DISAMBIGUATION OF HUMAN NAMES IN TEXT

ALF.Shanaz 1[1] and R.Ragel 2[2]

[1] Department of Computer Science & Engineering, Faculty of Engineering, South Eastern University of Sri Lanka, Sri Lanka
[1,2] Department of Computer Engineering, Faculty of Engineering, University of Peradeniya, Sri Lanka

shanazlathif@gmail.com, ragelrg@gmail.com

ABSTRACT: *The aim of this paper is to implement an entity linking system for news recommendation. Which can automatically recognize Person entities (humans) from input English text (news article), and link them to the best-matched entities in Wikidata knowledge base. That is, for each specific mention of a person entity found in a text, the developed Named Entity Disambiguation (NED) algorithm was applied to search for candidate entities (in Wikidata) and return either the best candidate or a NIL reference if the spotted person entity does not match any Human in Wikidata. In a nutshell, our system maps mentions of ambiguous human names (people mention) in text onto Wikidata unique identifier (Q number). We extensively evaluated the performance of our system over manually annotated AIDA CoNLL-YAGO Dataset, and the experimental results show that our system achieves the top-5 precision of 84.4%.*
Keywords: Named Entity Linking, Named Entity Disambiguation

1. INTRODUCTION

Online news reading has become general among people as the web provides access to news articles from various sources around the world (Liu, Dolan, and Pedersen 2010). Finding relevant news articles to readers is a non-trivial task that requires access to information about the user, the news item and the general context. Hence, the personalized news recommendation has attracted more research attention. News aggregator sites such as Google news provides personalization service to its a substantial amount of online users by aggregating news articles from various news sources worldwide. The news recommender systems (NRS) are built to provide the most relevant news articles to online news readers based on their interest and needs without any manual search effort by the users. Besides, news recommender systems are different from other recommender systems like books, music and movies.

One of the most critical challenges the NRS has to deal with is user profiling (Knowledge of user preferences): In order to make more individual specific recommendations, it is needed to construct a user profile (Özgöbek, Gulla, and Erdur 2014). News reading preferences of a user may change over time (Abel et al. 2011; Lu et al. 2016; Özgöbek, Gulla, and Erdur 2014). Hence, NRS should be able to modify itself as per the needs of the user by employing different recommendation methods. Which helps to recognize user requirements and behave in accordance with them (Kanoje, Girase, and Mukhopadhyay 2014; Özgöbek, Gulla, and Erdur 2014; Schiaffino and Amandi 2009).

User profiling for News recommendation

A typical user profile can include information about user's interests and preferences, and for each user of a website, it can also contain various user characteristics, such as age, gender, ethnicity, location, etc. (Adomavicius and Tuzhilin 2005). Modelling users is usually an application dependent approach (Plumbaum 2015). In an online newspaper domain, the user profile would contain information such as the types of news (topics) the user likes to read, the kinds of news (topics) the user does not wish to read, the newspapers he usually reads, and the user's reading habits and patterns (Schiaffino and Amandi 2009).

There are many approaches to build user profiles for news recommendation. However, accurate user profiling is essential for an online recommender system to provide proper personalized recommendations to its users (Lu et al. 2016).
The growth of Web 2.0 tools allows users to interact and collaborate as content creators. Researchers make use of information obtained from Twitter to build user profiles for news recommendation (Abel et al. 2013; Lee et al. 2014). Twitter also acts as an entry point to news for many readers (Dimitrova et al. 2017). When Twitter users' find an interesting news article they tend to share it with their followers. If the followers consider it as an interesting article then, they like and/or comment and/or retweet it. This kind of tweets usually provides a web link to the news article page. Hence, the news article links given in tweets can be utilized to extract the user's favorite entities such as person, place, product, etc. This paper will focus on extracting user's favourite personalities from the news articles they shared on Twitter. Because person entity can be used infer user's news topic interest such as sports, entertainment, politics, health and much more. Also, in news aggregator domain, extracted person entities from user favourite news articles can also be used to infer the geographic location interest of the user and much more.

