Improved Entity Linking with User History and News Articles

Soyun Jeong, Youngmin Park, Sangwoo Kang, Jungyun Seo
Department of Computer Science and Engineering,
Sogang University,
Seoul, Korea
{soyun.j.nlp,pymnlp,gahng.sw}@gmail.com
seojy@sogang.ac.kr

Abstract

Recent researches on EL(Entity Linking) have attempted to disambiguate entities by using a knowledge base to handle the semantic relatedness and up-to-date information. However, EL for tweets using a knowledge base, leads to poor disambiguation performance, because the data tend to address short and noisy contexts and current issues that are updated in real time. In this paper, we propose an approach to building an EL system that links ambiguous entities to the corresponding entries in a given knowledge base through the news articles and the user history. Using news articles, the system can overcome the problem of Wikipedia coverage, which does not handle issues in real time. In addition, because we assume that users post tweets related to their particular interests, our system can also be effectively applied to short tweet data through the user history. The experimental results show that our system achieves a precision of 67.7% and outperforms the EL methods that only use knowledge base for tweets.

1 Introduction

Recent development of the internet and computing technologies makes the amount of information increasing rapidly. Therefore, many long-term studies have been conducted on retrieving the needed information from the huge data. Named entity recognition(NER) and entity linking(EL) to specific entities as a part of information extraction now are actively attempt to extract meaningful knowlege in the huge information. The EL is the task of linking entity mentions int text to entities in a knowlege base.

As shown in Figure 1, the goal of entity linking is to map an ambiguous entity to its corresponding

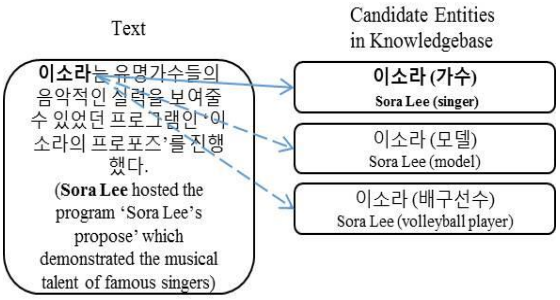

Figure 1: An example of entity linking. The bold type indicates an ambiguous named entity is in the text; the correct mapping entity is linked with the solid arrow

entity in knowledgebase. By leveraging the context information around an entity and knowledge base, '이소라 (Sora Lee)' in the left box 'Text' in Figure 1 can be identified as the singer '이소라 (가수) (Sora Lee (singer))'. Context information can be a noun phrase; for example, '이소라의 프로포즈 (Sora Lee's propose)', which is the name of a music program hosted by '이소라 (가수) (Sora Lee (singer))', can be known by knowledgebase.

Researchers have recently begun studying the problem of addressing named entities in informal and short texts. For example, Twitter, a popular microblogging platform, is updated and posted by users succinctly describing their current status within a limit of 140 characters. Java et al. (2007) showed that tweets address contents ranging from daily life to current events, news stories, and other interests. EL on Twitter has been used to identify entities from a structural knowledge base, e.g., Wikipedia, to enrich the task with additional features. To consider the characteristic of Twitter, state-of-the-art researches collectively link all entities in all tweets posted by a user via modeling the user's interest (Shen et al., 2013; Bansal et al.,

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 19 - 26
Shanghai, China, October 30 - November 1, 2015

2014). However, such methods cannot cover a EL for tweet task completely, because the posting the latest issues on tweet mentions, the most important characteristic of Twitter, cannot be applied.

In this paper, we first propose an EL method that considers Twitter contents addressing current issues and user interests through news articles and user history tweets besides knowledge base. In section 2, we describe recent EL studies. Section 3 provides our improved EL model with user history and news articles. Next, in section 4 we describe an experimental analysis in which we generated a Korean Twitter corpus and compared the contributions of each feature of our proposed method. Finally, we summarize our study with some concluding remarks in section 5.

