Web Semantics: Science, Services and Agents on the World Wide Web 37–38 (2016) 75–89

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

CohEEL: Coherent and efficient named entity linking through random walks

Toni Gruetze *, Gjergji Kasneci, Zhe Zuo, Felix Naumann

Hasso Plattner Institute, Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
29 January 2016
Accepted 2 March 2016
Available online 11 March 2016

Keywords:
Entity linking
Named entity disambiguation
Random walk
Machine learning

ABSTRACT

In recent years, the ever-growing amount of documents on the Web as well as in digital libraries led to a considerable increase of valuable textual information about entities. Harvesting entity knowledge from these large text collections is a major challenge. It requires the linkage of textual mentions within the documents with their real-world entities. This process is called *entity linking*.

Solutions to this entity linking problem have typically aimed at balancing the rate of linking correctness (precision) and the linking coverage rate (recall). While entity links in texts could be used to improve various Information Retrieval tasks, such as text summarization, document classification, or topic-based clustering, the linking precision is the decisive factor. For example, for topic-based clustering a method that produces mostly correct links would be more desirable than a high-coverage method that leads to more but also more uncertain clusters.

We propose an efficient linking method that uses a random walk strategy to combine a precision-oriented and a recall-oriented classifier in such a way that a high precision is maintained, while recall is elevated to the maximum possible level without affecting precision. An evaluation on three datasets with distinct characteristics demonstrates that our approach outperforms seminal work in the area and shows higher precision and time performance than the most closely related state-of-the-art methods.

1. Named entity linking

Semi-structured and collaboratively created Web platforms, such as Wikipedia, have motivated a wide variety of research projects aiming at knowledge harvesting. For instance, large semantic knowledge bases, such as DBpedia [1] and YAGO [2], are based on the structured assets of Wikipedia, such as infoboxes and categories. However, harvesting implicit knowledge about entities in text without consistent structure, i.e., on websites or digital libraries, is still a major challenge. To address it, reliable techniques for *Named Entity Detection* and *Disambiguation* in natural language text are needed. Especially the disambiguation process is a critical step; it involves the correct grouping of textual mentions of the same real-world entity. If these groups are linked to corresponding entities in a knowledge base, the process is referred to as *Named Entity Linking* (NEL). Quote 1 shows an example text of a news article about the Deflategate scandal. The mentions of different entities within the text (e.g., *U.S. District Court*) are emphasized. Note that only named entities have been highlighted, where general entities, such as, concepts (e.g., judge) or temporal expressions (e.g., week) have not.

> U.S. District Court, New York: Judge Richard Berman expects to rule this week on Brady's four-game suspension appeal. Will the NFL finally prevail over Brady in the Deflategate saga?

Quote 1: Running example of a Named Entity Linking task with emphasized entity mentions.

The task is to link these highlighted mentions to corresponding knowledge base entities. Dredze et al. identify three key challenges for NEL [3]:

(i) Name variations occur for various reasons, for instance to reduce the length of the actual mention, such as *"NFL"* as abbreviation for *"National Football League"*.

(ii) The absence of entities is another important challenge. For instance, the mentioned *"Deflategate"* is not covered by Wikipedia versions from 2014 or earlier. A linking algorithm has to cautiously handle such mentions and should not

* Corresponding author.
E-mail addresses: toni.gruetze@hpi.de (T. Gruetze), gjergji.kasneci@hpi.de (G. Kasneci), zhe.zuo@hpi.de (Z. Zuo), felix.naumann@hpi.de (F. Naumann).

http://dx.doi.org/10.1016/j.websem.2016.03.001
1570-8268/

link them to similarly named entities, such as the data compression algorithm *DEFLATE*.

(iii) Furthermore, the entity ambiguity concerns cases where different entities are referred to by the same name, such as *"Brady"*. An algorithm with the goal to link this mention to an entity from Wikipedia has to choose between various geographical entities, such as a city in Texas and a village in Nebraska, hundreds of Persons (e.g., the famous American football quarterback and 4 time super bowl winner, the award-winning film director and producer, the American judge and Associate Justice, …).

There are various application areas that could benefit from reliable NEL techniques. Digital libraries are usually managed by information systems that enable textual keyword search. However, state-of-the-art keyword-based search engines are not able to deal with name variations or ambiguity. Hence, previously identified entity mentions might enable users to identify documents containing information about specific entities. Furthermore, relationships between entities can be inherited from co-occurrences in documents and aggregated into large semantic relationship graphs. The identified entity mentions as well as the relationships between them might serve as a starting point for topic-based clustering and document classification to enable a categorization of the document collection.

