International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July 2016

Knowledge based systems text analysis

Vatan Choudhary[1] & Dipesh kumar Sharma[2]
[1]M-Tech Scholar Dept. of CSE RIT, Raipur (C.G)
[2]H.o.D Dept. of CSE RIT, Raipur (C.G)

Abstract

The astronomically immense number of potential applications from bridging Web data with cognizance bases has led to an incrementation in the entity linking research. Entity linking is the task to link entity mentions in text with their corresponding entities in a knowledgebase. Potential applications include information extraction, information retrieval, and cognizance base population. However, this task is challenging due to designate variations and entity ambiguity. In this survey, we present an exhaustive overview and analysis of the main approaches to entity linking, and discuss sundry applications, the evaluation of entity linking systems, and future directions.

Keywords: knowledge based systems, text analysis.

1. Introduction

Tidally and the Web has become one of the most sizably voluminous HE amount of Web data has incremented exponent data repositories in the world in recent years. Plenty of data on the Web is in the form of natural language. However, natural language is highly equivocal, especially with reverence to the frequent occurrences of denominated entities. A denominated entity may have multiple names and a designation could denote several different denominated entities. On the other hand, the advent of cognizance sharing communities such as Wikipedia and the development of information extraction techniques have facilitated the automated construction of astronomically immense scale machine-readable cognizance bases. Cognizance bases contain affluent information about the world's entities, their semantic classes, and their mutual relationships. Such kind of eminent examples include DBpedia YAGO Freebase Ken tall Read the Web and Probase. Bridging Web data with cognizance bases is propitious for annotating the plethora of raw and often noisy data on the Web and contributes to the vision of Semantic Web. A critical step to achieve this goalis to link

denominated entity mentions appearing in Web text with their corresponding entities in a knowledgebase, which is called entity linking. Entity linking can facilitate many different tasks such as cognizance base population, question answering, and information integration. As the world evolves, incipient facts are engendered and digitally expressed on the Web. Consequently, enriching subsisting cognizance bases utilizing incipient facts becomes increasingly consequential. However, inserting incipiently extracted cognizance derived from the information extraction system into a subsisting cognizance base ineluctably needs a system to map an entity mention associated with the extracted erudition to the corresponding entity in the cognizance base. For example, cognation extraction is the process of discovering utilizable relationships between entities mentioned in text and the extracted cognation requires the process of mapping entities associated with the cognation to the erudition base afore it could be populated into the cognizance base. Furthermore, an immensely colossal number of question answering systems rely on their fortified cognizance bases to give the answer to the user's question. To answer the question "What is the birth date of the famous basketball player Michael Jordan?", the system should first leverage the entity linking technique to map the queried "Michael Jordan" to the NBA player, in lieu of for example, the Berkeley edifier; and then it retrieves the birth date of the NBA player designated "Michael Jordan" from the cognizance base directly. Adscititiously, entity linking avails puissant join and coalescence operations that can integrate information about entities across different pages, documents, and sites.

The entity linking task is challenging due to designate variations and entity ambiguity. A denominated entity may have multiple surface forms, such as its full denomination, partial denominations, aliases, abbreviations, and alternate spellings. For example, the designated entity of "Cornell University" has its abbreviation "Cornell" and the denominated entity of "New York City" has its moniker "Big Apple". An entity linking system has to identify the correct mapping entities for entity mentions of sundry surface forms. On the other hand, an entity mention could possibly denote different denominated entities. For instance, the entity mention "Sun" can refer to the star at the center of the Solar System, a multinational computer company, a fictional character named "Sun-Hwa Kwon" on the ABC television series "Lost" or

many other entities which can be referred to as" Sun". An entity linking system has to disambiguate the entity mention in the textual context and identify the mapping entity for each entity mention.

Task Description

Given an erudition base containing a set of entities E and a text amassment in which a set of denominated entity mentions M are identified in advance, the goal of entity linking is to map each textual entity mention ∈ M to its corresponding entity e ∈ E in the cognizance base. Here, a designated entity mention mis a token sequence in text which potentially refers to some denominated entity and is identified in advance. It is possible that some entity mention in text does not have its corresponding entity record in the given erudition base. We define this kind of mentions asunlinkable mentions and give NIL as a special label denoting "un linkable". Ergo, if the matching entity e for entity mention m does not subsist in the cognizance base (i.e., e / ∈ E), an entity linking system should label m as NIL. For un linkable mentions, there are some studies that identify their fine-grained types from the cognizance predicate which is out of scope for entity linking systems. Entity linking is adscititiously called Denominated Entity Disambiguation (NED) in the NLP community. In this paper, we just fixate on entity linking for English language, rather than cross lingual entity linking .Typically, the task of entity linking is preceded by a designated entity apperception stage, during which boundaries of denominated entities in text are identified. While designated entity apperception is not the focus of this survey, for the technical details of approaches utilized in the designated entity apperception task, you could refer to the survey paper and some concrete methods. In integration, there are many publicly available denominated entity apperception implements, such as Stanford NER1, OpenNLP2, and LingPipe3. Finke let al. introduced the approach utilized in Stanford NER. They leveraged Gibbs sampling to augment ant subsisting Conditional Desultory Field predicated system with long-distance dependency models, enforcing label consistency and extraction template consistency.

