Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia

Milan Dojchinovski[1,2] and Tomáš Kliegr[2]

[1] Web Engineering Group
Faculty of Information Technology
Czech Technical University in Prague
milan.dojchinovski@fit.cvut.cz

[2] Department of Information and Knowledge Engineering
Faculty of Informatics and Statistics
University of Economics, Prague, Czech Republic
tomas.kliegr@vse.cz

Abstract. Targeted Hypernym Discovery (THD) performs unsupervised classification of entities appearing in text. A hypernym mined from the free-text of the Wikipedia article describing the entity is used as a class. The type as well as the entity are cross-linked with their representation in DBpedia, and enriched with additional types from DBpedia and YAGO knowledge bases providing a semantic web interoperability. The system, available as a web application and web service at entityclassifier.eu, currently supports English, German and Dutch.

1 Introduction

One of the most significant challenges in text mining is the dimensionality and sparseness of the textual data. In this paper, we introduce Targeted Hypernym Discovery (THD), a Wikipedia-based entity classification system which identifies salient words in the input text and attaches them with a list of more generic words and concepts at varying levels of granularity. These can be used as a lower dimensional representation of the input text.

In contrast to the commonly used dimensionality reduction techniques, such as PCA or LDA, which are sensitive to the amount of data, THD provides the same quality of output for all sizes of input text, starting from just one word. Since THD extracts these types from Wikipedia, it can also process infrequent, but often information-rich words, such as named entities. Support for live Wikipedia mining is a unique THD feature allowing coverage of "zeitgeist" entities which had their Wikipedia article just established or updated.

THD is a fully unsupervised algorithm. A class is chosen for a specific entity as the one word (concept) that best describes its type according to the consensus of Wikipedia editors. Since the class (so as the entity) is mapped to DBpedia, the semantic knowledge base, one can traverse up the taxonomy to the desired class granularity. Additionally, the machine-readable information obtainable on the disambiguated entity and class from DBpedia and YAGO can be used for feature enrichment.

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part III, LNAI 8190, pp. 654–658, 2013.

2 Architecture

THD is implemented in Java on top of the open source GATE framework[1].

Entity extraction module identifies entity candidates (noun phrases) in the input text. Depending on setting, entities can be restricted to named entities ("Diego Maradona") or common entities ("football").

Disambiguation module assigns entity candidate with a Wikipedia entry describing it. This module combines textual similarity between the entity candidate and article title with the importance of the article.

Entity classification module assigns each entity with one or more hypernyms. The hypernyms are mined with the THD algorithm (see Sec. 3) from the Wikipedia articles identified by the Disambiguation module. This mining is performed either on-line from live Wikipedia or from a Wikipedia mirror. The default option is to use the Linked Hypernyms Dataset, which contains 2.5 million article-hypernym pairs precomputed from a Wikipedia mirror.

Semantization module maps the entity as well as the class to `DBpedia.org` concepts. A "semantic enrichment" is also performed: once the entity is mapped, additional types are attached from DBpedia [1] and YAGO [2], the two prominent semantic knowledge bases. The final set of types returned for an entity thus contains the "linked hypernym" (hypernym mapped to DBpedia obtained with THD), and a set of DBpedia and YAGO types.

