DBpedia Spotlight at the MSM2013 Challenge

Pablo N. Mendes[1], Dirk Weissenborn[2], and Chris Hokamp[3]

[1] Kno.e.sis Center, CSE Dept., Wright State University
[2] Dept. of Comp. Sci., Dresden Univ. of Tech.
[3] Lang. and Inf. Tech., Univ. of North Texas
pablo@knoesis.org,dirk.weissenborn@mailbox.tu-dresden.de
christopherhokamp@my.unt.edu

1 Introduction

DBpedia Spotlight [5] is an open source project developing a system for automatically annotating natural language text with entities and concepts from the DBpedia knowledge base. The input of the process is a portion of natural language text, and the output is a set of annotations associating entity or concept identifiers (DBpedia URIs) to particular positions in the input text. DBpedia Spotlight provides programmatic interfaces for phrase recognition and disambiguation (entity linking), including a Web API supporting various output formats (XML, JSON, RDF, etc.)

The annotations generated by DBpedia Spotlight may refer to any of 3.77 million things in DBpedia, out of which 2.35 million are classified according to a cross-domain ontology with 360 classes. Through identity links, DBpedia also provides links to entities in more than 100 other languages, and tens of other data sets. This paper describes our application of DBpedia Spotlight to the challenge of extracting Person (`PER`), Location (`LOC`), Organization (`ORG`) and Miscellaneous (`MISC`) entities from microposts (e.g. tweets) as part of the MSM2013 Challenge at WWW2013.

All of the code used in this submission is available as Open Source Software, and all of the data used is shared as Open Data. A description of the software, data sets and more detailed evaluations are available from our supporting material page at `http://spotlight.dbpedia.org/research/msm2013/`.

Table 1. Comparison between NER approaches on the MSM2013 Challenge Training Set.

Syst./NERType	PER	LOC	ORG	MISC	**Average**
	P / R / F1	P / R / F1	P / R / F1	P / R / F1	P / R / F1
Unsup. (1)	0.95 0.50 0.65	0.62 0.58 0.60	0.62 0.38 0.47	0.22 0.21 0.21	0.60 0.42 0.48
CRF (2)	0.86 0.66 0.75	0.82 0.7 0.76	0.73 0.56 0.63	0.49 0.29 0.36	0.72 0.53 0.61

· **#MSM2013 Workshop Concept Extraction Challenge Proceedings** ·
available online as CEUR Vol-1019, at: `http://ceur-ws.org/Vol-1019`
Making Sense of Microposts Workshop @ WWW'13, May 13th 2013, Rio de Janeiro, Brazil

2 Datasets

DBpedia Spotlight's annotation model was constructed based on a number of datasets derived mainly from DBpedia and Wikipedia. First, for each DBpedia resource r, we extracted from Wikipedia all paragraphs containing wiki links with target on r's Wikipedia article. Second, from the collection of wiki links, disambiguation pages and redirects, we extracted a number of *lexicalization examples* – words that have been used to express a given DBpedia entity. Third, we use the community-maintained DBpedia Ontology mappings to collect a list of ontology classes (and superclasses) for each DBpedia resource. More details on this preliminary extraction process are available from Mendes et al., 2011 [5] and Mendes et al. 2012 [4].

To adapt this framework to the challenge, we also extended the coverage of known instance types by importing extra `rdf:type` statements between DBpedia and the DBpedia Ontology from Aprosio et al., 2013 [1], between DBpedia and Freebase[4] and between DBpedia and OpenCyc[5] by Pohl, 2012 [7].

Subsequently, we extended our lexicalization examples with a number of person and organization names based on 'naming' ontology properties such as `foaf:givenName`, `foaf:name`, `foaf:familyName`, etc.We further extended our lexicon with gazeteers from BALIE [6] including names for association, company designator, company, government, military, first name, last name, person title, celebrity, month, city, state province, country.

To allow our tool to output the target types of the challenge, we manually browsed through the ontology and created mapping from the types used in the MSM2013 Challenge, and the ontology types in the DBpedia Ontology, Freebase and OpenCyc. We refer to this set as "Manual Mappings."

Evaluation Corpus Pre-processing. The version of the MSM2013 Challenge corpus used in our evaluation contains a number of undesirable artifacts, presumably resulting from pre-processing parsing and tagging steps. The text was seemingly pre-tokenized, including spaces between tokens and punctuation, although not consistently so throughout the data set.

In our pre-processing, we attempted to reconstruct original sentences by adding extra markers as token separators (\/), as well as removing parsing artifacts (`-[LR]RB-`, `#B-ORG/`), Twitter markers (`RT`,`#\S+`), and other artifacts included in the training set for anonymization (`_URL_`, `_MENTION_`, `_Mention_`, `<NEWLINE>` and `_HASHTAG_`). For the sentence reconstruction, we also reverted the separation from the left-neighboring token of punctuation such as commas, apostrophes and exclamation marks. We will refer to this corpus as "reconstructed sentences".

