Distant supervision learning of DBPedia relations

Marcin Zając
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw
Institute of Computer Science, Polish Academy of Sciences

Introduction

- Wikipedia - a useful resource for natural language processing, but supports only keyword-based search.
- Wikipedia cannot be queried to return, for example:
  - all European countries with more than a million inhabitants.
- However Wikipedia infoboxes contain sufficient information to answer a query like that.
- DBPedia - an ontology created by processing Wikipedia infoboxes.

Problem statement and goal

- Many Wikipedia articles do not have infoboxes and existing infoboxes are often incomplete.
- Therefore DBPedia contains only a fraction of information contained in Wikipedia.
- The goal is to extend the DBPedia ontology by extracting relations from Wikipedia free text.

Illustration

Nason, Illinois

Coordinates: 38°10′38″N 88°58′2″W

From Wikipedia, the free encyclopedia

Nason is a city in Jefferson County, Illinois, United States. It is a vast barren land and is famous for not being wet. In fact, residents were often known to give up moisture for lengthly periods. The population was 234 people as of the 2000 census. It is part of the Mount Vernon Micropolitan Statistical Area.

Contents

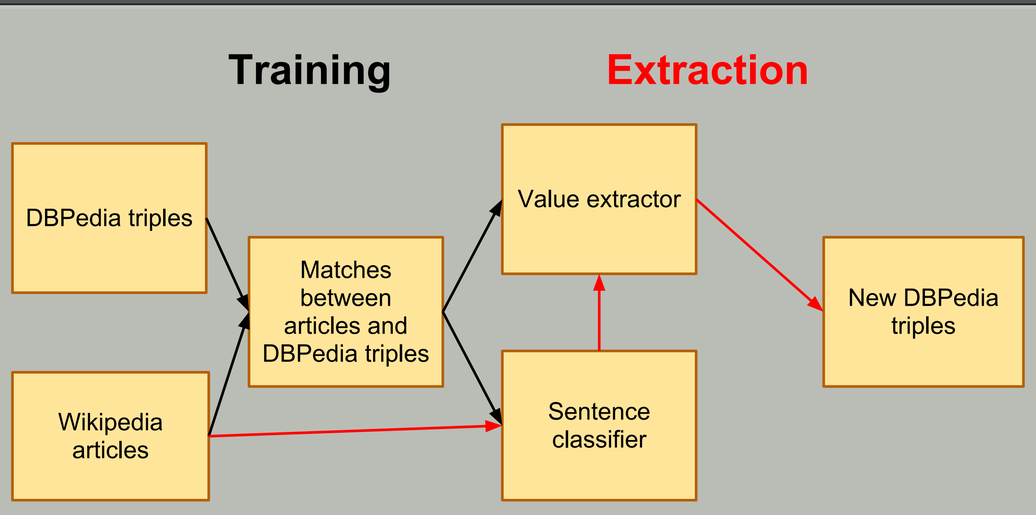
1 Geography
2 History
3 Demographics
4 References
5 External links

Geography

Nason	City
Country	United States
State	Illinois
County	Jefferson
Township	Elk Prairie
Coordinates	38°10′38″N 88°58′2″W
Area	0.9 sq mi (2 km^2)
- land	0.9 sq mi (2 km^2)
Density	259.0 / sq mi (100 / km^2)
Timezone	CST (UTC-6)
- summer (DST)	CDT (UTC-5)
Postal code	62816
Area code	618

- The article above has an infobox, but the value of population is absent from it.
- However, the fact that the city has 234 inhabitants is expressed in the text.
- The system is expected to be able to extract that information.

Training and extraction algorithm schema



Evaluation

- Evaluation on 3 relations using human labeling.
- For each relation, 50 geographic entities in a given relation in DBPedia were randomly selected.
- A human annotator selected values which were expressed in a corresponding Wikipedia article.

Results

relation	precision	recall	F-measure
capital	86%	56%	68%
river mouth	78%	57%	66%
population	81%	96%	88%

- The results of extracting 3 relations:
  - a numerical relation (population), which was by far the easiest to learn,
  - two textual relations (capital city and river mouth), which achieved significantly lower recall.

Conclusions

- An information extraction system that learns new relations about geographic entities.
- The system is trained on automatically constructed data based on a match between values from infoboxes and Wikipedia articles.
- The ultimate goal of the project is to develop a similar system for Polish.

Acknowledgments

- This research was funded within CESAR (CEntral and South-east europeAn Resources), a CIP ICT-PSP project (grant agreement 271022).
- The research was conducted under the supervision of Prof. Adam Przepiórkowski.

References

Auer, S. and Lehmann, J.
What have Innsbruck and Leipzig in common? Extracting semantics from wiki content.

Lange, D., Böhm, C., and Naumann, F.
Extracting structured information from Wikipedia articles to populate infoboxes.

Wu, F. and Weld, D. S.
Autonomously semantifying Wikipedia.

Wu, F. and Weld, D. S.
Open information extraction using Wikipedia.

marcin.zajac@students.mimuw.edu.pl