News articles are mostly available in the form of natural language text, and they contain mentions of different named entities such as people, places, organizations, etc. These mentions are often ambiguous: the same mention can refer to different entities, and an entity can also be referred to by multiple mentions. For instance a surface form "Bush" can refer to multiple entities; for example two former Presidents of U.S., American football player "Reggie Bush" and British singer "Kate Bush" and an entity George W. Bush, 43rd President of the United States, might refer by multiple surface forms (such as "George Bush", "George Walker Bush", "Bush, George W."). Therefore, a systematic approach is required to map mentions in the text to actual entities. This problem is called Named Entity Disambiguation (NED). Building an entity linking system requires mention detection and entity disambiguation. The following sections describe the two main steps in named entity disambiguation.

1.1. NER

Named entity recognition is an essential task of information extraction systems. Its goal is the identification of mentions of entities in text such as people, locations, organisations and

products, and label them with one of several entity type labels. Various NER tools have been developed in the last decade. Some publicly available and well-established NER tools include Stanford NER, spaCy and NLTK.

1.2. Named Entity Disambiguation (NED)

Named entity disambiguation (NED) is the task to link entity mentions in the text to the actual entities in a knowledge base (Shen, Wang, and Han 2015). This task is also referred to as Named Entity Linking (NEL). The focus around entity linking system has increased significantly in the last few years, with effective algorithmic approaches to solve the mention-entity match problem, possibly using other knowledge bases such as DBpedia, Freebase or Yago (Cornolti, Ferragina, and Ciaramita 2013).

A knowledge base is a fundamental component for the entity linking task. Knowledge bases provide information about the world's entities, their semantic categories and the mutual relationships between entities (Shen, Wang, and Han 2015). The knowledge base will be described in detail in Section 1.3.

Some of the publicly available entity linking systems include AIDA, Illinois Wikifier, TagMe, and DBpedia Spotlight; which currently define the state-of-the-art for the entity-annotation task (Cornolti, Ferragina, and Ciaramita 2013).

According to (Cornolti, Ferragina, and Ciaramita 2013; Shen, Wang, and Han 2015), an entity disambiguation system consists of the following three modules:

1. **Candidate Entity Generation**: Entity linking systems try to include possible entities from the knowledge base that matches entity mention in the text. Approaches to candidate entity generation are largely based on string comparison between the surface form of the mention in text and the name of the entity existing in a knowledge base. Some main approaches that have been applied for generating the candidate entity set are listed below:
    a. Name Dictionary Based Techniques: Most of entity linking systems have leveraged this technique. An offline name dictionary is built by leveraging features available in the knowledge base, which maps various names to their possible entities. (Wikidata alias feature makes this step easier for us).
    b. Surface form expansion approach: As we described in our earlier example, some entity mentions are part of full names. Therefore, this technique is used to expand the surface form of an entity mention into a richer form from the local document where the entity mention appears.
    c. Search engine based approach: leverage the whole Web information to identify candidate entities via Web search engines. Researchers have leveraged search engines such as Google and Wikipedia. Search engines were queried using entity mentions, and the top N number of Wikipedia pages in the search result were taken as candidate entities.

2. **Candidate Entity Ranking**: Disambiguation of mentions, which is the task of selecting the most appropriate candidate entity that best describes each mention; The Candidate Entity Ranking module is a key component for the entity linking system. There are two types of features found to be useful for candidate entity ranking: context-independent features and context-dependent features. Context-independent features rely on the surface form of the entity mention and the knowledge about the candidate entity. Context-dependent features are based on the context where the entity mention appears.
3. **Unlinkable Mention Prediction**: some entity mention does not have its corresponding record in a knowledge base. Therefore, they have to deal with the problem of predicting unlinkable mentions. (Shen, Wang, and Han 2015) states that more research on the area of entity linking problem is required for the emergence of more effective and efficient entity linking systems.