2. Related Work

The information extraction system into a subsisting cognizance base ineluctably needs a system to map an entity mention associated with the extracted cognizance to the corresponding entity in the cognizance base. On the other hand,

an entity mention could possibly denote different designated entities. For instance, the entity mention "Sun" can refer to the star at the center of the Solar System, a multinational computer company, a fictional character named "Sun-Hwa Kwon" on the ABC television series "Lost "or many other entities which can be referred to as "Sun". An entity linking system has to disambiguate the entity mention in the textual context and identify the mapping entity for each entity mention.

Literature Survey

A Survey on" DBpedia: A Nucleus for a Web of Open Data"

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. We describe the extraction of the DBpedia datasets, and how the resulting information is published on the Web for human- and machine consumption. We describe some emerging applications from the DBpedia community and show how website authors can facilitate DBpedia content within their sites. Finally, we present the current status of interlinking DBpedia with other open datasets on the Web and outline how DBpedia could serve as a nucleus for an emerging Web of open data.

DBpedia is a major source of open, royalty-free data on the Web. We hope that by interlinking DBpedia with further data sources, it could serve as a nucleus for the emerging Web of Data.

A Survey on "Coupled semi-supervised learning for information extraction,"

We consider the problem of semi-supervised learning to extract categories (e.g., academic fields, athletes) and relations (e.g., Plays Sport(athlete, sport)) from web pages, starting with a handful of labeled training examples of each category or relation, plus hundreds of millions of unlabeled web documents. Semi-supervised training using only a few labeled examples is typically unreliable because the learning task is under constrained. This paper pursues the thesis that much greater accuracy can be achieved by further constraining the learning task, by coupling the semi-supervised training of many extractors for different categories and relations. We characterize several ways in which the training of category and relation extractors can be coupled, and present experimental results demonstrating significantly improved accuracy as a result.

We have presented methods of coupling the semi supervised learning of category and relation instance extractors and demonstrated empirically that coupling forestalls the problem of semantic drift associated with bootstrap learning methods. This empirical evidence leads us to advocate large-scale coupled training as a strategy to significantly improve accuracy in semi-supervised learning.

A Survey on "Snowball: Extracting relations from large plain-text collections"

Text documents often contain valuable structured data that is hidden in regular English sentences. This data is best exploited if available as a relational table that we could use for answering precise queries or for running data mining tasks. We explore a technique for extracting such tables from document collections that requires only a handful of training examples from users. These examples are used to generate extraction patterns, that in turn result in new tuples being extracted from the document collection. We build on this idea and present our Snowball system. Snowball introduces novel strategies for generating patterns and extracting tuples from plain-text documents. At each iteration of the extraction process, Snowball evaluates the quality of these patterns and tuples without human intervention, and keeps only the most reliable ones for the next iteration. In this paper we also develop a scalable evaluation methodology and metrics for our task, and present a thorough experimental evaluation of Snowball and comparable techniques over a collection of more than 300,000 newspaper documents.

This paper presents Snowball, a system for extracting relations from large collections of plain-text documents that requires minimal training for each new scenario. We introduced novel strategies for generating extraction patterns for Snowball, as well as techniques for evaluating the quality of the patterns and tuples generated at each step of the extraction process. Our large-scale experimental evaluation of our system shows that the new techniques produce high-quality tables, according to the scalable evaluation methodology that we introduce in this paper. Our experiments involved over 300,000 newspaper articles.

A Survey on "Discovering relations among named entities from large corpora"

Discovering the significant relations embedded in documents would be very useful not only for information retrieval but also for question answering and summarization. Prior methods

for relation discovery, however, needed large annotated corpora which cost a great deal of time and effort. We propose an unsupervised method for relation discovery from large corpora. The key idea is clustering pairs of named entities according to the similarity of context words intervening between the named entities. Our experiments using one year of newspapers reveals not only that the relations among named entities could be detected with high recall and precision, but also that appropriate labels could be automatically provided for the relations.