2 Related Works

Traditional approaches have addressed the EL by dividing the task into two steps. The first step is NER, and the second step is entity disambiguation. Knowledge-based NER problem is different from the traditional NER. In the Traditional NER, while defining a class of such "PER" or "ORG" to entity, knowledge-based NER is to extract candidates of the fully qualified names of entities in knowledge base. For example, when recognizing the entity "이소라 (Sora Lee)", a common NER classifies it into classes such as "PER", while knowlege-based NER links it to specific entity such as "이소라 (모델) (Sora Lee (model))". Early models of EL had tried a method of extracting only those corresponding to the named entity existing in knowledge base of all possible n-gram terms within document(Mihalcea and Csomai, 2007). Milne and Witten (2008) tried to utilize machine learning methods to recognize entities. Kim et al. (2014) uses hyperlinks within the Korean Wikipedia and a small amount of text manually annotated with entity information as training data. It employs a SVM model trained with chracter-based features to recognize entity. Liu et al. (2011) proposed an alternating two-step approach that alternates between the KNN classifier and CRF labeler in tweets. The KNN classifier models global features which span over long range of words. The CRF models the localized features among consecutive words. After recognizing entities in document, the next step is entity disambiguation. In section 2.1, we describe previous works about entity disambiguation based on Wikipedia as knowledge base. Next, in section 2.2 describes state-of-the-art researches on EL via user modeling on tweets.

2.1 Entity Linking based on Wikipedia

Approaches leveraging Wikipedia for entity disambiguation started with Bunescu and Pasca (2006) and have been proposed in Cucerzan (2007), Han and Zhao (2009), Milne and Witten (2008) , Charton et al. (2014). Bunescu and Pasca (2006) defined a sematic relatedness by similarity measure using Wikipedia categories. Later studies developed methods using richer structural features from the Wikipedia. The semantic relatedness is measured through the co-occurrence of links in Wikipedia articles. Milne and Witten (2008) have proposed to compute the mention to entity compatibility by leveraging the interdependence between EL. The system proposed that referent entity of a name mention should be coherent with its unambiguous contextual entities. Han and Zhao (2009) demonstrated how to leverage the semantic knowledge in Wikipedia, so the performance of named entity disambiguation can be enhanced by obtaining a more accurate similarity measure between name observations. Charton et al. (2014) built a representation of named entities that do not appear same as the knowledge base named entities.

2.2 Entity Linking via User Modeling

For EL on tweets aimed at short and noisy texts, the system should cover the insufficient context information contained in a tweet. To overcome such a problem, Shen et al. (2013) and Bansal et al. (2014) proposed an EL system via user modeling. Shen et al. (2013) suggested the KAURI system, a graph-based framework to collectively link all named entity mentions in all tweets posted by a user via modeling the user's topics of interest. They assumed that each user has an underlying topic interest distribution over various named entities. Bansal et al. (2014) attempted to combine contextual and user models by analyzing a user's tweeting behavior from previous tweets. This approach can be used for modeling users and disambiguating entities in other streaming documents. EL applied through user modeling systems outperforms systems using only a knowledge base.

In this paper, we adopted the traditional method that extracting only those corresponding to the named entity existing in knowledge base of all possible n-gram terms within document in the NER step. By proposing three models considering characteristics of the tweets, we focus on entity disambiguation step.

3 Entity Linking System with User History and News Articles

3.1 Notation Framework

Our system is applied based on the user's interest and current issues, and by considering which contents their Twitter mentions address. In this section, we introduce our proposed system, which consists of three scoring model systems, a Context modeling system, a User modeling system, and an Issue modeling system, as well the Linking model as shown in Figure 2. We adopted an existing method into the Context modeling system, by considering the context information around an ambiguous entity. The User modeling system uses the history mentions of user who posted the targeted mention. Targeted mention means the tweet mention with ambiguous entity, which we have to disambiguate. The Issue modeling system enriches the information from news articles, and can extract information that Wikipedia does not handle. The following section focuses on how our User modeling system works well in comparison to a Context modeling system. Furthermore we describe how the Issue modeling system improves the entire system to obtain a high level of performance.

- E – Target entity that should be linked
- e^j – j-th non-ambiguous entity, which corresponds to an Wikipedia entity
- c_j – j-th candidate entity for entity mention E that can be linked
- ⟨D⟩ – Sets of all entities in a document D
- $D_{c_j}^i$ – i-th news article that includes c_j as a topic, D_{c_j} means the set of all $D_{c_j}^i$
- [e] – Sets of all links in Wikipedia article whose title corresponds to entity e
- $S_c(c_j)$, $S_u(c_j)$, $S_i(c_j)$ – j-th candidate entity score of the Context modeling system, User modeling system, and Issue modeling system respectively
- Wikipedia article(page) title – Synonym of Wikipedia entity. Each article(page) in Wikipedia describes a specific entity and its title can be used to represent the entity it describes. Each article includes links which have semantic relation to its title. In other words, Wikipedia entities are considered to be semantic related if there are links between them.