1.3. Knowledgebase

The success of Wikipedia and the proposed vision of Linked open data (Bizer, Heath, and Berners-Lee 2009) have enabled the construction of large-scale machine-understanding knowledge base about the world's entities, their semantic categories and the relationships between them. Four of such knowledge bases which have been widely exploited in the field of entity linking include Wikipedia[1], DBpedia (Auer et al. 2007), YAGO (Suchanek et al. 2017) and Freebase (Bollacker et al. 2008). Knowledge bases are a prerequisite for entity disambiguation. However, exploitation of knowledgebases to create efficient and effective entity linking system is yet to be explored.

Knowledge bases provide information about the world's entities, their semantic categories and the mutual relationships between entities. Very few works has been done on entity linking system that makes use of wikidata as a knowledge base. Which is a free knowledge base contains over 50,329,084 data items. The following section will describe wikidata in detail.

1.4. WIKIDATA

Wikidata[2] is the community-created knowledge base operated by the Wikimedia Foundation. It was launched on October 30, 2012. It is intended to provide a common source of data to be used by Wikimedia projects such as Wikipedia, Wikinews, Wikisource, and by anyone under a public domain licence. Data is collect in a structured form, this allows easy reuse of that data by Wikimedia projects and third parties and will enable computers to easily process and "understand" it. The free knowledge base contained 50,329,084 data items on 27th September 2018.

[1] https://www.wikipedia.org/
[2] https://www.wikidata.org

The data is also completely free and open. It is the richness of the data that makes Wikidata unique. Data is strongly connected to external datasets in many domains, and all of the data is multi-lingual by design.

There are many ways to access structured content from Wikidata.

1. Wikidata offers copies of the available content for anyone to download. A complete database dump of all entities in the Wikidata in a single JSON array can be downloaded.[3] JSON is the recommended dump format. Where each entity is placed on a separate line in the JSON file. These dumps are released once a week.
2. We can also access data per item via dereferenceable URIs, or we can query the data in Wikidata through the Wikidata Query Service. It can be used both as an interactive web interface and by application programming interface.

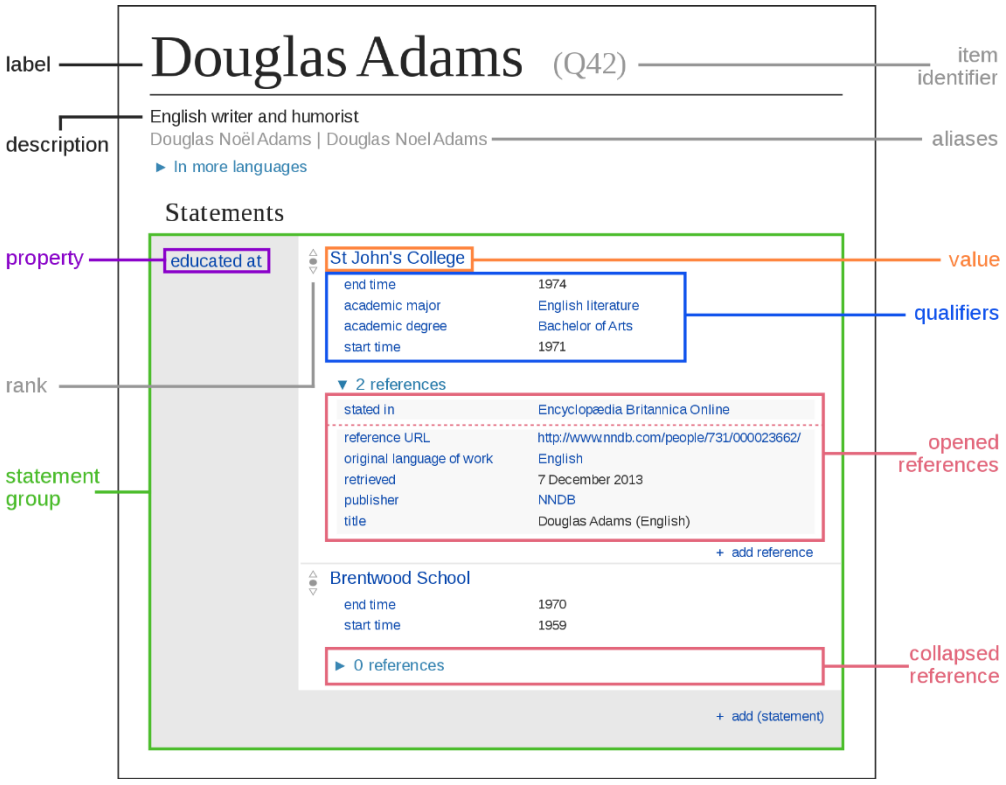

Figure 8 Datamodel in Wikidata[4] - shows the most important terms used in Wikidata

Followings are some of important Wikidata concepts defined in Wikidata glossary page[5].

