Information Sciences 307 (2015) 18–38

Contents lists available at ScienceDirect

Information Sciences

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

Resolving polysemy and pseudonymity in entity linking with comprehensive name and context modeling

Ming Zhao-Yan [a,*], Chua Tat Seng [b]

[a] Department of Computer Science, Digipen Institute of Technology, Singapore
[b] School of Computing, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 29 November 2013
Received in revised form 6 February 2015
Accepted 14 February 2015
Available online 20 February 2015

Keywords:
Entity linking
Coreference
Polysemy
Pseudonymity
Name modeling
Context modeling

ABSTRACT

Names are important atomic information carriers in unstructured text. Matching names that refer to the same entities is an important issue in text analysis and a key component in many real world applications. Generally referred to as *entity linking*, it is defined as a task that aligns a name mentioned in free text to its corresponding entry in a Knowledge Base (KB). The difficulty of the task lies in the many-to-many correspondence between names and entities, causing the pseudonymity and polysemy issues. Existing work usually focuses on resolving polysemy by aggregating large numbers of loosely arranged features in supervised learning frameworks, with very few targeting the pseudonymity or both issues with the same depth. In this work, we tackle both issues by comprehensive modeling of an entity's name and context: we tackle the pseudonymity by modeling name variants on the query name and the KB title; and polysemy by modeling heterogeneous aspects of the query and KB context. Specially, we harness entity coreferences within query and KB documents together with the external alias resources for modeling name variants, and further use the name variants to identify focused context. Moreover, we propose a recall-boosted retrieval method for efficient candidate entity generation. Experimental results show that our proposed approach outperforms the state-of-the-art systems on the benchmark data.

1. Introduction

Names, such as person, organization and location names, are important atomic information carriers in unstructured text, such as newspaper articles. The referents that these names (or the rigid designators, as defined by Kripke [39]) stand for are called "Named Entities" in text processing research. On the other hand, the definitions and other detailed information about the entities are usually manually compiled and stored as entries in structured Knowledge Base (KB) such as dictionaries or encyclopedias. Entity linking is therefore a task that automatically aligns a name mentioned in unstructured text to its corresponding entry in a knowledge base.

Entity linking has been found useful in many real-world applications. A direct application is in an educational environment where entity linking can provide fast access to reference knowledge in study materials such as lecture notes and assignments. In encyclopedia itself, a new knowledge entry's content can be cross-referenced to existing entries, so as to build comprehensive knowledge interlinking. Wikify [51] is a successful system toward the goal of enriching Wikipedia articles with interlinks. Other systems include the Microsoft Smart Tags in later versions of Microsoft Word and the "Instant

* Corresponding author.
E-mail address: mingzhaoyan@gmail.com (Z.-Y. Ming).

http://dx.doi.org/10.1016/j.ins.2015.02.025
0020-0255/© 2015 Elsevier Inc. All rights reserved.

Lookup" feature of the Trillian[1] instant messaging client. Toward the building of the semantic web [50], entity linking can extend its application scope to providing instant references to any type of web document, such as microblogs, reviews, forum discussion threads, and more.

Entity linking is also an important topic in the text analysis community and the data management community [59]. Toward populating structured knowledge base, Text Analysis Conference 2009 introduced the *entity linking* task [47] that takes an entity mention and the document it appears in as the query and the Wikipedia as the KB. When no entity in the KB can be matched to the query, the query is predicted as NIL. This happens when the KB is not big enough, or the query is a newly emerged concept. Prior tasks such as Web People Search (WePS) [1] and Global Entity Detection and Recognition (GEDR) in Automatic Content Extraction focus on specific types of entities and less structured documents as knowledge bases. With entity linking enriched text, other text processing tasks such as summarization [56], entailment, and text categorization [53,70,65] can also benefit from the additional information attached to the original documents. In Fig. 1, we use an example consisting of a linking query and its KB entry to illustrate the TAC entity linking task [46,32].