We proposed an unsupervised method for relation discovery from large corpora. The key idea was clustering of pairs of named entities according to the similarity of the context words intervening between the named entities. The experiments using one year's newspapers revealed not only that the relations among named entities could be detected with high recall and precision, but also that appropriate labels could be automatically provided to the relations. In the future, we are planning to discover less frequent pairs of named entities by combining our method with bootstrapping as well as to improve our method by tuning parameters.

3. Methodology

After careful analysis the system has been identified to have the following modules:

1. **Entity linking**

2. Knowledge base

3. Candidate Entity Ranking.

1. Entity linking

Entity linking can facilitate many different tasks such as cognizance base population, question answering, and information integration. As the world evolves, incipient facts are engendered and digitally expressed on the Web. Ergo, enriching subsisting erudition bases utilizing incipient facts becomes increasingly paramount. However, inserting incipiently extracted erudition derived from the information extraction system into a subsisting cognizance base ineluctably needs a system to map an entity mention associated with the extracted cognizance to the corresponding entity in the erudition base. Adscititiously, entity linking avails puissant join and coalescence operations that can integrate information about entities across different pages, documents, and sites. The entity linking task is challenging due to denominate variations and entity ambiguity.

2. Candidate Entity Generation

Candidate Entity Generation module, for each entity mention m ∈ M, entity linking systems endeavor 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 predicated on string comparison between the surface form of the entity mention and the designation of the entity subsisting in an erudition base. This module is as paramount as the Candidate Entity Ranking module and critical for a prosperous entity linking system according to the experiments conducted by Hachey et al. In the remnant of this section, we review the main approaches that have been applied for engendering the candidate entity set Em for entity mention m.

3. Candidate Entity Ranking

In most cases, the size of the candidate entity set me is more sizably voluminous than one. Researchers leverage different kinds of evidence to rank the candidate entities in Me and endeavor to find the entity e ∈Which is the most likely link for mention m. Inspection we will review the main techniques utilized in this ranking process, including supervised ranking methods and To deal with the quandary of presaging unlinkablementions, some work leverages this module to validate whether the top-ranked entity identified in the Candidate Entity Ranking module is the target entity for mention m. Otherwise, they return NIL for mention m. In, we will give an overview of the main approaches for presaging unlikable mentions. For this ranking we utilize Vector Space Model.

Methods Based on Search Engines:

Some entity linking systems [61,69,73,83] try to leverage the whole Web information to identify candidate entities via Web search engines (such as Google). Specifically, Han and Zhao [61] submitted the entity mention together with its short context to the Google API and obtained only Web pages within Wikipedia to regard them as candidate entities. Dredze et al. [83] queried the Google search engine using the ntity mention and identified candidate entities whose Wikipedia pages appear in the top 20 Google search results for the query. Lehmann et al. [69] and Monahan et al. [73] stated that the Google search engine is very effective at identifying some of the very difficult mappings between surface forms and entities. They performed the query using the Google API limited to the English Wikipedia site and filtered results whose Wikipedia titles are not significantly Dice or acronym based similar to

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July 2016

the query. Lastly, they utilized the top three results as candidate entities.

In addition, Wikipedia search engine is also exploited to retrieve candidate entities which can return a list of relevant Wikipedia entity pages when you query it based on keyword matching. Zhang et al. [65] utilized this feature to generate infrequently mentioned candidate entities by querying this search engine using the string of the entity mention.

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 |Em|, and use $1 \leq i \leq |Em|$ 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 (TAC-KBP tracks and data sets will be introduced in Section 5.2) 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 utilizedin [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 rankingmethods into two categories:

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.

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.

4. Result and Discussion

Fig 2: Search Result Page.

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July 2016

Fig 3: View Total Article Page.

The large number of features introduced here reflects the large number of aspects an entity linking system could consider when dealing with the entity linking task. Unfortunately, there are very few studies that compare the effectiveness of the various features presented here. However, we emphasize that no features are superior to others over all kinds of data sets. Even some features that demonstrate robust and high performance on some data sets could perform poorly on others. Hence, when designing features for entity linking systems, the decision needs to be made regarding many aspects, such as the tradeoff between accuracy and efficiency, and the characteristics of the applied data set.

