ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Presence Combining With Familiarity Shoddy

Avinash Konduri
Assistant Professor,
Christu Jyothi Institute of Technology and Science.

N.Uma Rani
Assistant Professor,
Christu Jyothi Institute of Technology and Science.

K.Mamatha
Assistant Professor,
Christu Jyothi Institute of Technology and Science.

Abstract:

The presence of a large number of potential applications of the Internet data reduction with the knowledge of the rules has led to increased research liaison body. Connectivity is an important entity to connect the entity referred to in the text with the appropriate authorities in the knowledge base. Potential applications include data mining, information retrieval, knowledge base population. However, this difficult task due to the differences and the name of the entity mystery. In this study, an overview and analysis of the main methods of linking entity is presented, and the various applications and evaluation systems that are attached to the entity, and future directions are discussed.

I.INTRODUCTION:

The task of linking the entity is a challenge because of the different name and mystery entity. There may be multiple forms of surface named entities, such as full name, partial name, aliases and shortcuts and alternative spellings. For example, the name of the entity, “Cornell University” is the abbreviation “Cornell” and the entity called “the city of New York” has his nickname “The Big Apple”. Link entity institutions determine the correct mapping the entity system recalls the different forms of the surface. On the other hand, it can be said of the entity may indicate the various entities mentioned. For example, it can be said entity “Sun” to star in the center of the solar system, a company of multinational computer, which is a fictitious name character “Sun Hua Kwan” in the television series on ABC’s “Lost” or many other entities may be referred to as “the sun”. Entity link to disambiguate entity referred to in the context of the text and to determine the allocation of each entity referred system. I’ve done an exhaustive study of the entity relationship. Specifically, he surveyed the main methods used in the three units of the systems of the body are connected (any generation of candidates entity, the entity available candidate, and said Unlinkable predict), we also provide other important aspects of the entity, such as applications and connection functions and evaluation.

Although there are many methods proposed to address the link entity is not clear offline technologies and systems that are the current state of the art, since these systems differ in the length of the multiple dimensions and evaluated different sets of data. System entity and a link usually leads to different groups and different data fields. Although the means of settlement under the supervision appears to be doing much better than the focus is controlled with respect to a candidate entity classification, also it affected the overall performance of the system installation that joins significantly techniques adopted in the other two units (is namely the generation entity candidate, said prediction Unlinkable). Under observation techniques require many training examples and annotated examples of important signaling expensive. Moreover, the task of linking entity is data dependent and it is unlikely that technology controls all others in all data sets. A task link entity in particular is difficult to determine which technologies are best suited.

II.RELATED WORK:
A.Entity lining System:

Candidate entity generation. For each entity mention m 2 M, the entity linking system aims to filter out irrelevant entities in the knowledge base and retrieve a candidate entity set Em which contains possible entities that entity mention m may refer to. To achieve this goal, a variety of techniques have been utilized by some state-of-the-art entity linking systems, such as name dictionary based techniques, surface form expansion from the local document, and methods based on search engine. Candidate entity ranking. In most cases, the size of the candidate entity set Em is larger than one. Researchers leverage different kinds of evidence to rank the candidate entities in Em and try to find the entity e 2 Em which is the most likely link for mention m. To deal with the problem of predicting unlinkable mentions, some work leverages this module to validate whether the topranked entity identified in the Candidate Entity Ranking module is the target entity for mention m. Otherwise, they return NIL for mention m. An overview of the main approaches for predicting unlinkable mentions.

B.Candidate entity generation:

Entity Generation module, for each entity mention m 2 M, entity linking systems try to include possible entities that entity mention m may refer to in the set of candidate entities Em. Approaches to candidate entity generation are mainly based on string comparison between the surface form of the entity mention and the name of the entity existing in a knowledge base. This module is as important as the Candidate Entity Ranking module and critical for a successful entity linking system according to the experiments conducted by Hachey et al. [33]. In the remainder of this section, we review the main approaches that have been applied for generating the candidate entity set Em for entity mention m.