[4] `http://downloads.dbpedia.org/3.8/links/freebase_links.nt.bz2`

[5] `http://opencyc.org`

3 Methodology

The Concept Extraction task proposed is very similar to the task performed by Named Entity Recognition (NER). The task can be broken down into two problems. First, a segmentation problem requires finding boundaries of entity names within sentences; and second, a classification problem requires correctly classifying the segment into one of the entity types. We have tested approaches that perform each task separately, as well as approaches that perform both tasks jointly.

First, we tested an unsupervised approach – i.e. one that does not use the training set provided in the challenge. It uses DBpedia Spotlight's phrase recognition and disambiguation to perform NER in a two-step process of segmentation and classification (dbpedia_spotlight_1.tsv). For this approach, the reconstructed sentences were sent through DBpedia Spotlight's lexicon-based recognition, and subsequently through the disambiguation algorithm. Based on the types of the entities extracted, we used our manual mappings to classify the names into one of the NER types.

Our joint segmentation/classification method is a supervised-machine learning approach enhanced by knowledge-based distant supervision from DBpedia. We use lexicalizations from DBpedia to indicate that a given token may be within an entity or concept name. This feature is intended to help with the segmentation task, particularly in cases where morphological characteristics of a word are not informative. Moreover, we use the ontology types for DBpedia resources to create a battery of features which further bias the classification task, according to the types predicted by DBpedia Spotlight.

We collected all our best features and created a Linear-Chain Conditional Random Fields (CRF) model to act as our NER (dbpedia_spotlight_2.tsv). We used Factorie [3] to implement our CRF. Our features include morphological (e.g. punctuation, word shape), context-based (e.g. surrounding tokens) and knowledge-based characteristics. Our knowledge-based features include the presence of a token within a name in our knowledge base, as well as the types predicted for this entity.

Given those features and the provided training corpus, the model is trained using stochastic gradient ascent. Gibbs sampling is used to estimate the posterior distribution for each label during training. We also added a small post-processing filter to remove whole entities that contain less than two letters or digits in them as well as entities with name "the" and "of".

Finally, we included Stanford NER [2] as our third baseline (dbpedia_spotlight_3.tsv), since it is a well known NER implementation.

4 Evaluation and Discussion

Table 1 presents our evaluation results on the training set. Precision, recall and F1 on Table 1 were computed based on the overlap (using exact name and type matches) between the set of entities we extracted and the set of annotated entities. The scores shown for our supervised method are our averaged 10-fold cross-validation scores.

We also report token-based precision, recall and F1 averaged over a 10-fold cross-validation on the training set. For Stanford NER (Vanilla) (with default features), we obtain P: 0.77, R: 0.54 and F1: 0.638. For Stanford NER (Enhanced), after adding our knowledge-based features, we observe improvements to P: 0.806, R: 0.604 and F1: 0.689. The same evaluation approach applied to DBpedia Spotlight CRF yields P:0.91, R:0.72, F1:0.8.

We found the segmentation to be far harder than classification in this dataset. First, as expected in any task that requires agreement between human experts, some annotation decisions are debatable. Second, inconsistent tokenization was a big issue for our implementation.

In some cases, our model found annotations that were not included by the human-annotators, such as `ORG/twitter`, where "twitter account" could be (but was not) interpreted as an account within the `ORG` Twitter. In other cases, our model trusted the tokenization provided in the training set and predicted `MISC/Super Bowl-bound` while the human-generated annotation was `MISC/Super Bowl`.

However, in general, after guessing correctly the boundaries, the type classification seemed an easier task. Our manual mappings already obtain an average accuracy over 82%. After training, those numbers are improved even further.

However, in some cases, there seems to be some controversial issues in the classification task. Is "Mixed Martial Arts" a `Sport` or a `SportEvent`? Is "Hollywood" an organization or a location? Depending on the context, the difference can be subtle and may be missed even by the human annotators.

By far, the toughest case to classify is `MISC`. Perhaps, such a "catch all" category may be too fuzzy, even for human annotators. The annotations often contain human languages like `MISC/English;MISC/Dutch;` where the guidelines stated that only Programming languages would be annotated.

In future work we plan to carefully evaluate the contribution of each of our features, further expand our evaluations within the MISC type, and conduct a reannotation of the dataset to normalize some of the issues found.

References

1. A. P. Aprosio, C. Giuliano, and A. Lavelli. Automatic expansion of DBpedia exploiting Wikipedia cross-language information. In *ESWC'13*, Montpellier, France, 2013 (to appear).
2. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *In ACL*, 2005.
3. A. McCallum, K. Schultz, and S. Singh. FACTORIE: Probabilistic programming via imperatively defined factor graphs. In *NIPS*, 2009.
4. P. N. Mendes, M. Jakob, and C. Bizer. DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. In *LREC'12*, Istanbul, Turkey, 2012.
5. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding light on the web of documents. In *I-Semantics*, Graz, Austria, 2011.

6. D. Nadeau, P. Turney, and S. Matwin. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. *Advances in Artificial Intelligence*, 4013:266–277, 2006.
7. A. Pohl. Classifying the Wikipedia Articles into the OpenCyc Taxonomy. In *WoLE'12 at ISWC'12*, 2012.