**Entity Linking: A Problem to Extract Corresponding Entity with Knowledge Base**

**Abstract**

Entity linking is a task to extract query mentions in documents and then link them to their corresponding entities in a knowledge base. It can improve the performances of knowledge network construction, knowledge fusion, information retrieval, natural language processing and knowledge base population. In this paper, we introduce the difficulties and applications of entity linking and focus on the main methods to address this issue. At last, we list the knowledge bases, datasets, the evaluation criterion and some challenges of entity linking.

**Architecture**

****

**ExistingSystem**

Entity linking also known as named entity disambiguation. The task of named entity linking refers to map named entity mentions in the text to their corresponding entities in a knowledge base. After overcoming the problem of recognizing an entity in a text, the attentions of researches are changed to entity disambiguation. As early as the coreference resolution and word sense disambiguation are the important tasks of Natural Language Processing. They deal with disambiguation of words in text so as to achieve the purpose of accurately understanding the meaning of the text. propose a method which uses the Wikipedia knowledge base to do entity disambiguation. Different from word sense disambiguation, it adds the treatment of proper nouns and considers the meanings of target entities. This work is considered to be one of the early works of entity linking.

**Proposed System**

We can help search engines to disambiguate so that the correct searching results are closer to the top of pages of the searching results. In the field of knowledge fusion, entity linking also plays an important role. When we fuse entities from different databases to a unified database, actually, some entities are the same with each other in the expressive meaning but have different surface forms, so it requests us to map these entities into the same entity in the knowledge base first. Then, the information of the entities can be fused. Meanwhile, the retrieval efficiency can be also improved. In addition, knowledge base population has become a popular topic in recent years. The mission of it is to extract new information scattered on the web and then fill the relevant entities into the existing knowledge base. For this purpose, the first phase is exactly to complete the entity linking task.

**Future Work**

It can promote the integration of knowledge among different languages. However, this task is limited by the problems of translation between languages, the portability of the models and gaining the large-scale training corpus, which lead to the low accuracy of multilingual entity linking. Thus, it is very meaningful to study entity linking among different languages. To improve the accuracy of entity linking, many systems exploit some complex models to address this problem. As compensation, they will have a higher time complexity. It can be regarded as a new challenge to balance accuracy and computing complexity in the future work.

**Module Implementation**

1. **Query mention**: the surface form of named entity in text. We also call it mention as alternative. Text: some query texts which contain many named entities. It can be divided into long texts and short texts. The long texts usually contain more than 400 words, such as news articles and the short texts contain an average of less than 200 words, such as tweets.
2. **The candidate entity generation module**: this module is to select the candidate entities for each query mention in text. It first uses named entity recognition tools to identify the entity mentions and then use the name of entities in conjunction with other features to find the candidate entities in the knowledge base. At last, we can get a series of related entities, for example, the candidate entities for Li Na including Li Na (tennis player), Li Na (professor of the Peking University) and Li Na (singer).
3. **The candidate entity disambiguation module**: this module is an important stage of entity linking. It utilizes different methods fusing various features of entities to rank the candidate entities. Features used in the candidate disambiguation phase include: entity popularity, entity type, the similarity between a query mention and the name of candidate entity, the similarity between context of query mention and candidate entity text, topic similarity and combination of several features. For example, exploit 27 different features in their article, and classify these features into three kinds of features: singleton, asymmetric feature and symmetric feature.
4. **The linking result module**: this module is to select the target entity through the ranking result of the candidate entity disambiguation phase. As the last module, when the score of all candidate entities is below the threshold, the system will return NIL. However, the threshold is usually set manually, which will easily lead to the problem that the correct target entity is judged to be below the threshold. Thus, some researches work for this issue. A machine learning method to classify the highest ranked candidate entity, judging whether the highest ranked entity is the target entity. When the highest ranked entity is classified as the positive instance, it will be regarded as the target entity, otherwise, there is no corresponding target entity mapping with the query mention. In another case, the candidate entity with the highest score is selected as the target entity among those candidate entities which satisfy the threshold.

**Algorithm**

1. **Hierarchical Agglomerative Clustering Algorithm**

Ward’s is the only one among the agglomerative clustering methods that is based on a classical sum-of-squares criterion, producing groups that minimize within-group dispersion at each binary fusion. In addition, Ward’s method is interesting because it looks for clusters in multivariate Euclidean space.This method is a common method to deal with documents clustering problem. The NIL clustering based on a hierarchical agglomerative clustering method usually initializes entities to several different clusters according to the mentions of entities and then merge entities in each clusters until the distance between clusters is smaller than the threshold.

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL

**Conclusion**

In this paper, we review and summarize the significance and application of entity linking. Then, we analyze some main methods of entity linking. At last, the knowledge base, datasets, and the evaluation criterion are described. Through the research of the entity linking technology, we put forward the future challenges as follows. The knowledge bases utilized in the entity linking systems are offline databases or extracted from the online database but lacking in automatic update mechanism. Therefore, many knowledge bases are incomplete because of the slow updating. It requires us to mine more evidences for the entity with little information.