

Graph-based Entity Linking using Shortest Path

Yongsun Shim¹, Sungkwon Yang¹, Hyunwhan Joe¹, Hong-Gee Kim¹

¹ Biomedical Knowledge Engineering Laboratory,
Seoul National University, Seoul, Korea
{yongsun0926, sungkwon.yang, hyunwhanjoe, hgkim}@snu.ac.kr

Abstract. Entity Linking (EL) is a technique to link named entities in a given context to relevant entities in a given knowledge base. Generally, EL consists of two important task. But, there are limitations of these tasks. To overcome these limitations, we tried to solve the problem of EL through the interdependencies between not only named entities but also common nouns and verbs appearing in the context. The second approach is that the words appeared in context are more closely related and the distance between those is closer. In this paper, we proposed the approaches to overcome these limitations.

Keywords: Entity Linking, Graph-based knowledge base, Shortest Path.

1 Introduction

Entity Linking (EL), which is one of the various natural language processing methods, is also known as Named Entity Disambiguation (NED). EL is a technique to link named entities in a given context to relevant entities in a given knowledge base. Here, the context means natural language sentences we want to analyze. EL is used in various fields such as information extraction, semantic search, and question answering.

Generally, EL consists of two important task. First one is candidate entity look up task which lists some relevant entities in the given knowledge base that corresponds to the named entity of interest. Second task is called candidate ranking which ranks the entities by relevance. There are two main approaches for candidate ranking. The first one is local compatibility based approach. Local compatibility approaches compares similarity between the context and descriptions of each candidate entity. The limitation of this approach is that there is no consideration of relations between every name mentions appeared in the context. The second one is pairwise based approaches. Pairwise based approaches considers interdependencies among the candidate entities of named entities appearing in the context. However, in these approaches only computes similarity between two candidate entities. Sometimes it lies wrong result, because of the lack of direct relation between two candidate entities, even though they are indirectly connected. In other words, there is a shortage of global interdependence.

To overcome these limitations, we tried to solve the problem of EL through the interdependencies between not only named entities but also common nouns and verbs appearing in the context.

2 Our Approach

In our approach, a context means a sentence. Our approach tries to use every words in a context. Every words means that not only the named entity but also including general nouns and verbs which are appeared in context. This approach differs from the traditional approaches which use only the named entity. In this approach, our assumption is that every words appeared in a context is related to each other.

The second approach tries to use ordering of every words. This approach means that every words appeared in context has association between words. In other words, when a word appears in context, the words which related it will be closely appeared in context.

The third approach expand these approaches to graph-based knowledge base. It means that vertices which related each other are closely located in graph-based knowledge base. That is, relevant vertices will be located close each other in a knowledge base such as related words are appeared in a context.

By using our approaches, we can apply to find shortest path for solving EL in graph-based knowledge base. That is, if the words appear at the same time in a context, and the distance between words is close, the distance between the words and the vertices mapped in the graph-based knowledge base will be close.

3 Workflow

In this paper, we tried to use all the information of words used in a context to prevent information loss. And we considered the ordering of words in a context, assuming that there is a close relationship between each other words. Thus, in a context, every words appeared by pre-processing are grouped into a sequence. We then applied this sequence to a pre-constructed weighted-graph to apply our approaches.

This chapter describes the workflow of this paper. The workflow of this paper is shown in Fig1. The workflow is divided into 5 steps.

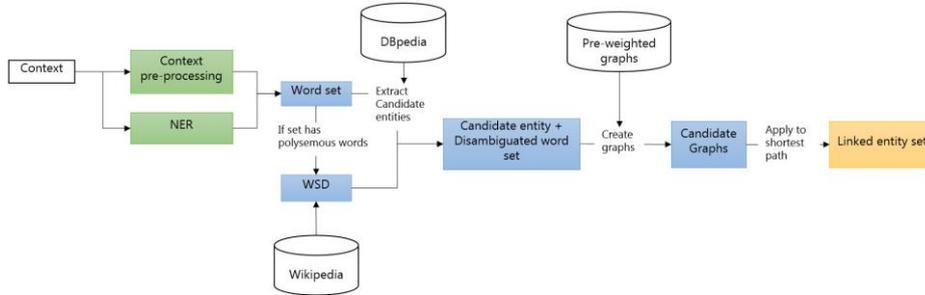


Fig. 1. A workflow of our system.

