Information Processing and Management 52 (2016) 1247–1257

Contents lists available at ScienceDirect

Information Processing and Management

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

Entity disambiguation to Wikipedia using collective ranking

Gang Zhao[a], Ji Wu[a,*], Dingding Wang[b], Tao Li[b]

[a] Department of Electronic Engineering, Tsinghua University, Beijing, China
[b] School of Computing and Information Sciences, Florida International University, Miami, FL, USA

ARTICLE INFO

Article history:
Received 19 May 2015
Revised 18 April 2016
Accepted 10 June 2016
Available online 23 June 2016

Keywords:
Named entity disambiguation
Feedback-query-expansion
Re-ranking

ABSTRACT

Entity disambiguation is a fundamental task of semantic Web annotation. Entity Linking (EL) is an essential procedure in entity disambiguation, which aims to link a mention appearing in a plain text to a structured or semi-structured knowledge base, such as Wikipedia. Existing research on EL usually annotates the mentions in a text one by one and treats entities independent to each other. However this might not be true in many application scenarios. For example, if two mentions appear in one text, they are likely to have certain intrinsic relationships. In this paper, we first propose a novel query expansion method for candidate generation utilizing the information of co-occurrences of mentions. We further propose a re-ranking model which can be iteratively adjusted based on the prediction in the previous round. Experiments on real-world data demonstrate the effectiveness of our proposed methods for entity disambiguation.

1. Introduction

Named entity identification is an essential component in many Natural Language Processing (NLP) applications, which enables information extraction from large-amount of documents. However, named entities such as dates, locations, names, and organizations often involve challenging ambiguity. For example, different people may share the same name, and on the other hand one person may have different names. Thus it is crucial to disambiguate these entities for effective information retrieval.

A general way for entity disambiguation is to link entities to an existing knowledge base such as Wikipedia, the largest encyclopedia. This task is called Entity Linking (EL), which is the foundation of knowledge base population and widely used in many information extraction and retrieval systems. Since 2009, Text Analysis Conference (TAC) has included the EL task Ji (2014); Ji, Grishman, Dang, Griffitt, and Ellis (2010) in its Knowledge Base Population (KBP) track. A typical EL system at TAC KBP track includes three function modules: (1) query expansion which aims to find out as many alias as possible according to the context of the queries, (2) candidate generation which generates candidates for linking, and (3) candidate ranking which ranks the candidates according to their relationships with queries. The typical methods used in an EL system usually deal with one query at a time and do not consider the relationships among different queries for the same document. However if two mentions appear in the same document, they should not be treated as independent to each other. For example, in the description of *"Gianni Infantino will take over from David Taylor as secretary general of UEFA from October 1, European football's governing body announced on Tuesday."*, we know that *"Gianni Infantino"* and *"David Taylor"* should share some common working experiences. The typical EL systems mainly rely on the text similarity calculation to

[*] Corresponding author.
E-mail addresses: zhaogang12@126.com (G. Zhao), wuji_ee@mail.tsinghua.edu.cn (J. Wu), wangd@fau.edu (D. Wang), taoli@cs.fiu.edu (T. Li).

http://dx.doi.org/10.1016/j.ipm.2016.06.002
0306-4573/© 2016 Elsevier Ltd. All rights reserved.

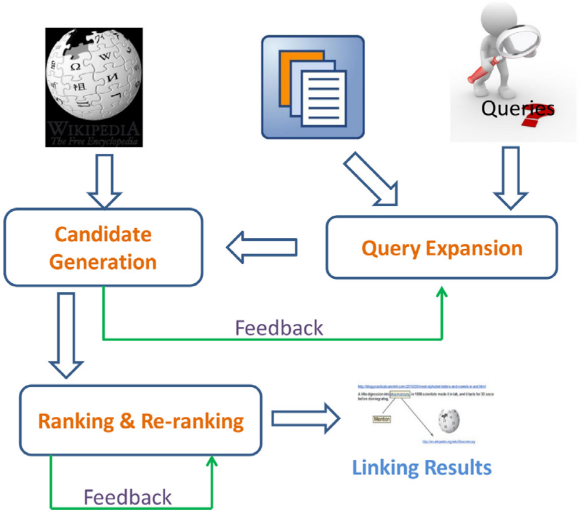

Fig. 1. The system framework.

disambiguate entities. However if a query mention is highly ambiguous, the set of candidates can be very large with more than thousands of candidates to rank.

Recent research shows that collective approaches, in which the name mentions in the same document are jointly linked by exploiting the relationships among them, can improve the EL performance significantly. For example, Cucerzan (2007) used a sum vector of all candidate vectors in a document to represent the document. This approach adds some relation features of different mentions in a document, but also introduces additional noises. Ratinov, Roth, Downey, and Anderson (2011) proposed a simple and efficient approach that can reduce the irrelevant information by using local features to predict global features. Han, Sun, and Zhao (2011) proposed a graph-based method which can model the global interdependence between different mentions in one document. They also used a collective inference algorithm to estimate the importance of each candidate node. Shen, Wang, Luo, and Wang (2012) linked entities by leveraging the semantic knowledge in the taxonomy of the knowledge base and used a sum vector (similar to Cucerzan (2007)) to estimate the unknown entity.

In this paper, we propose a re-ranking algorithm which iteratively adjusts the model based on the prediction from the previous round. We also add feedback procedures in the query expansion and candidate generation modules of the EL system. The rest of the paper is organized as follows. Section 2 introduces the proposed entity linking system framework. Section 3 proposes the candidate generation using query expansion with feedbacks. Section 4 discusses the proposed ranking algorithms. Section 5 conducts experiments to examine the performance of the proposed methods. Section 6 concludes the paper.

2. The system framework

Fig. 1 shows the framework of the proposed named entity linking system. Given a collection of queries and documents, we first conduct a query expansion. Then we search the expanded queries in the knowledge base corpus from Wikipedia to generate candidates. Different from existing methods, we add a feedback loop to iteratively refine the candidate sets to obtain a more accurate results. Finally, we rank the candidates using both local and global features and re-rank them using another feedback procedure to generate the final linking results.

3. Query expansion and candidate generation

3.1. Basic query expansion

The difficulties of the EL task lie in two aspects: name variation and name ambiguity. Name variation means that one entity may be described by different alias. For example, we can call the famous NBA player Michael Jordan for Jordan or MJ. Name ambiguity means that a name may express different entities. For example, Michael Jordan can also be used to refer a famous professor in machine learning at UC Berkeley.

Query Expansion aims to address the problem of name variation. In a document, the full names of entities are usually mentioned at the beginning of the document, and these entities are often mentioned using abbreviations or part of the names in the rest of the articles. The purpose of query expansion is to find the full names of the mentions which can increase the recall of the target entities. There are mainly two general matching strategies. (1) For queries composed of