Unsupervised Information Extraction using BabelNet and DBpedia

Amir H. Jadidinejad

Islamic Azad University, Qazvin Branch,
Qazvin, Iran.
amir@jadidi.info

Abstract. Using linked data in real world applications is a hot topic in the field of Information Retrieval. In this paper we leveraged two valuable knowledge bases in the task of information extraction. BabelNet is used to automatically recognize and disambiguate concepts in a piece of unstructured text. After extracting all possible concepts, DBpedia is leveraged to reason about the type of each concept using SPARQL.

Keywords: Concept Extraction, Linked Data, BabelNet, DBpedia, SPARQL.

1 BABELNET

BabelNet[1] is a multilingual lexicalized semantic network and ontology. It was automatically created by linking the largest multilingual Web encyclopedia – i.e. Wikipedia[1] – to the most popular computational lexicon of the English language – i.e. WordNet[2]. It contains an API for programmatic access of 5.5 million concepts and a multilingual knowledge-rich Word Sense Disambiguation (WSD) [3]. With the aid of this API, we can extract all possible concepts in a piece of text. These concepts are linked to DBpedia, one of the more famous parts of the Linked Data project.

2 DBPEDIA

DBpedia[4] is a project aiming to extract structured content from the information created as part of the Wikipedia project. This structured information is made available on Semantic Web formats. DBpedia allows users to query relationships and properties associated with Wikipedia concepts. In this paper we used SPARQL to query DBpedia. It's possible to reason about the type of each concept (PER, LOC, ORG, MISC) with the aid of a classic deductive reasoning using classes and subclasses. For example, "Settlement" is defined as a subclass of "Place" (although maybe not directly). That means that all Things that are "Settlements" are also "Places". "Tehran" is a "Settlement", so it is also a "Place". Using the following query:

[1] http://www.wikipedia.org

Making Sense of Microposts Workshop @ WWW'13, May 13th 2013, Rio de Janeiro, Brazil

```
ASK {
  {
    ?thing a ?p .
    ?p rdfs:subClassOf dbpedia-owl:Place OPTION (transi-
tive).
  }
 UNION
  {
    ?thing a dbpedia-owl:Place .
  }
}
```

It's possible to reason about the type of every "`?thing`" such as: `http://dbpedia.org/resource/Tehran`. A similar query is used for LOCATION and ORGANIZATION.

3 IMPLEMENTATION DETAILS

Our proposed solution shows in Figure 1. The input text is passed to "Text2Concept" module. This module is used "BabelNet" and "Knowledge-rich WSD" algorithm to recognize a list of concepts. Finally, "Text Reasoner" module reason about the type of each concept with the aid of DBpedia using a simple deductive reasoning.

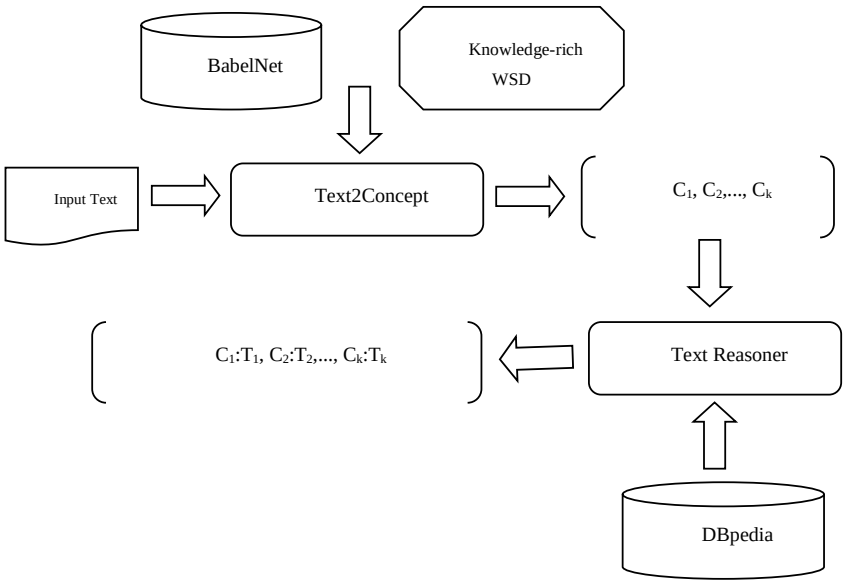

Fig. 1. Different parts of the proposed method.

Table 1 shows the impact of the proposed solution in on the training data set. Our proposed solution is achieved $F_1 = 0.50$ on training set and $F_1 = 0.52$ on testing set (See Fig. 2).

Table 1. Concept Extraction using the proposed solution on training data set

Data Set	Precision	Recall	F1
Train	0.5099	0.5003	0.5050

Rank	Best Run per Submission	PER	ORG	LOC	MISC	ALL
1	submission_14_1	**0.92**	**0.64**	0.74	0.38	**0.67**
2	submission_21_3	0.91	0.61	0.72	**0.41**	0.66
3	submission_15_3	0.92	0.57	**0.79**	0.36	0.66
4	submission_20_1	0.83	0.61	0.62	0.38	0.61
5	submission_25_1	0.83	0.49	0.74	0.30	0.59
6	submission_03_3	0.87	0.56	0.74	0.19	0.59
7	submission_29_1	0.76	0.54	0.59	0.36	0.56
8	submission_28_1	0.81	0.41	0.71	0.24	0.54
9	submission_32_1	0.73	0.35	0.59	0.41	0.52
10	submission_30_1	0.71	0.38	0.58	0.31	0.49
11	submission_33_3	0.85	0.37	0.62	0.14	0.49
12	submission_35_1	0.82	0.42	0.60	0.12	0.49
13	submission_23_1	0.83	0.52	0.50	0.04	0.47
14	submission_34_1	0.54	0.37	0.53	0.16	0.40

Fig. 2. Overall results between different participants.

4 REFERENCES

[1] Navigli, R., Ponzetto, S. P. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, 217-250.

[2] George A. Miller (1995). WordNet: A Lexical Database for English. *Communications of the ACM*. 38, 11, 39-41.

[3] Navigli, R., Ponzetto, S. P. 2012. Multilingual WSD with Just a Few Lines of Code: the BabelNet API. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics* (ACL 2012), Jeju, Korea, 67-72.

[4] Bizer, C., Lehmann, J., Kobilarov, G., Becker, C., Cyganiak, R., Hellmann, C. 2009. DBpedia – A Crystallization Point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 7, 154–165.