However, entity linking is not a trivial task, due to the fact that names are often not the unique identifiers for entities. In other words, the relation between names and entities is not one-to-one but a many-to-many mapping. Specifically, this means that a name may stand for multiple entities (**polysemy**), such as *ABC* may refer to *American Broadcasting Company*, *Australian Broadcasting Corporation*, *ABC (newspaper)*; and an entity may also have multiple names (**pseudonymity**), such as full name, acronym, spelling variations, metaphorical names, and other aliases. For example, *American Broadcasting Company* (an American commercial broadcasting television network created in 1943) can be called as *Alphabet Network* (its alias), *ABC* (its call sign), and *American Broadcasting Company* (its official name). Therefore, the difficulty of entity linking lies in the many-to-many correspondence between names and entities, causing synonym and ambiguity, or the pseudonymity and polysemy issues.

Most existing works have successfully resolved the polysemy (ambiguity) issue [47,34,68], with very few works targeting the pseudonymity issue or both with the same depth. To resolving ambiguity, a typical approach in current work is to aggregate large numbers of loosely arranged features in a supervised learning framework. While the framework works well, it deserves more principled modeling of the problem in order to generate structured features. To tackle the pseudonymity issue, current work usually adopts an external alias list or equivalent. However, this method is subject to availability of such a list and the quality and quantity of items it covers. In view of the above, in this work, we propose to tackle both of the issues by comprehensive modeling of an entity's name and context.

Modeling name variants. As the name is the primary identifier of an entity, the modeling of pseudonymity, or name variants, plays an important part in entity linking. Most existing work explored name variants at the knowledge base side [47,68,2], namely, acquiring name variants for entries in KB. We take this one step further. Besides the external name mapping resources for enriching knowledge base entries, we harness entity coreferences within both query and knowledge base documents for expanding the query mention and the KB entry title respectively. In other words, we solve the pseudonymity issue by modeling name variants of both the query name and the KB title. Specifically, we propose a rule-based entity coreference method based on the Stanford multi-pass sieve framework to find in-document variants of the names. For example, we can find *North Queensland Cowboys*, which appears at the beginning of a query document, as a referent to the query mention *Cowboys*, where the full name is exactly the title of the KB entry linked to the query. At the knowledge base side, we extract alias lists from sources such as "titles of entity pages", "disambiguation pages", "redirect pages", and "anchor texts", which is also widely adopted name variations mining methods in the literature [68,67].

Modeling context. Though names are key identifiers of entities, they are not the unique ones. The fact that many names can refer to multiple entities causes the polysemy issue. Tackling polysemy by appropriate disambiguation models is important for the entity linking task. While most systems have an explicit or implicit context modeling component for disambiguation purpose [6,10,3], we propose a novel method that uses coreferences to identify the more focused context within a document. Besides the surrounding text, some novel aspects of entities, such as the attributes, the popularity, the categories, are also modeled.

Overall, our entity linking system consists of two stages: candidate generation and entity disambiguation. For candidate generation, we propose to use a recall-oriented retrieval model. For candidate disambiguation, we cast the linking between a query mention and a candidate entity in a learning-to-rank framework. Our proposed models for name variants and contexts are embodied as features which characterize matching between query name and entity name, query name and entity context, query context and entity name, and query context and entity context.

We empirically evaluate our proposed method for entity linking with the official Knowledge Base Population track entity linking task data. The experimental results show that the proposed retrieval based entity candidate generation method greatly enhances the recall, which raises the upper bound and reduces the cost of the follow-on computationally intensive entity disambiguation process. For entity disambiguation, the proposed name variant expansion model and context model outperform the state-of-the-art learning-to-rank models with uniform features. We show that the coreference enhanced name matching and context matching models are effective in resolving the pseudonymity and the polysemy issues in entity linking. The contributions of this work are threefold:

[1] http://www.trillian.im/.