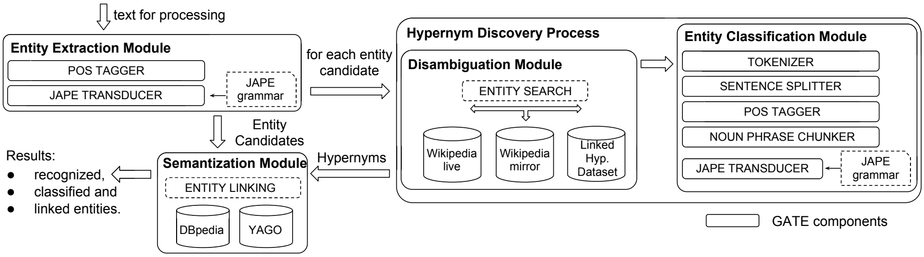


Fig. 1. Architecture overview

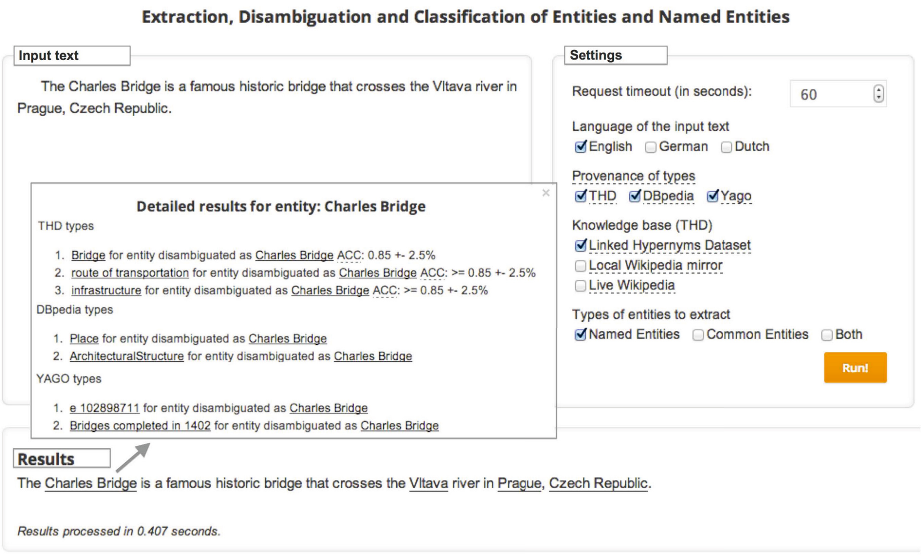
3 Hypernym Discovery Algorithm and Benchmark

Hypernym discovery is performed with hand-crafted lexico-syntactic patterns. These were in the past primarily used on larger text corpora with the intent to discover all word-hypernym pairs in the collection [7]. With *Targeted* Hypernym Discovery we apply lexico-syntactic patterns on a *suitable document* (Wikipedia article) with the intent to extract *one hypernym* at a time (details in [3,4]).

THD performance was measured on the following benchmarks independent on the input text: a) discovering correct hypernym given a Wikipedia article, b) linking hypernym to a semantic web identifier. The outcome of the evaluation[2]

[1] `http://gate.ac.uk`

[2] The results and the "High accuracy dataset" are available at `http://ner.vse.cz/datasets/linkedhypernyms/`.



Extraction, Disambiguation and Classification of Entities and Named Entities

Input text

The Charles Bridge is a famous historic bridge that crosses the Vltava river in Prague, Czech Republic.

Settings

Request timeout (in seconds): 60

Language of the input text

☑English □German □Dutch

Provenance of types

☑THD ☑DBpedia ☑Yago

Knowledge base (THD)

☑Linked Hypernyms Dataset

□Local Wikipedia mirror

□Live Wikipedia

Types of entities to extract

☑Named Entities □Common Entities □Both

Detailed results for entity: Charles Bridge

THD types

1. Bridge for entity disambiguated as Charles Bridge ACC: 0.85 +- 2.5%
2. route of transportation for entity disambiguated as Charles Bridge ACC: >= 0.85 +- 2.5%
3. infrastructure for entity disambiguated as Charles Bridge ACC: >= 0.85 +- 2.5%

DBpedia types

1. Place for entity disambiguated as Charles Bridge
2. ArchitecturalStructure for entity disambiguated as Charles Bridge

YAGO types

1. e 102898711 for entity disambiguated as Charles Bridge
2. Bridges completed in 1402 for entity disambiguated as Charles Bridge

Results

The Charles Bridge is a famous historic bridge that crosses the Vltava river in Prague, Czech Republic.

Results processed in 0.407 seconds.

