Document Analytics through Entity Resolution[*]

João Santos, Bruno Martins, and David S. Batista

Instituto Superior Técnico and INESC-ID, Lisboa, Portugal
{joao.d.santos,bruno.g.martins,dsbatista}@ist.utl.pt

Abstract. We present a prototype system for resolving named entities, mentioned in textual documents, into the corresponding Wikipedia entities. This prototype can aid in document analysis, by using the disambiguated references to provide useful information in context.

Keywords: Text Mining, Information Extraction, Entity Resolution.

1 Introduction

Even as structured databases and semantic knowledge bases become prevalent, a substantial amount of human knowledge is still available as free-form text. News articles can, for instance, contain information about entities (i.e., people, organizations, and locations) and their relationships. For humans, reading and understanding large volumes of text is a time consuming task, and so automatically extracting and organizing this information is in high demand.

We present a prototype system that enhances textual documents by identifying references to people, places, or organizations, afterwards disambiguating these references by linking them to identifiers in a knowledge base such as Wikipedia. This paper briefly introduces a web-based demonstration of this prototype system, also outlining its main components.

2 Entity Resolution

Named entity resolution is an important text analytics problem that has been getting an increasing attention [2]. The problem involves two separate sub-tasks, namely (i) entity identification, and (ii) entity disambiguation. The first subtask is deeply related to Named Entity Recognition (NER), a problem that has been thoroughly studied in the Natural Language Processing community. In our system, entity identification is performed through the Stanford NER system, with models trained for the English (i.e., the standard model distributed with

[*] This research draws on work from two projects supported by Fundação para a Ciência e Técnologia (FCT), namely the projects with references PTDC/EIA-EIA/109840/2009 (SInteliGIS) and UTA-Est/MAI/0006/2009 (REACTION). The work reported on this paper was also supported through INESC-ID's multianual funding provided by the FCT (PEst-OE/EEI/LA0021/2013).

X. Lin et al. (Eds.): WISE 2013, Part II, LNCS 8181, pp. 531–534, 2013.

the tool[1]), Spanish (trained with the CoNLL-02 data[2]) and Portuguese (trained with data from the CINTIL corpus[3]) languages.

The second sub-task involves re-expressing the identified entity references into a standard unambiguous format (i.e., mapping each entity reference, previously recognized by the NER system, to an identifier specific to the real-world concept that is being referred to in the text). In our system, the mappings from entity references to real-world concepts are made through a knowledge base built from Wikipedia[4] and DBpedia[5]. Wikipedia is a collaborative wiki-based encyclopedia that covers almost all areas of human knowledge, with articles written in standard prose that are mostly intended for human consumption. On the other hand, DBpedia is a project concerning with the extraction of structured information from Wikipedia articles, representing this information in a machine-readable semantic graph using Resource Description Framework triples. For building the knowledge base, we essentially used the entities from the English, Portuguese, or Spanish versions of Wikipedia, categorized in the DBpedia structured ontology as entities corresponding to either people, organizations, or locations.

The method for performing entity disambiguation follows the general methodology from systems participating in the TAC-KBP yearly-challenge on named entity disambiguation [2], and it involves the following main tasks:

1. **Query Expansion**: Entities may be referenced by several alternative names, some of which more ambiguous than others. Given a reference, we apply expansion techniques that try to identify other names, in the source document, that reference the same entity. We specifically consider two simple mechanisms, namely one that finds alternative names by looking for a textual pattern that corresponds to a set of capital words followed by the alternative name inside parentheses (i.e., finding expressions like *United States (US)*), or vice-versa, and another that looks for longer entity mentions in the source text (i.e., *Union of Soviet Socialist Republics* as an expansion for *USSR*).
2. **Candidate Generation**: This step filters the Knowledge Base (KB) entries that might correspond to the query, based on string similarity. Some of Wikipedia's link structure (e.g., disambiguation pages, redirects, anchors, etc.) is also used to obtain alternative names. We specifically return the top 50 most likely entries in the KB (i.e., those whose name(s) are more similar to the entity reference, and whose textual descriptions are similar to the support text), according to a retrieval model supported by a Lucene[6] index.
3. **Candidate Ranking**: This step sorts the retrieved candidates according to the likelihood of being the correct referent, using the LambdaMART learning to rank algorithm as implemented in the RankLib[7] software library. Three

