synset
  - https://babelnet.io/v3/getEdges

First experiment

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Algorithm	RI	ARI	JI	F1
Lingo	62.52	18.09	30.76	49.01
STC	66.95	23.05	28.10	53.08
K-means	62.79	7.69	12.83	49.79

Second experiment

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Improvement	RI	ARI	JI	F1
Lingo	62.52	18.09	30.76	49.01
synsets+	**63.52**	**18.61**	29.21	**49.76**
categories+	**63.04**	17.01	27.46	**49.36**
categories+1	61.73	16.48	29.55	48.65
categories+2	62.17	17.44	30.30	48.80
glosses+	**62.69**	12.27	21.30	47.24
hypernyms+	61.52	16.35	29.44	48.32

Third experiment

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Approach	RI	ARI	JI	F1
Lingo	62.52	18.09	30.76	49.01
babelCl1	50.60	1.67	26.87	41.53
babelCl2	50.44	1.56	27.06	40.41

Conclusions

- We introduced new semantic features from BabelNet/Babelfy (as disambiguated synsets, categories/glosses describing synsets, or semantic edges) in order to verify how they influence on the clustering quality of the representative search result clustering algorithm.
- The best improvements concern synsets expansions and are relatively weak (what seems to be suprising outcome?).
- Snippets clustering, only with the use of Babelfy/Babelnet interfaces, report the results still below the Lingo measures.
- Problem connected with the time performance issues.
- In the future we plan to introduce more sophisticated improvements based on graph theories, because there must be a way to drastically improve the quality measures consuming such well-defined and organized semantic network as BabelNet.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Dziękujemy za uwagę

Merci

Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy
al. Niepodległości 188B, 00-608 Warszawa
tel. 22 570 14 00
www.opi.org.pl