C.Name Dictionary Based Techniques:

Name dictionary based techniques are the main approaches to candidate entity generation and are leveraged by many entity linking systems.The structure of Wikipedia provides a set of useful features for generating candidate entities, such as entity pages, redirect pages, disambiguation pages, bold phrases from the first paragraphs, and hyperlinks in Wikipedia articles. These entity linking systems leverage different combinations of these features to build an offline name dictionary D between various names and their possible mapping entities, and exploit this constructed name dictionary D to generate candidate entities. This name dictionary D contains vast amount of information about various names of named entities, like name variations, abbreviations, confusable names, spelling variations, nicknames, etc. Specifically, the name dictionary D is a hkey, valuei mapping, where the key column is a list of names. Suppose k is a name in the key column, and its mapping value k:value in the value column is a set of named entities which could be referred to as the name k. The dictionary D is constructed by leveraging features from Wikipedia as follows:

i.Entity pages. Each entity page in Wikipedia describes a single entity and contains the information focusing on this entity. Generally, the title of each page is the most common name for the entity described in this page, e.g., the page title “Microsoft” for that giant software company headquartered in Redmond. Thus, the title of the entity page is added to the key column inD as a name k, and the entity described in this page is addedas k:value.

ii.Redirect pages. A redirect page exists for each alternative name which could be used to refer to an existing entity in Wikipedia. For example, the article titled “Microsoft Corporation” which is the full name of Microsoft contains a pointer to the article of the entity Microsoft. Redirect pages often indicate synonym terms, abbreviations, or other variations of the pointed entities. Therefore, the title of the redirect page is added to the key column in D as a name , and the pointed entity is added as k:value.

iii.Disambiguation pages. When multiple entities in Wikipedia could be given the same name, a disambiguation page is created to separate them and contains a list of references to those entities. For example, the disambiguation page for the name “Michael Jordan” lists thirteen associated entities having the same name of “Michael Jordan” including the famous NBA player and the Berkeley professor. These disambiguation pages are very useful in extracting abbreviations or other aliases of entities. For each disambiguation page, the title of this page is added to the key column in D as a name k, and the entities listed in this page are added as k:value.

iv.Bold phrases from the first paragraphs. In general, the first paragraph of a Wikipedia article is a summary of the whole article. It sometimes contains a few phrases written in bold. Varma et al. observed that these bold phrases invariably are nick names, alias names or full names of the entity described in this paper. For instance, in the first paragraph of the entity page of Hewlett-Packard (HP), there are two phrases written in bold (i.e., “Hewlett-Packard Company” and “HP”) which are respectively the full name and the abbreviation for the entity Hewlett-Packard. Thus, for each of the bold phrases in the first paragraph of each Wikipedia page, it is added to the key column in D as a name k, and the entity described in this page is added as k:value.

v.Hyperlinks in Wikipedia articles. An article in Wikipedia often contains hyperlinks which link to the pages of the entities mentioned in this article. The anchor text of a link pointing to an entity page provides a very useful source of synonyms and other name variations of the pointed entity, and could be regarded as a name of that linked entity. For example, in the entity page of Hewlett-Packard, there is a hyperlink pointing to the entity William Reddington Hewlett whose anchor text is “Bill Hewlett”, which is an alias name of the entity William Reddington Hewlett.

Hence, the anchor text of the hyperlink is added to the key column in D as a name k, and the pointed entity is added as k:value. Using these features from Wikipedia described above, entity linking systems could construct a dictionary D. Besides leveraging the features from Wikipedia, there are some studies that exploit query click logs and web documents to find entity synonyms, which are also helpful for the name dictionary construction.

D.Surface Form Expansion from the Local Document:

Since some entity mentions are acronyms or part of their full names, one category of entity linking systems use the surface form expansion techniques to identify other possible expanded variations (such as the full name) from the associated document where the entity mention appears. Then they could leverage these expanded forms to generate the candidate entity set using other methods such as the name dictionary based techniques introduced above. We categorize the surface form expansion techniques into the heuristic based methods and the supervised learning methods.

E.Candidate entity ranking:

In the previous section, we described methods that could generate the candidate entity set Em for each entity mention m. We denote the size of Em as jEmj, and use 1 _ i _ jEmj to index the candidate entity in Em. The candidate entity with index i in Em is denoted by ei. In most cases, the size of the candidate entity set Em is larger than one. For instance, Ji et al. [89] showed that the average number of candidate entities per entity mention on the TAC-KBP2010 data set is 12.9, and this average number on the TAC-KBP2011 data set is 13.1. In addition, this average number is 73 on the CoNLL data set utilized in [58]. Therefore, the remaining problem is how to incorporate different kinds of evidence to rank the candidate entities in Em and pick the proper entity from Em as the mapping entity for the entity mention m. The Candidate Entity Ranking module is a key component for the entity linking system. We can broadly divide these candidate entity ranking methods into two categories:

i.Supervised ranking methods. These approaches rely on annotated training data to “learn” how to rank the candidate entities in Em. These approaches include binary classification methods, learning to rank methods, probabilistic methods, and graph based approaches.