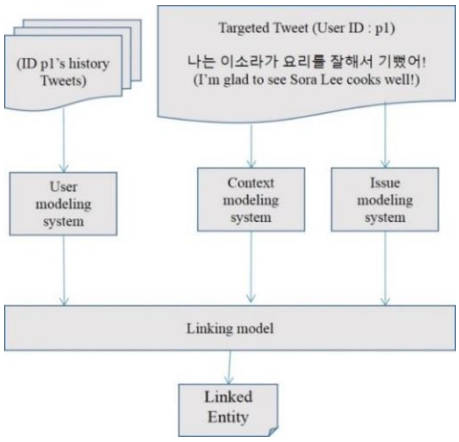

Figure 2**:** EL system

3.2 Context Modeling System

The Context modeling system uses contextual information, meaning the non-ambiguous entities in a tweet including an ambiguous named entity. Most researches use this feature with Wikipedia category information, but we found that categories are often noisy(Milne and Witten, 2008) and that Korean Wikipedia pages provides insufficient category information compared with English Wikipedia pages. We therefore we adopted Eric Charton's scoring method "mutual relation score" (Charton et al., 2014) without category information, as defined by the following formula :

$$S_c(c_j) = \partial\, dsr_{score}(e^j, c_j) + (1-\partial) csr_{score}(e^j, c_j) \quad (1)$$

$$dsr_{score}(e^j, c_j) = \left|e^j \cap [c_j]\right| \quad (2)$$

$$csr_{score}(e^j, c_j) = \frac{|[e^j] \cap [c_j]|}{|[e^j]| + |[c_j]|} \quad (3)$$

Figure 3: Korean Wikipedia disambiguation page of named entity "이소라 (Sora Lee)" which appeared in user p1's tweet mention shown on Figure 1

A Context modeling system fundamentally addresses the contextual features and links with a specific entity in Wikipedia. In Wikipedia, there is a "disambiguation page" that describes entities with the same name. As shown in Figure 3, the disambiguation page for 이소라 (Sora Lee) lists three other people with the same name. The first 이소라 (Sora Lee) listed is a famous Korean model. The second 이소라 (Sora Lee) is a famous Korean singer. The last 이소라 (Sora Lee) is a member of the Korean national volleyball team. In this paper, we use the term S_c to indicate the "calculate score" in (Charton et al., 2014), which we use as our baseline system, as compared to systems with other added features. S_c implies simply exploiting a tweet as a context, not considering the properties of the tweet.

3.3 User Modeling System

As attributes of Twitter mentions, tweet contents range from daily life to current issues. Because the User modeling system understands the above property, it handles the user's behaviors and interests. To address this concept, we utilize the past tweets of the user. We assumed that if a particular named entity is mentioned in a tweet, the user tends to have an interest in this named entity.

$$S_u(c_j) = \sum_{e^j \epsilon \langle \mathrm{D} \rangle} \partial \, dsr_{score}(e^j, c_j) + (1-\partial) csr_{score}(e^j, c_j) \quad (4)$$

$$\langle \mathrm{D} \rangle = \{e^j | j \; ranges \; from \; user's \; past \; tweets \; to \; the \; present \; tweet\} \quad (5)$$