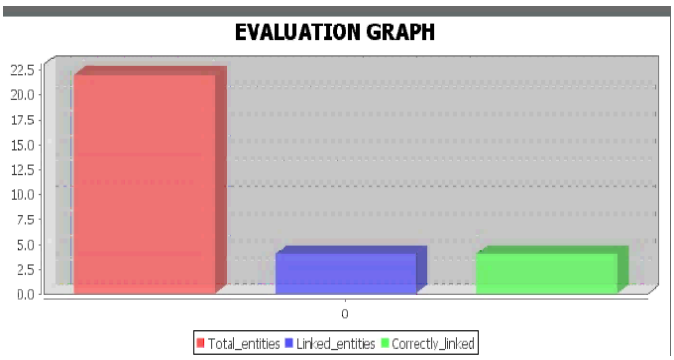

Fig 5: Result Evolution Graph.

Graph Based Approaches

In the meantime, Hoffart et al. also proposed a graph based approach for collective entity linking. This model combines three features into a graph model: entity popularity, textual context similarity as well as coherence between mapping entities. They built a mention-entity graph, a weighted and undirected graph with entity mentions and candidate entities as nodes. In this mention-entity graph, a mention-entity edge is weighted with a combination of the entity popularity feature and the textual context similarity feature, and an entity-entity edge is weighted by the Wikipedia hyperlink structure based coherence (see Section 3.1.2.2). Given this constructed graph, their goal is to compute a dense subgraph that contains exactly one mention-entity edge for each entity mention. However, this problem is NP-hard as it generalizes the well studied Steiner-tree problem. To solve this problem, Hoffart et al. developed a greedy algorithm with the extension of the algorithm proposed. The experimental results show that it outperforms the collective entity linking system and the approach of Cucerzan , and achieves 81.8% accuracy over their own CoNLL data set.

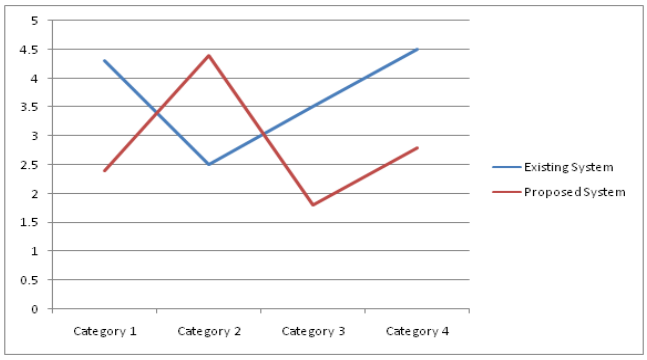

Fig 6: Graphical Representation.

5. Conclusion

Although our survey has presented much effortsin 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. Firstly, most of the current entity linking systems focuses on the entity linking task where entity mentions are detected from unstructured documents (such as news articles and blogs). However, entity mentions may also appear in other types of data and these types of data also need to be linked with the knowledgebase, such as Web tables Web lists and tweets. As different types of data have various characteristics (e.g., Web tables are semi structured text and have almost no textual context, and tweets are very short and noisy), it is very meaningful and necessary to develop specific techniques to deal with linking entities in them. Although some researchers have preliminarily addressed the entity linking task in Web tables Web lists and tweets respectively, we believe there is still much room for further improvement. Moreover, a repository of benchmark data sets with these different types should be made available to researcher's in order for them to develop and evaluate their methods for linking entities in these diverse types of data.

7. References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in ISWC, 2007,pp. 11–15.

[2] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A coreof semantic knowledge unifying wordnet and wikipedia," inWWW, 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 SIGMOD, 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. Yates, "Web-scaleinformation extraction in knowitall: (preliminary results)," inWWW, 2004, pp. 100–110.

[5] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, Jr.,and T. M. Mitchell, "Coupled

semi-supervised learning forinformation extraction," in WSDM, 2010, pp. 101–110.

[6] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: a probabilistic taxonomy for text understanding," in SIGMOD, 2012,pp. 481–492.

[7] T. Berners-Lee, J. Hendler, and O. Lassila, "The semanticweb," Scientific American, 2001.

[8] E. Agichtein and L. Gravano, "Snowball: Extracting relationsfrom large plain-text collections," in ICDL, 2000, pp. 85–94.

[9] D. Zelenko, C. Aone, and A. Richardella, "Kernel methodsfor 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 ACL,2004, pp. 415–422.

[11] W. Shen, J. Wang, P. Luo, M. Wang, and C. Yao, "Reactor: a framework for semantic relation extraction and tagging over enterprise data," in WWW, 2011, pp. 121–122.

[12] W. Shen, J. Wang, P. Luo, and M. Wang, "A graph-based approach for ontology population with named entities," inCIKM, 2012, pp. 345–354.