ii.Unsupervised ranking methods. These approaches are based on unlabeled corpus and do not require any manually annotated corpus to train the model. These approaches include vector space model (VSM) based methods and information retrieval based methods. In this section, all candidate entity ranking methods are illustrated according to the above categorization. In addition, we could also categorize the candidate entity ranking methods into another three categories:

iii.Independent ranking methods. These approaches consider that entity mentions which need to be linked in a document are independent, and do not leverage the relations between the entity mentions in one document to help candidate entity ranking. In order to rank the candidate entities, they mainly leverage the context similarity between the text around the entity mention and the document associated with the candidate entity.

iv.Collective ranking methods. These methods assume that a document largely refers to coherent entities from one or a few related topics, and entity assignments for entity mentions in one document are interdependent with each other. Thus, in these methods, entity mentions in one document are collectively linked by exploiting this “topical coherence”.

v.Collaborative ranking methods. For an entity mention that needs to be linked, these approaches identify other entity mentions having similar surface forms and similar textual contexts in the other documents. They leverage this cross-document extended context information obtained from the other similar entity mentions and the context information of the entity mention itself to rank candidate entities for the entity mention.

CONCLUSION:

In this paper, we have presented a comprehensive survey for entity linking. Specifically, we have surveyed the main approaches utilized in the three modules of entity linking systems (i.e., Candidate Entity Generation, Candidate Entity Ranking, and Unlinkable Mention Prediction), and also introduced other critical aspects of entity linking such as applications, features, and evaluation.

Although there are so many methods proposed to deal with entity linking, it is currently unclear which techniques and systems are the current state-of-the-art, as these systems all differ along multiple dimensions and are evaluated over different data sets. A single entity linking system typically performs very differently for different data sets and domains. Although the supervised ranking methods seem to perform much better than the unsupervised approaches with respect to candidate entity ranking, the overall performance of the entity linking system is also significantly influenced by techniques adopted in the other two modules (i.e., Candidate Entity Generation and Unlinkable Mention Prediction). Supervised techniques require many annotated training examples and the task of annotating examples is costly. Furthermore, the entity linking task is highly data dependent and it is unlikely a technique dominates all others across all data sets. For a given entity linking task, it is difficult to determine which techniques are best suited. There are many aspects that affect the design of the entity linking system, such as the system requirement and the characteristics of the data sets. Although our survey has presented many efforts in entity linking, we believe that there are still many opportunities for substantial improvement in this field. In the following, we point out some promising research directions in entity linking.

REFERENCES:

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives, “DBpedia: A nucleus for a web of open data,” in Proc. 6th Int. Semantic Web 2nd Asian Conf. Asian Semantic Web Conf., 2007, pp. 11–15.

[2] F. M. Suchanek, G. Kasneci, and G. Weikum, “Yago: A core of semantic knowledge unifying wordnet and wikipedia,” in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 697–706.

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: A collaboratively created graph database for structuring human knowledge,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2008, pp. 1247–1250.

[4] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. ates, “Web-scale information extraction in knowitall: (preliminary results),” in Proc. 13th Int. Conf. World Wide Web, 2004, pp. 100–110.

[5] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, Jr, and T. M. Mitchell, “Coupled semi-supervised learning for information extraction,” in Proc. 3rd ACM Int. Conf. Web Search Data Mining, 2010, pp. 101–110.

[6] W. Wu, H. Li, H. Wang, and K. Q. Zhu, “Probase: A probabilistic taxonomy for text understanding,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2012, pp. 481–492.

[7] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web,” Sci. Am., vol. 284, pp. 34–43, 2001.

[8] E. Agichtein and L. Gravano, “Snowball: Extracting relations from large plain-text collections,” in Proc. ACM Int. Conf. Digital Libraries, 2000, pp. 85–94.

[9] D. Zelenko, C. Aone, and A. Richardella, “Kernel methods for relation extraction,” J. Mach. Learn. Res., vol. 3, pp. 1083–1106, 2003.

[10] T. Hasegawa, S. Sekine, and R. Grishman, “Discovering relations among named entities from large corpora,” in Proc. 42nd Ann. Meeting Assoc. Comput. Linguistics, 2004, pp. 415–422.