Fig. 2. Screenshot of the system (edited to fit the page)

altogether on 16.500 entity articles (English, German, Dutch) is reported in [3]. The best results were obtained for the German person subset, with precision 0.98 and recall 0.95. This is on par with the the best results in the respective metrics recently reported in [5]: 0.97 precision for lexico-syntactic patterns and 0.94 recall for Syntactic-Semantic Tagger. The overall accuracy of discovering plain text (linked) hypernyms for English is 0.95 (0.85), for Dutch 0.93 (0.88) and German 0.95 (0.77). These numbers provide a lower bound on the error of THD, since they do not include the entity recognition error and particularly the disambiguation error (matching entity with a Wikipedia article).

4 Comparison with Related Systems

While techniques for Named Entity Recognition and classification (NER) are well-researched, NER classifiers typically need to be trained on large labeled document corpora, which generally involve only several labels, making them unsuitable for dimensionality reduction. Replacement of “Maradona” with “Person” loses too much meaning for most applications. The recent shift from human-annotated corpora to Wikipedia in some systems allows to provide types with finer granularity, and also broadening of the scope to “common” entities. In this section (and accompanying screencasts), we present a comparison with two best-known academic systems DBpedia Spotlight [6] and AIDA [8].

Real-time Mining. THD directly incorporates a text mining algorithm. Once an entity is disambiguated to a Wikipedia article, the system retrieves the article from Wikipedia and extracts the hypernym from its free text. The mining speed is about 1 second per entity including network overhead. This allows to discover types for entities, which had their article only recently added to Wikipedia, or adapt to changes in Wikipedia. The authors are not aware of any other system that incorporates query-time Wikipedia mining. AIDA and DBpedia Spotlight lookup the disambiguated entity in a database of types.

Complementarity to other Systems. Since THD extracts the types from *free text*, the results are largely complementary to types returned by other Wikipedia-based systems. These typically rely on DBpedia or YAGO knowledge-bases, which are populated from article *categories* and *"infoboxes"*, the semistructured information in Wikipedia. As a convenience, THD returns types from DBpedia and YAGO in addition to the mined hypernym. The complementary character of the results can be utilized for classifier fusion.

Right Granularity. For many entities DBpedia and YAGO-based systems provide a long list of possible types. For example, DBpedia assigns Diego Maradona with 40 types including `dbpedia-owl:SoccerManager`, `foaf:Person` as well as the highly specific `yago:1982FIFAWorldCupPlayers`. THD aids the selection of the "right granularity" by providing the most frequent type, as selected by Wikipedia editors for inclusion into the article's first sentence. For Maradona, as of time of writing, THD returns "manager".[3]

Multilinguality. System currently supports English, Dutch and German, extensibility to a new language requires only providing two JAPE grammars and plugging in correct POS tagger (ref. to Fig. 2). DBpedia Spotlight and AIDA support only English.

Acknowledgements. This research was supported by the European Union's 7th Framework Programme via the LinkedTV project (FP7-287911) and CTU in Prague grant (SGS13/100/OHK3/1T/18).

References

1. Bizer, C., et al.: DBpedia - a crystallization point for the web of data. Web Semant 7(3), 154–165 (2009)
2. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence 194, 28–61 (2013)
3. Kliegr, T., Dojchinovski, M.: Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery (Submitted)
4. Kliegr, T., et al.: Combining captions and visual analysis for image concept classification. In: MDM/KDD 2008. ACM (2008)
5. Litz, B., Langer, H., Malaka, R.: Sequential supervised learning for hypernym discovery from Wikipedia. In: Fred, A., Dietz, J.L.G., Liu, K., Filipe, J. (eds.) IC3K 2009. CCIS, vol. 128, pp. 68–80. Springer, Heidelberg (2011)

[3] As demonstrated in [4], the algorithm used can also return multi-word hypernyms ("soccer manager"). This feature is not yet available in THD.

6. Mendes, P.N., Jakob, M., Garcia-Silva, A., Bizer, C.: DBpedia spotlight: Shedding light on the web of documents. In: I-Semantics (2011)
7. Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. In: Advances in Neural Information Processing Systems, vol. 17, pp. 1297–1304. MIT Press, Cambridge (2005)
8. Yosef, M.A., et al.: AIDA: An online tool for accurate disambiguation of named entities in text and tables. PVLDB 4(12), 1450–1453 (2011)