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

[2] http://www.cnts.ua.ac.be/conll2002/ner/

[3] http://cintil.ul.pt/cintilfeatures.html#corpus

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

[5] http://dbpedia.org/

[6] http://lucene.apache.org/

[7] http://people.cs.umass.edu/~vdang/ranklib.html

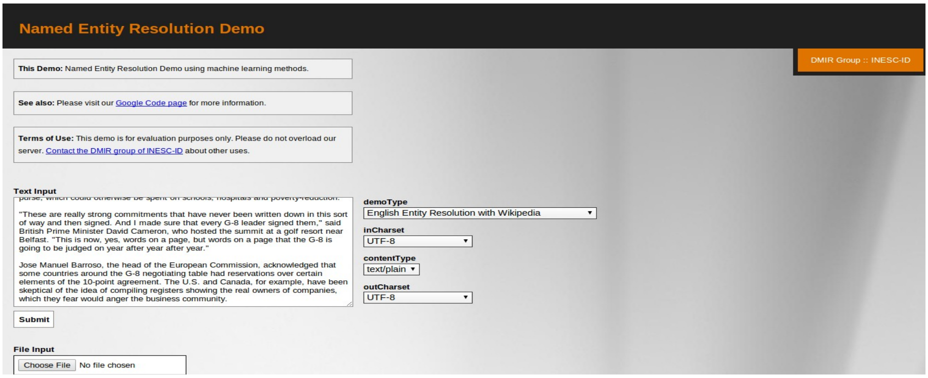
ranking models (i.e., for disambiguating entities in English, Spanish and Portuguese texts) were trained to optimize accuracy over sets of disambiguation examples that were automatically gathered from Wikipedia (i.e., we used hypertext anchors from links towards entities in the knowledge base, occurring in Wikipedia documents different from those in the knowledge base). These models leverages on a rich set of features for representing each candidate, including (i) candidate authority features, (ii) textual similarity features, (iii) topical similarity features, (iv) name similarity features, (v) entity-based features, (vi) geospatial features, and (vii) document coherence features. Space restrictions prevent us from showing the complete set of features here, but the reader can refer to a previous paper describing our participation in the 2011 edition of TAC-KBP, where an extensive set of experiments is reported with an early version of the system [1].

4. **Candidate Validation**: This step decides whether the top ranked referent is an error, resulting from the fact that the correct referent is not given in the knowledge base, through a Random Forest classifier that reuses the features from the ranking model, and that also considers some additional features for representing the top ranked result (e.g., the candidate ranking score, or results from outlier tests over the ranked lists of candidates).

The most innovative aspects of our named entity resolution system relate to the usage of the LambdaMART state-of-the-art learning to rank method, and to the extensive set of features that was considered.

Our prototype presents the entity resolution functionality through a web-based interface, through which users can input the text where entities are to be resolved. The system replies with an XML document encoding the entities occurring in the text, together with the results for their disambiguation. A stylesheet builds a web page from this XML document, where the named entities in the original text appear linked to the corresponding Wikipedia page. Through tooltips, the user can quickly access overviews on the entities that were referenced in the text (e.g., we show photos representing the referenced entities, elementary metadata attributes about the entities, and maps with pins for the latitude and longitude coordinates of the referenced locations).

Figures 1 and 2 present two screenshots of the web-based interface for our entity recognition and disambiguation system. Figure 1 shows the main screen of the service. There are two options to introduce the desired input text to be disambiguated, namely an option to paste it into a text box, and another to upload a file containing the target text. Before submitting this text to the entity linking system, the language needs to be chosen, and the system supports English, Spanish, and Portuguese. Figure 2 presents a screenshot for the output of the system, i.e. the result that is produced from the input textual document. All entities recognized by the NER model apperar highlighted, with each color representing a different entity type. These entities are also linked to the corresponding Wikipedia page, according to the disambiguations made by the system. A tooltip with relevant additional information about each entity pops up when the user moves the mouse over the entity reference.