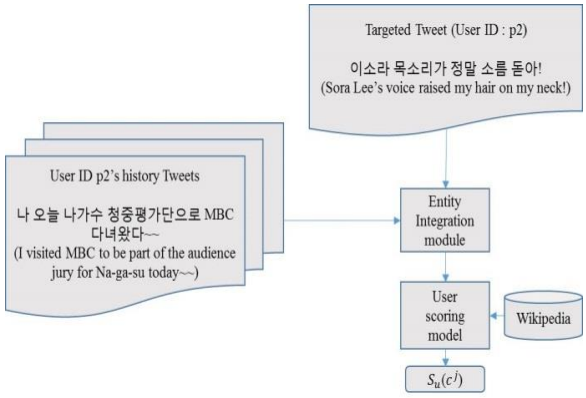

Figure 4 : User Modeling System

Figure 4 describes the process of the User modeling system. When the User modeling system detects a certain tweet of a user that includes an ambiguous entity, then we extract the user's tweet history. The Entity Integration module then detects whether the feature e^j exists in Wikipedia entity by using the left-longest-match-preference method with *eojeol* uni-gram and bi-gram, then generate ⟨D⟩. j ranges from user's past tweets to their last present tweet which to be disambiguated as shown in (5). We exploit the left-longest-match-preference method with *eojeol* uni-gram and bi-gram because tweet mentions tend to be grammatically incorrect and because in Korean, a noun always appears on the left-side in an *eojeol*, which consists of one or more morphemes comprising a spacing unit (Kang et al., 2014). Finally, S_u in the User scoring model is evaluated as in (4) and (5). In the example shown in Figure 4, the system has detected an ambiguous entity '이소라 (Sora Lee)' in user p2's tweet and extracted p2's tweet history. The Entity Integration model then collect entities such as "나가수 (Na-ga-su)", the TV program, and "MBC", which is the broadcast station from p2's tweet history. These features enhance the S_u(이소라 (가수)), because [이소라 (가수) (Sora Lee (singer))] includes "나가수 (Na-ga-su)" and "MBC".

3.4 Issue Modeling System

The model described in section 3.3 has a disadvantage in that it cannot consider current issues, which is one of the characteristics of Twitter. A knowledge base focuses on major issues

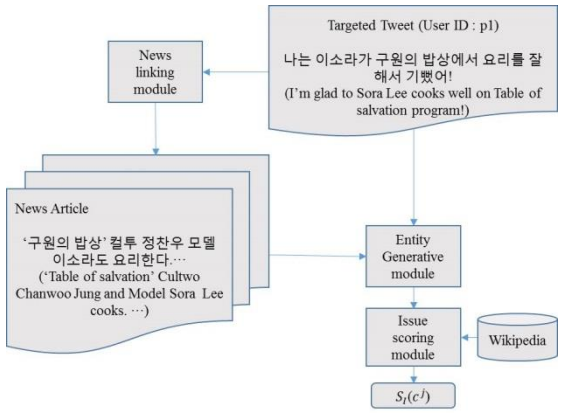

Figure 5: Issue Modeling System

and cannot provide all current issues in real time, for example, the recent events of a celebrity or trivial real-time events, which tend to be mentioned by Twitter users. Our Issue modeling system solves this problem by leveraging news articles. Because news articles address issues and events in real time, the Issue modeling system can extract current information.

As shown in Figure 5, when the Issue modeling system detects the ambiguous entity, "이소라 (Sora Lee)", the news linking module extracts current news articles issued in within k days from the date of the tweet posted. The news articles have title included the detected entity E, 이소라 (Sora Lee). The module links the news articles to c_j by applying a cosine similarity between the article and Wikipedia page contents. In this example, c_1, c_2, and c_3 are "이소라 (가수) (Sora Lee (singer))", "이소라 (모델) (Sora Lee (model))" and "이소라 (배구선수) (Sora Lee (volleyball player))" respectively. Accordingly, the Issue modeling system exploits the news articles as Wikipedia articles. Each article can be a $D_{c_j}^i$, which means the i-th news article which addresses c_j as a topic. The Entity Generative model generates the Wikipedia links in each news article. Single quotes in newspaper articles has the ability to display title or name. For example, the title of the book, the title of the movie, the title of the album and the title of the drama can be placed in single quotes[13]. Therefore we assume that news articles include important entities explicitly notated by punctuation. The Entity Generative model recognizes phrases or words in single punctuation marks and nouns as entities, similar to Wikipedia links.

Table 1 describes the generation rules of the Entity Generative model and the example news article from Figure 5. Because the entity "구원의 밥상 (table of salvation)" is not actually included in the Wikipedia page on "이소라 (모델) (Sora Lee (model))", it can be important information to link E , "이소라" to the correct answer, c_2 , "이소라 (모델) (Sora Lee (model))". The second generation rule, extracting entities that exist in Wikipedia article titles, the Entity generative model exploits a morpheme analyzer to extract nouns from news articles. This model uses a noun uni-gram and bi-gram to match the entity in Wikipedia article titles. Finally Issue scoring module determines the score S_i, as shown in (6).

$$S_i(c_j) = \frac{\sum_{i=0}^{|\boldsymbol{D}_{c^j}|} \partial\, dsr_{score}\left(e^j, D_{c_j}^i\right) + (1-\partial) csr_{score}\left(e^j, D_{c_j}^i\right)}{|\boldsymbol{D}_{c^j}|} \quad (6)$$

Generation Rule	Extracted Example
phrases or words in single punctuation marks and nouns	"구원의 밥상" ("table of salvation")
entities that exist in Wikipedia article titles	"컬투", "정찬우", "모델" ("Cultwo", "Chanwoo Jung", "model")

