WIKIDATA BASED PERSON NAME DISAMBIGUATION FOR USER PROFILING IN NEWS RECOMMENDATION

F. Shanaz[1*] and R.G. Ragel[2]

[1]*Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka*
[2]*Department of Computer Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka*
[*]*shanazlathif@gmail.com*

Entity linking (EL) is a process of extracting entity mentions in documents and linking them to their corresponding actual entities in a Knowledge Base (KB), such as Wikipedia or Wikidata. This task is challenging due to name variations, incompleteness of the KB and high ambiguity of entity mentions. News articles generally contain mentions of entities, such as persons, organizations and locations, which are an excellent resource for understanding readers' news interest. However, an entity mention can refer to different real-world entities. Furthermore, person entities are generally more ambiguous and critically important to understand the readers' news interest. This paper aims to design an EL system for disambiguating person mentions in news articles that can be used for user profiling in news recommendation. Current EL methods do not focus on improving EL performance of person entity mentions. Wikidata KB is chosen based on accessibility, completeness of the relations and timeliness of the data. The proposed method includes three main steps. In the first step, a new approach is proposed to generate candidate entities based on name dictionary-based technique. Partial string-matching technique is adopted when person names are referred as part of their full names. In the second step, top N candidates are selected using features from previous studies: entity popularity, textual similarity and the contextual similarity between a mention and the KB entity. In the last step, the best-matched entity is chosen from the top N candidates based on the semantic relatedness between the entities in a news article. The performance of the proposed methods is evaluated over a manually annotated AIDA-CoNLL news dataset. Experimental results show that the proposed candidate entity generation algorithm achieves the highest precision, and the recall of 98.91% on AIDA-CoNLL testb dataset. The best overall performance is achieved by correctly linking 90.05% of mentions in testb corpus. The results show that the proposed approach to generate candidate entities leads to achieve the highest precision at best matched entity selection stage. Thus, it can significantly benefit the user interest profiling in news recommendation.

Keywords: Entity linking, News recommendation, Person name disambiguation, User profiling, Wikidata