Accepted Manuscript
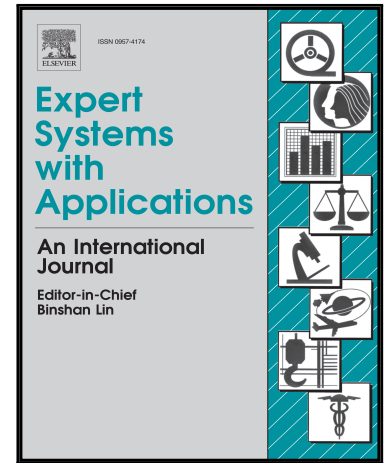
Exploiting Semantic Similarity for Named Entity Disambiguation in Knowledge Graphs

Ganggao Zhu, Carlos A. Iglesias

Please cite this article as: Ganggao Zhu, Carlos A. Iglesias, Exploiting Semantic Similarity for Named Entity Disambiguation in Knowledge Graphs, *Expert Systems With Applications* (2018), doi: 10.1016/j.eswa.2018.02.011

Exploiting Semantic Similarity for Named Entity Disambiguation
in Knowledge Graphs

Ganggao Zhu*, Carlos A. Iglesias

Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros de Telecomunicación, Departamento de Ingeniería de Sistemas Telemáticos, Avenida Complutense 30, Madrid, Spain

Abstract

With the increasing popularity of large scale Knowledge Graph (KG)s, many applications such as semantic analysis, search and question answering need to link entity mentions in texts to entities in KGs. Because of the polysemy problem in natural language, entity disambiguation is thus a key problem in current research. Existing disambiguation methods have considered entity prominence, context similarity and entity-entity relatedness to discriminate ambiguous entities, which are mainly working on document or paragraph level texts containing rich contextual information, and based on lexical matching for computing context similarity. When meeting short texts containing limited contextual information, such as web queries, questions and tweets, those conventional disambiguation methods are not good at handling single entity mention and measuring context similarity. In order to enhance the performance of disambiguation methods based on context similarity with such short texts, we propose SCSNED method for disambiguation based on semantic similarity between contextual words and informative words of entities in KGs. Specially, we exploit the effectiveness of both knowledge-based and corpus-based semantic similarity methods for entity disambiguation with SCSNED. Moreover, we propose a Category2Vec embedding model based on joint learning of word and category embedding, in order to compute word-category similarity for entity disambiguation. We show the effectiveness of these proposed methods with illustrative examples, and evaluate their effectiveness in a comparative experiment for entity disambiguation in real world web queries, questions and tweets. The experimental results have identified the effectiveness of different semantic similarity methods, and demonstrated the improvement of semantic similarity methods in SCSNED and Category2Vec over the conventional context similarity baseline. We further compare the proposed approaches with the state of the art entity disambiguation systems and show the performances of the proposed approaches are among the best performing systems. In addition, one important feature of the proposed approaches using semantic similarity, is the potential application on any existing KGs since they mainly use common features of entity descriptions and categories. Another contribution of the paper is an updated survey on background of entity disambiguation in KGs and semantic similarity methods.

Keywords: Entity Linking, Named Entity Disambiguation, Context Similarity, Semantic Similarity, Word Embedding, Knowledge Graph

1. Introduction

The increasing availability of Linked Open Data (LOD) has given birth to the notion of large scale KGs [8], with popular examples such as Freebase [11], DBpedia [9], and YAGO [52]. Named Entity Linking (NEL) is a fundamental module for developing KG-based applications, including text analysis [67], document retrieval [65], knowledge base population [56, 24], semantic search and question answering [95]. In general, a NEL system needs to detect a sequence of words (spots or mentions) in a given text, and to identify those mentions to entities registered in the given KG. The latter process of entity identification is not

a trivial task because it needs to tackle two difficult problems, namely synonymy and polysemy. To address synonymy problem, a NEL system needs to match an entity despite its diverse name variations such as abbreviations, spelling variations, nicknames to name a few. The main approach to solve the synonymy problem is to construct entity name dictionaries as complete as possible in order to cover diverse name variations [97], and to apply approximate string matching [24]. Thus, the performance of these techniques is mainly concerned with the quality of name dictionaries and approximate matching algorithms. The polysemy problem is caused by the fact that multiple entities in KGs might have the same name, and this is quite common for named entities. The task of addressing the polysemy problem for named entities is called Named Entity Disambiguation (NED), and there is a large body

---

*Corresponding author

Email addresses: `gzhu@dit.upm.es` (Ganggao Zhu),
`cif@dit.upm.es` (Carlos A. Iglesias)