Table 1: Generation rule of the Entity Generative Model in Issue Modeling System

3.5 Linking Model

The Linking model finally disambiguates the entities based on the three models above, computes the Total Relatedness score, TR, and combines the scores $S_c(c_j)$, $S_u(c_j)$ and $S_i\ (c_j)$ with parameter α, β, and γ, which are defined empirically. Equation (7) shows how this works.

$$TR(E, c_j) = \alpha\, S_c(c_j) + \beta\, S_u(c_j) + \gamma\, S_{\,i}(c_j) \quad (7)$$
$$(\alpha + \beta + \gamma = 1)$$

4 Experiments

4.1 Dataset and Framework

In the experiments, we evaluate our proposed method on the disambiguation of personal names, which is the most common type of named entity disambiguation. We created data set by collecting 50~60 tweets per 300 users who use twitter actively. Finally we collected 16,367 tweets in total. Then we selected tweets that contain the one person entity which exists in the list of entities in the Wikipedia's disambiguation page. Finally we selected 248 tweets annotated same entity by 3 different annotators to verify reliability. The data set consists of 248 tweets including 248 disambiguous entities and they represent 33 ambiguous PERSON named entities. 33 ambiguous entities have 4.75 disambiguation pages on average in Wikipedia, 3.45 in data set. We conducted our experiments using k = 3. We defined α,β, and γ empirically because they depend on the dataset. We used Korean Wikipedia as a knowledge base, the contents of which can be downloaded from http://download.wikipedia.org/kowiki. In our experiment, we dumped the latest Korean Wikipedia dump file, kowiki-2015-6-2-pages-articles. In the Issue modeling system, we adopted a Korean morpheme analyzer, "Jhannanum" [1]. In addition, we constructed a Wikipedia PER entity dictionary using the category information.

4.2 Experimental Result

Table 2 shows the performance of our proposed system. In the first row, the system is evaluated using only the Context modeling system. For the second row, we applied the User modeling system along with the Context modeling system. The system with the complete entity linking system obtained the results provided in the third row. We applied the accuracy score(number of true positives + number of true negatives/number of data set) to evaluate the system. We observe that the results of complete entity linking system are shown in the third row. We applied a precision score to evaluate the system. We observed that the complete algorithm provides the best results for our created test set. Considering that the 33 PER entities in our test set have 4.75 disambiguation pages on average, our proposed system performed well within a 67.7% level of accuracy. We also measured how precisely the Issue modeling system linked between news articles and Wikipedia pages. A 70.2% level of precision was achieved using only the cosine similarity. We showed that the complete system improves the performance of the Context modeling system when using user history and news articles.

System	Accuracy
Baseline (Context Modeling System)	31.5
Baseline +User Modeling System	58.9
Baseline +User Modeling System +Issue Modeling System	**67.7**

Table 2: Experimental Results

Table 3 shows the performance of News Linking module in Issue Modeling system. Ambiguous entity in collected news articles's title were annotated by 2 different annotator. The Linking module performed within a 70.2% level of accuracy.

#news article	#entity type	# same name in news article	# same name in Wikipedia	Accuracy
836	20	1.4	3.35	0.72

Table 3: Performance of News Linking module in Issue Modeling system

Table 4 shows the extracted entities from each systems. First example shows improved performance by extracted entities from User Modeling System. This tweet user likes baseball as usual because he mentioned "삼성(Samsung)", "롯데(Lotte)", "야구장(ballpark)", "조성환(sunghwan Cho)" in previous tweets. Among them, "삼성(Samsung)" and "롯데(Lotte)" appear in

1. Semantic Web Research Center , JHannanum, http://semanticweb.kaist.ac.kr/home/index.php/HanNanum

Tweet	Extracted entities
@DooBoo_2 - 김태균이랑 동선이라닠ㅋㅋㅋㅋ (“Same line with Taegyun Kim kkk”)	Context modeling system
	″김태균(**Taegyun Kim**)″
	User modeling system
	″삼성(**Samsung**)″, ″롯데(**Lotte**)″, ″야구장(**ballpark**)″, ″조성환(sungh wan Cho)″ ...
	Issue modeling system
	″**SK wyverns**″, ″ **Samsung Lions**″, ″우승**(victory)**″, ″한국프로야구(**K orean Baseball**)″...
@myhomenamsan - 어제 조인성을 봤다. 화장실에서 거울 봤는데, 우리가 엄마가 잘못했다 (“I saw Insung Cho yesterday. After I saw my face in the mirror. My mother’s fault.”)	Context modeling system
	″조인성(**Insung Cho**)″
	User modeling system
	″축구(**soccer**)″, ″일본(**Japan**)″, ″중국(**China**)″...
	Issue modeling system
	″드라마(**drama**)″, ″영화(**movie**)″,″ **SBS**″, ″배우(**actor**)″, ″태국(**Thailan d**)″...