Named Entity Resolution Demo

DMIR Group :: INESC-ID

This Demo: Named Entity Resolution Demo using machine learning methods.

See also: Please visit our Google Code page for more information.

Terms of Use: This demo is for evaluation purposes only. Please do not overload our server. Contact the DMIR group of INESC-ID about other uses.

Text Input

purse, which could otherwise be spent on schools, hospitals and poverty-reduction.

"These are really strong commitments that have never been written down in this sort of way and then signed. And I made sure that every G-8 leader signed them," said British Prime Minister David Cameron, who hosted the summit at a golf resort near Belfast. "This is now, yes, words on a page, but words on a page that the G-8 is going to be judged on year after year after year."

Jose Manuel Barroso, the head of the European Commission, acknowledged that some countries around the G-8 negotiating table had reservations over certain elements of the 10-point agreement. The U.S. and Canada, for example, have been skeptical of the idea of compiling registers showing the real owners of companies, which they fear would anger the business community.

demoType: English Entity Resolution with Wikipedia ▾

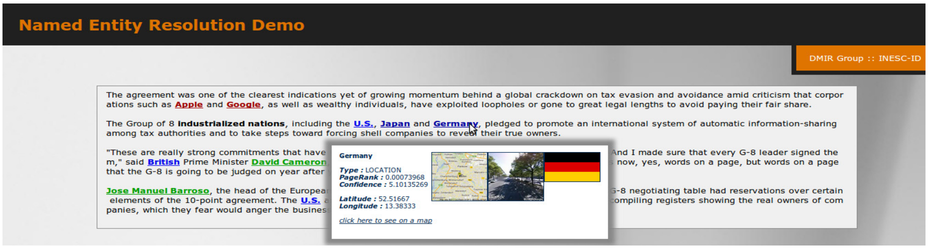
inCharset: UTF-8 ▾

contentType: text/plain ▾

outCharset: UTF-8 ▾

File Input: No file chosen

Fig. 1. Data entry form for the named entity resolution system



Named Entity Resolution Demo

DMIR Group :: INESC-ID

The agreement was one of the clearest indications yet of growing momentum behind a global crackdown on tax evasion and avoidance amid criticism that corporations such as Apple and Google, as well as wealthy individuals, have exploited loopholes or gone to great legal lengths to avoid paying their fair share.

The Group of 8 **industrialized nations**, including the U.S., Japan and Germany, pledged to promote an international system of automatic information-sharing among tax authorities and to take steps toward forcing shell companies to reveal their true owners.

"These are really strong commitments that have … And I made sure that every G-8 leader signed the
m," said British Prime Minister David Cameron … now, yes, words on a page, but words on a page
that the G-8 is going to be judged on year after

Jose Manuel Barroso, the head of the Europea… G-8 negotiating table had reservations over certain
elements of the 10-point agreement. The U.S. a… compiling registers showing the real owners of com
panies, which they fear would anger the busines

Germany

Type : LOCATION
PageRank : 0.00073968
Confidence : 5.10135269

Latitude : 52.51667
Longitude : 13.38333

click here to see on a map

Fig. 2. Output generated by the named entity resolution system

3 Conclusions

We demonstrate an end-to-end system for extracting and disambiguating named entities in textual documents, which combines recent developments in information extraction and natural language engineering. This system integrates many freely available open-source components (e.g., Stanford NER, Lucene, etc.) and offers scalability for processing large datasets.

Many opportunities exist for extending the system. For future work, we intend to incorporate relationship extraction into our information extraction pipeline. We also plan to integrate graph visualization methods into our system, in order to support the visual analysis of co-occurrences or of other types of relations between entities, as extracted from large document collections.

References

1. Anastácio, I., Calado, P., Martins, B.: Supervised learning for linking named entities to wikipedia pages. In: Proceedings of the Text Analysis Conference (2011)
2. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Poibeau, T., et al. (eds.) Multi-source, Multi-lingual Information Extraction and Summarization. Springer (2011)