3.1 Create to Pre-weighted Graph

The first step is to assign a weight to the graph-based knowledge base. This step is performed only once at first. When weights are given to the entire graph, we will target the edges of triangles and corresponding vertices. The reason for using the triangle shape is that we start from the assumption that we can know the meaning of a specific vertex by using directly linked vertices and vertex information beyond a one depth.

3.2 Context Pre-processing & NER

The second step is a context pre-processing that extracts the whole word set from the context. This step consists of two parts.

The first part is context pre-processing which is the tokenization, stopwords remove, and stemming. Among the Stanford NLP tools, we use tokenization tools to break up words, and use the stopwords removal tool to remove the stopwords such as I, am, are, etc. After eliminating the stopwords, the remained words will be converted into the original form through the stemming process.

The second part is Named Entity Recognition (NER) which extracts the named entity using NER technique. Among the Stanford NLP tools, NER tools will be used to perform entity recognition.

3.3 Extract Candidate Entities

The third step is extracting candidates of named entity using the whole word set. The third stage consists of two parts.

The first part is the Word Sense Disambiguation which connects the proper meaning of the word when there is a common noun with ambiguity meaning in the whole word set extracted in the first step. In this paper, we apply Lesk algorithm [3],

which is well known as a knowledge-based approach. When applying to the Lesk algorithm, we will compute the similarity between the context and descriptions of the word. And then the highest similarity is selected to proper meaning.

The second part is that extracts candidates of the named entity when there is a named entity with ambiguity meaning in the whole word set. When extracting candidates, named entities in DBpedia are extracted through keyword matching.

That is, the whole word set extracted through steps 2 and 3 is extended to candidate word sets of a combination of a disambiguated words and candidate entities.

3.4 Apply Candidate Word Sets to Pre-weighted Graph

The fourth step is creating a graph of each candidate word set by applying a weighted-graph constructed in step 1. When the candidate word set is applied to the weighted-graph, it gradually expands from the vertex of the first word to the other vertices connected to the vertex. If we find the second vertex of the candidate word set while expanding, we extend the graph again based on the second vertex. In this way, when the last vertex appears while the graph is gradually expanded, the process is stopped.

3.5 Find the Shortest Distance Graph in the Candidate Graph

In step 5, we will compute the shortest distance to the vertex of each candidate word set based on the graph created for each candidate word set. When moving from one vertex to another vertex, select the vertex with the highest edge weight between the two vertices. In this way, we go to the path until the last vertex is reached. For each candidate graph, we choose the shortest path graph among the candidate graphs. The candidate named entity of the selected graph will be considered the correct answer for the named entity used in the context.

4 Discussion & Conclusion

In this paper, we performed for solving EL using the graph-based knowledge base. We try to solve the problem of EL with three approaches. Our first approach used every words appeared in a context. This is because words appeared in a context are related to each other. That is, it can be said that information of simultaneous appearance of words is utilized. The second approach is that words are more closely related and the distance between words is closer. That is, because the order of the words is important, we will represent the set of words in a sequence. The third approach seeks that strongly related words will be closer to the graph-based knowledge base.

In the future, we will simplify the process of extending the graph for the candidate word set, and apply various algorithms such as dynamic programming to apply the shortest distance.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No. 2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform).

References

1. Andrea, M., Alessandro, R., Robert, N.: Entity Linking meets Word Sense Disambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics. 2 (2014)
2. Xianpei, H., Le, S., Jun, Z.: Collective Entity Linking in Web Text: A Graph-Based Method. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 765-774. SIGIR '11. Beijing, China (2011).
3. Michael, L.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th annual international conference on Systems documentation. pp. 24-26. SIGDOC '86. Toronto, Ontario, Canada (1986).