Table 4: Extracted entity examples from three modeling systems

“김태균(1971) (Taegyun Kim (1971))”. Further, entities extracted by Issue Modeling System enhance the system to resolve into “김태균(1971) (Taegyun Kim (1971))”. User Modeling system extracted entities in second example does not support the “조인성(배우) (Insung Cho(Actor))”. Instead, Issue Modeling system extracted “드라마(drama)”, “영화(movie)”, “SBS”, “배우(actor)” that support the system can link to “조인성(배우) (Insung Cho(Actor))”.

5 Conclusion

In this paper, we propose an entity linking system that, consists of three scoring model systems, a Context modeling system, a User modeling system and an Issue modeling system, along with a Linking model to integrate the three systems. We adopted an existing entity linking method as a baseline for Korean tweets, and by applying the User modeling system and Issue modeling system, it outperforms the baseline system, just using knowledge base. Our system handles the characteristics of tweet mentions, such as current issues or trivial events that not described in a knowledge base, by using the User modeling system and Issue modeling system effectively.

However, because our work is the first to link entities with three different scoring model systems, the User modeling system does not use the additional features such as Twitter hashtag information. Furthermore, because only the left-longest-match-preference model and a noun unigram and bigram were used to detect entities in news articles in this experiment, the accuracy was not very high and should be improved later.

Further analysis is required for the user modeling and Issue modeling aspects of the system. Future work will also involve applying statistical methods to identify entities in news articles and using additional features appearing in Twitter for the User modeling system. Furthermore, we will adopt an efficient method to link news articles with tweets. We also plan to experiment on larger datasets and adopt our system to English tweets such as TAC_KBP.

Acknowledgment

This work was supported by the ICT R&D program of MSIP/IITP. [R0126-15-1112, Development of Media Application Framework based on Multi-modality which enables Personal Media Reconstruction]

References

Akshay Java, Xiadan Song, Tim Finin and Belle Tseng. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. *In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 56-65.

Cucerzan Silviu. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 7: 708–716.

David Milne and Ian H. Witten. 2008. *Learning to Link with Wikipedia*. *In Proceedings of the 18th conference on Information and knowledge management*, 215-224.

Donghyuk Lee. 2008. The Function of Single Quotation Marks on the Newspaper Articles. *Journal of Urimal*, (23):139-162.

Eric Charton, Marie-Jean Meurs, Ludovic Jean-Louis and Michel Gagnon. 2014. Mutual Disambiguation for Entity Linking. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 476–481.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. *In Proceedings of the 16th conference on Conference on information and knowledge management* , 233-242

Razvan Bunescu and Marius Pasca. 2006. Using Encyclopedic Knowledge for Named entity Disambiguation. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 6: 9-16.

Romil Bansal, Sandeep Panem., Manish Gupta and Vasudeva Varma. 2014. EDIUM: Improving Entity Disambiguation via User Modeling. *Journal of Advances in Information Retrieval*, 8416:418-423.

Sangwoo Kang, Harksoo Kim, Hyun-Kyu Kang and Jungyun Seo. 2014. Lightweight morphological analysis model for smart home applications based on natural language interfaces. *International Journal of Distributed Sensor Networks*, 2014:1-9

Wei Shen, Jianyong Wang, Ping Luo and Min Wang. 2013. Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling. *In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 68-76.

Xianpei Han and Jun Zhao. 2009. Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge. *In Proceedings of the 18th conference on Information and knowledge management*, 215-224.

Xiaohua Liu, Ming Zhou, Xiangyang Zhou, Zhongyang Fu and Furu Wei. 2011. Recognizing Named Entities in Tweets. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1:359–369.

Youngsik Kim, Youngkun Hamn, Jisung Kim, Dosam Hwang and Ki-Sun Choi. 2014. A Non-morphological Approach for DBpedia URI Spotting within Korean Text, *In Proceedings of the 26th HCLT*, 100-106.