Collective Entity Disambiguation Based on Hierarchical Semantic Similarity

Bingjing Jia, Beijing University of Posts and Telecommunications and Anhui Science and Technology University, Beiging and Huainan, Anhui, China

Hu Yang, Beijing University of Posts and Telecommunications, Beijing China

Bin Wu, Beijing University of Posts and Telecommunications, Beijing, China

Ying Xing, Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

Entity disambiguation involves mapping mentions in texts to the corresponding entities in a given knowledge base. Most previous approaches were based on handcrafted features and failed to capture semantic information over multiple granularities. For accurately disambiguating entities, various information aspects of mentions and entities should be used in. This article proposes a hierarchical semantic similarity model to find important clues related to mentions and entities based on multiple sources of information, such as contexts of the mentions, entity descriptions and categories. This model can effectively measure the semantic matching between mentions and target entities. Global features are also added, including prior popularity and global coherence, to improve the performance. In order to verify the effect of hierarchical semantic similarity model combined with global features, named HSSMGF, experiments were carried out on five publicly available benchmark datasets. Results demonstrate the proposed method is very effective in the case that documents have more mentions.

KEYWORDS

Entity Disambiguation, Global Coherence, Hierarchical Semantic Similarity, Knowledge Discovery

1. INTRODUCTION

Due to the prosperity of the web, a large number of unstructured texts have emerged to represent web content. These texts contain many mentions which are names of people, places, and organizations, etc. Unfortunately, one mention may have several meanings without considering the context (Shen et al., 2014). For example, given a sentence ‘’Donald Trump has arrived in Washington ahead of his inauguration as the 45th President of the United States.’’ The mention ‘’Washington’’ may refer to the capital of the United States, the first President of the United States, or a football club in Washington. It is helpful for understanding sentences finding the real-world entities which mentions refer to. This process is called entity disambiguation (ED). Many researchers regard entries in a knowledge base (KB) as surrogates for real world entities. Therefore, the main purpose of ED is to link mentions in text to corresponding entities in a KB such as Wikipedia. ED is an essential step in knowledge discovery by combining unstructured with structured data, and is beneficial to many applications,

DOI: 10.4018/IJDWM.2020040101

including knowledge discovery, question answering and knowledge base population. In this paper, we target at disambiguating entities by a new neural network method.

Traditional ED methods usually take into account the context of a mention in the text, e.g., ''inauguration'' helps to comprehend that ''Washington'' is related to the capital of the United States ''Washington, D.C.''. Some researchers have explored many methods to model the context information such as vector space model (Mendes et al., 2011), TF-IDF vector and topic feature model (Ratinov et al., 2011). The similarity between a mention and an entity can be measured based on these features. The entity with highest similarity score is considered as the most possible result. However, hand-crafted features are insufficient to capture the semantic information embedded in the context because it relies on domain knowledge. Neural network approaches have been attempting to learn the context representation without any manual design efforts and have achieved promising results. For example, the input is the entire document and the internal structure is added into the context representation by Stacked Denoising Auto-encoders (He et al., 2013). However, they could not find the important information from context words and only take consideration of some features of the mention or entity. In addition, existing approaches do not jointly consider multiple mentions in the same document (Sun et al., 2015). In fact, candidate entities within the same document are highly related. In this paper, we design a new collective entity disambiguation method based on hierarchical semantic similarity model (HSSM) and global features, named HSSMGF, to overcome the above-mentioned shortcomings. First, HSSM uses attention mechanism to choose important information from multiple information sources. Second, the selected information is merged, and attention mechanism is reapplied to generate hierarchical representations of mentions and entities. Third, global features, including prior popularity and global coherence, are utilized to improve the results of ED. To reduce computational cost, all other mentions in the same document needn't to be considered when disambiguating a mention. Instead, the coherence of entities can be computed simply based on unambiguous mentions with less noises. HSSMGF not only makes use of all the available information for mentions and entities but is robust when missing supporting information. Our main contributions are illustrated as follows.

- We present a hierarchical semantic similarity model that generates hierarchical representations of mentions and entities. Our model is designed to fully utilize different kinds of information about mentions and entities to capture semantic similarity. Attention mechanism is also used to pick up important information which is more relevant to the mention or entity, and it can improve semantic matching.
- The semantic similarities between mentions and entities are combined with global features. The proposed method is a simple and effective collective ED algorithm which balances the influence of all features to achieve the best result.
- To estimate the effect of HSSMGF, experimental studies are conducted over five publicly available datasets. The experimental results demonstrate that HSSMGF is superior to the state-of-the-art methods in most cases. Some visualization cases demonstrate the interpretability of our method.

The structure of the rest of this paper is as follows. Section 2 states related work. The disambiguation algorithm is detailed in Section 3, and experimental results are described in Section 4. Section 5 gives conclusions and suggestions for future work.

2. RELATED WORKS

There are many entities in the text, but the abbreviation or ambiguous expressions of entities prevent us from understanding the meaning of the text (Chen et al., 2016). Hence, to improve the effect of knowledge discovery, ED can be studied deeply. Common ED methods can be categorized into