Efficient Entity Annotation for Large Scale Web Archives

Elena Demidova[1], Julian Szymanski[2] Sergej Zerr[1], and Karol Draszawka[2]

[1] L3S Research Center, Hannover, Germany,
{demidova, zerr}@L3S.de,
[2] Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Poland
{julian.szymanski, karol.draszawka}@eti.pg.gda.pl,

1 Motivation

Intended to capture the history of Internet development for further usage and research, Web archiving initiatives are collecting available online resources over long periods of time. Such archives are containing large amounts of textual data in form of Web pages and are forming important source for cultural research in digital humanities. However, such data collections are difficult to access and to process effectively due to their large sizes (growing continuously). One of the methods for improving document representation for machine based processing is the addition of semantic annotations such as appearance of named entities in the text. This task involves mapping a short sequence of characters from a source text into one or several indexes in a given dictionary. This problem, known as "Named Entity Recognition", requires effective dictionary table lookup and solving the unambiguous mappings using additional information acquired from the context. Especially the second part of the task entity disambiguation - is computationally expensive and requires efficient solutions in order to make such annotations in large amounts of data feasible.

2 Datasets

In our challenge we plan to focus on a subset of the German Web, defined by the top-level domain '.de', archived by the Internet Archive and provided to the L3S Research Center in the context of the ALEXANDRIA project. This dataset contains Web data gathered from 1996 to 2013. There is a wide range of other datasets (with Wikipedia as a most popular among them) that can profit from efficient entity disambiguation techniques for linking between the textual contents of resources[3].

[3] linkedup-project.eu/

3 Problem Statement

There exist a large number of algorithms for named entity recognition (NER) and disambiguation such as (Stanford NER [3], DBpedia Spotlight [4], Illinois Wikifier [8] Wikipedia Miner [6], Wikify! [5], TagMe [2]). Whilst each of them have different running time properties, they cannot be directly applied to large scale datasets due to the lack of efficiency, scalability and, for most of them, dedicated application for Wikipedia dictionary. More scalable approaches, like dictionary lookups can be implemented efficiently but are not able to perform imprecise matches.

4 Possible Solutions

Dictionary-based entity matching is a simple and scalable approach for finding exact matches in linear time with respect to the data size when using hash indexes. Such dictionaries can be provided by the user, or can be often also found on the Web. For our experiments we extracted a list of persons provided by Wikipedia. Unfortunately this method cannot be used directly for imperfect matching [7] between source text and items in the dictionary. The task of finding the most probable (similar) entity for a piece of source text can be treated as general classification task and solved using machine learning techniques. However, training of such classifiers is challenging due to lack of proper training data and the scale of the construction problem. The problem of efficient imperfect matching can be partially solved with adopted locality sensitive hashing [**?**]. LSH algorithms are able to obtain similar hash values for similar items (character sequences in our case), such that similar items obtain same hash value with high probability. A particular example is a MinHash algorithm [1] developed for efficient estimation of set similarity (a character sequence can be represented as a set of bi,- or trigrams), which is very close to Jaccard similarity measure. The challenge is to investigate, adopt, develop and test various LSH family algorithms for efficiently solving entity detection and disambiguation tasks in large text document datasets.

References

1. Broder, A.: On the resemblance and containment of documents. In: Compression and Complexity of Sequences 1997. Proceedings (1997)
2. Ferragina, P., Scaiella, U.: Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In: CIKM (2010)
3. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: ACL (2014)
4. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: ICSS (2011)
5. Mihalcea, R., Csomai, A.: Wikify!: Linking documents to encyclopedic knowledge. In: CIKM (2007)
6. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: CIKM (2008)

7. Navarro, G.: A guided tour to approximate string matching. ACM Comput. Surv. (2001)
8. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to wikipedia. In: ACL HLT (2011)