For Web retrieval tasks, such as person or product search, reliable disambiguation tools could enable the clustering of results, so that each cluster represents only one real-world entity, allowing the user to focus on the documents of interest [4]. Encyclopedic and scholarly search as well as question answering approaches would immensely benefit from reliable NEL techniques, by using the entity links in a document corpus to identify relevant texts [5]. Several academic projects in the realm of artificial intelligence, e.g., Read the Web [6] or YAGO-NAGA [7], as well as different scientific workshops and benchmarks, such as the TAC knowledge-base-population track or the ERD challenge [8], or GERBIL benchmark repository [9], have highlighted the importance of NEL for bootstrapping relationship extraction from natural language texts.

The effectiveness of those applications highly depends on the quality of the named entity detection and disambiguation step. This quality is in turn strongly influenced by the type of collected documents. For instance, digital libraries containing scientific publications share different text characteristics than collections containing news articles, social network posts, music reviews, or even entire books of fiction. Text structure, length, and topic variability are decisive for the vocabulary and the contextual information contained in the text. In our experimental evaluation, we show that most state-of-the-art algorithms lack precision, with values between 30% and 80% depending on the text type, only expensive coherence reasoning strategies can lead to more reliable results. The approach presented in this paper focuses on the efficient retrieval of reliable, i.e., high precision, alignments of approximately 90%.

Furthermore, the efficiency of NEL algorithms is a decisive factor that is usually neglected in current state-of-the-art algorithms. For instance, the runtime of NEL algorithms is important for use cases dealing with large text corpora, such as digital libraries with millions of texts, e.g., the Internet Archive, arXiv.org, or Google Books. NEL algorithms with a runtime of minutes per document would take two years to annotate such corpus. Another efficiency bound use-case are streaming applications. Given an infinite stream of incoming documents to be annotated, e.g., blog posts, news articles, or short messages, an NEL system has to process the texts approximately in the time window between two consecutive texts. Our contributions are the following:

1. We propose CohEEL, a supervised two-step model that enables the reliable NEL results. The model is designed to incorporate a knowledge base and (i) automatically adapts to different input text types, (ii) incorporates arbitrary mention-scoring functions, and (iii) considers known relations between mentioned entities within documents (i.e., coherence).
2. Based on the CohEEL model, we discuss a concrete configuration that provides reliable and coherent entity alignments to the knowledge bases YAGO and Wikipedia with a precision of approximately 90%.
3. Finally, we provide an exhaustive experimental comparison of our algorithm with state-of-the-art methods with respect to (i) linking quality and (ii) runtime.

2. Problem and prior work

In this section we introduce basic terms and discuss the related research for the Named Entity Linking problem. A specific discussion of the competitors used in the evaluation can be found in Section 5.

2.1. The named entity linking problem

The term named entity stands for a real-world instance and is distinct from a concept, which represents a category of named entities. Hence, NEL differs from the Wikification process, introduced in [10], which deals with the automatic identification of links to Wikipedia articles (i.e., both concepts and instances). It is also distinct from the word-sense disambiguation, which identifies the sense of a word in a sentence and is not focused on named entities [11,12]. A prerequisite for NEL is the discovery of textual mentions that might refer to named entities.

Definition 1. A **mention** $m = (D, p)$ is a textual reference occurring in position p within document D and referring to some named entity.

In the following we assume that such Named Entity Recognition (NER) techniques are readily available; we employ the Stanford NER tool [13] to discover mentions. Please note, that some related work follows another definition of NEL, that includes the NER step [14].

The subsequent NEL task for a given set $M(D)$ of mentions in document D and a set $E(K)$ of named entities in knowledge base K is to find a partial mapping $f_a : M(D) \rightarrow E(K)$, such that if $m \in M(D)$ is mapped to $e \in E(K)$ then indeed m refers to the real-world entity represented by e. For instance, the mention *"Richard Berman"* shall be mapped to the Wikipedia article of Richard M. Berman. The alignment function f_a remains partial, because in some cases K might not contain the entity that is referred to by a mention. For a total function, f_a can map the mention to a designated entity NIL in such cases. For better readability we use E' to refer to the extended knowledge base entity set $E(K) \cup \{\text{NIL}\}$, and E to refer to $E(K)$. Furthermore, we use M to address the set of mentions of a document $M(D)$ and refer to a mapping from mention m to entity e as an *alignment*, denoted (m, e).

So far, mentions are defined only as the occurrences of named entities in a document by means of unique positions. However, each mention m within a document covers different aspects of information about the entity it refers to, most importantly the surface and the context. In related research, these aspects are commonly referred to as local ranking features, because they are independent of other entity mentions in the document [15].

The **surface** $srfc(m)$ of a mention m is the actual word or phrase used to identify the referenced real-world instance in the text. For instance, surfaces in Quote 1 are $srfc(m_1) =$ "*U.S. District Court*" or $srfc(m_2) =$ "*New York*". The **context** $ctx_n(m)$ of mention m is a