**Collective List-Only Entity Linking:**

**A Graph-Based Approach**

**Abstract**

List-only entity linking (EL) is the task of mapping ambiguous mentions in texts to target entities in a group of entity lists. Different from traditional EL task, which leverages rich semantic relatedness in knowledge bases to improve linking accuracy, the list-only EL can merely take advantage of co-occurrences information in entity lists. State-of-the-art work utilizes co-occurrences information to enrich entity descriptions which are further used to calculate local compatibility between mentions and entities to determine results. Nonetheless, entity coherence is also deemed to play an important part in EL, which is yet currently neglected. In this paper, in addition to local compatibility, we take into account global coherence among entities. Specifically,we propose to harness co-occurrences in entity lists for mining both explicit and implicit entity relations. The relations are then integrated into an entity graph, on which personalized PageRank is incorporated to compute entity coherence. The final results are derived by combining local mention-entity similarity and global entity coherence. The experimental studies validate the superiority of our method. Our proposal not only improves the performance of the list-only EL, but also opens up the bridge between the list-only EL and conventional EL solutions.

**Architecture**

****

**Existing system**

Traditional EL methods leverage knowledge bases (KBs), which offer rich semantic information of entities, for robust and accurate disambiguation process. Nevertheless, despite the effectiveness of knowledge-based EL, it might not be applicable in situations where there is insufficient information of entities, such as entity lists. Entity lists can be found useful, for instance, in the scenario concerning detection of emerging stock names. When investors search new stock names in Wikipedia,a frequently updated KB, chances are that there are no corresponding items.

**Proposed system**

The dataset used for empirical study might be inappropriate and need a redesign. Current dataset is comprised of documents, which contain mentions to be disambiguated, and a group of entity lists, which include the true entities for mentions. However, each document only contains a single mention for disambiguation, which may not reflect the reality well. Where there are four mentions in one document. Additionally, the entities in different entity lists are dissimilar, making the task much easier to cope with since each mention may well only have one candidate entity. This also deviates from reality and simplifies the problem. In short, the shortcomings of the existing list-only EL solution are two-fold: Entity coherence within or across entity lists was overlooked and not leveraged; and Results were supportless for lack of appropriate dataset and deliberate experiment design. We close the gap and address the deficiencies in this article. In particular, we propose to solve list-only EL task by taking account of the correlations in entities and converting the disambiguation problem to a graph problem.

**Future work**

For future work, we plan to investigate two aspects. One is to consider the situation where an entity appears in more than one entity list. For instance,Washington, D.C. can appear in entity lists featured American Cities and Country Capitals. Another possible research direction is utilizing word embedding techniques and deep neural networks to better model mention-entity compatibility and entity coherence. Specifically, leveraging welltrainedword embedding vectors as inputs, Long Short-Term Memory (LSTM) with attention mechanism could be used to summarize semantic meanings of the contexts around mentions and the representative texts of entities, which can be further harnessed to calculate more accurate compatibility score.

**Module implementation**

1. **Pre Processing**

In the pre-processing step, mentions in the text are detected and the candidate entities are also generated. Specifically, the initial input for EL is a set of raw documents, either with specified mentions to be disambiguated or without. Under the circumstance where mentions are not pointed out, Named Entity Recognition (NER) should be harnessed to finish the mention detection task. State-of-the-art NER methods utilize Neutral Networks and Deep Learning techniques to achieve better performances, whereas they have not been widely used yet on account of the freshness and complexity. Instead, Stanford NER Tagger, a NER tool which is less accurate but maturer, embraces higher popularity in tasks involving but not focusing on NER. In our experiment, we have already extracted the mentions during dataset construction process.

1. **Entity Information Enrichment**

Solely relying on co-occurrences between entities is not enough to establish relations among entities, let alone semantically bridge mentions with candidate entities. Therefore, we enrich information on entity side by selecting representatives derived from input documents.

1. **Graph Construction**

Through the pre-processing step, mentions and their candidate entities are obtained. Then after enriching textual descriptions in the entity side, the compatibility score between each mention and corresponding candidate entity can be calculated in terms of text similarity. Previous listonly EL ranked the candidate entities for each mention merely based on mention-entity compatibility scores, thereby producing the results accordingly. We argue that the judgement simply depending on compatibility score is not convincing enough because the coherence among entities is ignored, which plays an indispensable role in the linking process.

1. **Mention Entity Matching Pairs**

Given a weighed entity graph of document, the target is to find the most likely entity from a group of entities for each mention in document. In line with popular methods proposed in KB-oriented, we propose graph-based list-only entity linking algorithm, namely Gloel, which utilizes Personalized PageRank to depict the coherence among candidate entities. Specifically, we assign a vector with length n to each node to represent the results of a PageRank process starting. To better capture the coherence among entities within the same document, instead of regarding the similarity between the vectors of nodes as the coherence score, we define it as how a candidate entity fits in the document. To enable the definition, a n-length vector p(di) is also assigned to document di, representing the results of the PageRank process initiating from a group of unambiguous nodes.

**Algorithm**

1. **Graph-BasedAlgorithm:**Gloel, which implements Personalized PageRank todetermine how likely an entity is the target entity by takinginto consideration both coherences among entities, and compatibilitiesbetween mentions and entities. The outputs are alist of pairs comprised of mentions and their most possibleentities.
2. **PageRank Algorithm:**We first elaborate the random walk process initiating froma single node, then extend it to calculating document PageRank vector. The PageRank algorithm, based on random walktheory, is firstly proposed to measure the importance of webpages by counting the number and quality of links to thispage. It has been applied to EL problems in recent years, andhas achieved great performance [3]\_[6]. The basic elementsof PageRank include initial vector *r*0, transition matrix *A*,and preference vector *s*. Note that in our method, *r*0 D *s*.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

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

**Conclusion**

List-only entity linking task, as a new form of traditional EL problem, distinguishes itself by the sparse information on the entity side. In this work, on the one hand, we propose to utilize entity co-occurrences information to mine both textual description of entities and relations among entities, so as to enrich entity information. On the other hand, inspired by conventional EL methods, we construct an entity graph to capture relations among entities, on which the newly proposed algorithm Gloel is applied to obtain results. Similar to the situation in traditional EL, our approach, a collective EL method based on graph, outperforms independent EL on the dataset we create for fair comparison.