DBPedia-extender manual

Marcin Zając
Institute of Computer Science, Polish Academy of Sciences

23 stycznia 2013

1 About

The DBPediaExtender is an information extraction system that extends an existing ontology of geographical entities by extracting information from text. The system uses distant supervision – training data is constructed based on matches between values a knowledge base (taken from DBPedia) and Wikipedia articles.

2 Licensing

DBPediaExtender is released under the GNU General Public License v3.

3 Prerequisites

Software (versions tested):

- Python 2.6 or 2.7
- scikit-learn 0.11
- OpenLink Virtuoso (Open-Source Edition) 6.1.4
- crfsuite 0.12
- pantera-tagger 0.9

Data:

- DBPedia dumps
- Wikipedia dumps

4 Installation

DBPedia-extender is a pure python system, therefore it doesn't require any installation. However it relies on availability of two resources: a DBPedia server and Wikipedia articles.

4.1 DBPedia

The program queries a given DBPedia server endpoint over http. It can use any endpoint, however in practice, it is necessary to create a local DBPedia server.
The process of installing and configuring the server is well described at apohllo.pl/blog/virtuoso-installation-in-debian.
The Polish DBPedia dumps are not available at the Polish DBPedia website (pl.dbpedia.org). However, the Polish DBPedia maintainer (the Knowledge Hives – knowledgehives.com) is willing to give access to the data.

4.2 Wikipedia

When run for the first time the program will download the pages-articles archive and then extract articles from it. This task may take even a few hours.

5 Usage

./dbpedia-enricher
The src/ directory contains a configuration file (config.py), in which the user can set:

- verbosity level,

- which predicates to learn,
- program mode (explained below).

The program can run in two modes:

- Extraction (default) - the program trains on available data and generates a list of RDF triples that can be imported into DBPedia.
  The triples extracted are stored in the results/ directory.
- Evaluation - the program trains on available data and then using manually annotated data, evaluates its performance.
  Program is distributed with manually annotated data for several predicates (in the tests/ directory).

6 Results

The system is distributed with text files containing RDF triples generated by running the program on several relations using the Polish DBPedia and Wikipedia.