[3] https://dumps.wikimedia.org/wikidatawiki/entities/
[4] https://commons.wikimedia.org/wiki/File:Datamodel_in_Wikidata.svg
[5] https://www.wikidata.org/wiki/Wikidata:Glossary

Entity is the content of a Wikidata page that can be either an *item* or a *property*. Every entity is identified by a unique entity ID, which is a prefixed number (items start with the prefix Q and the prefix P for property).

An entity is also identified by a unique combination of label and description. An entity may have alternate aliases in multiple languages.

Label is the main name given to identify an entity. Example, the item identifier Q42 has the English label “Douglas Adams”. Every entity has exactly one label in a given human language. Labels do not need to be unique. Multiple items can have the same label, however no two items have both the same label and the same description.

Description is a descriptive phrase for an item. Wikidata has set a constraint that, it must be unique together with the label.

Aliases are alternative names for items. Unlike labels there can be many aliases for an item. Multiple items can have the same alias, with different descriptions. Also, some items may not have any aliases.

A **statement** consists of a property-value pair, for example, "occupation: cricketer". The **property** in a statement describes the data value, and can be thought of as a category of data, for example "color" for the data value "blue". The **value** in the statement is the actual piece of data that describes the item.

As we can see in Figure 1, the Wikidata repository consists mainly of items, each one having a label, a description and any number of aliases. Items are uniquely identified by a Q followed by a number, such as Douglas Adams (Q42). Statements describe detailed characteristics of an Item and consist of a property and a value. Properties in Wikidata have a P followed by a number, such as with educated at (P69).

2. METHODOLOGY

Our approach to Person entity disambiguation can be grouped into the following steps: find named entity mentions in the given text, generate a set of candidates for each mention, and select the best candidate for each mention. Every step is performed separately using our proposed algorithms, where each step utilizes information from the input text and the output of previous steps. This work will only focus on NEL task and skip named entity recognition step. We propose two algorithms for the last two steps that are performed in the names entity disambiguation task.

Proposed Candidate Entity Generation algorithm

Algorithm 1 describes our candidate entity generation procedure. The goal of the candidate entity generation step is to create a limited set of best candidates for each mention. To accomplish the task, we use wikidata’s item label and aliases. The candidate set for each mention in a document

would contain the best-matched candidate entities in the knowledgebase in relation to the document. Two types of string matching have been utilized such as exact matching of lowercased strings and fuzzy string matching. Fuzzy string matching, also called approximate string matching, is the process of finding strings that approximately match a given pattern. The closeness of a match is often measured in terms of edit distance[6]. The proposed algorithm was evaluated on AIDA CoNLL- YAGO Dataset (Hoffart et al. 2011)

Algorithm 1: Candidate Entity Generation Algorithm

```
Input: list of mentions in a document
Output: dictionary, list of candidate entities (CEs) for each mention

Set CEs to empty dictionary
For each mention in a document
      Set list of Related entities (REs) to empty
      If a mention matches previous mention in document
            Set CEs of mention to CEs of previous mention

      If mention exactly matches any wikidata item label (case insentitive)
            Update RES with wikidata Q number

      If mention exactly matches any wikidata aliases
            Update REs with wikidata Q number

      Update CEs[mention] = REs

      If REs is empty or mention is a unigram
            Update CEs with wikidata Q numbers obtained from fuzzy string matching
            of mention and item label
```

Proposed Candidate Entity Ranking algorithm

Algorithm 2 describes the proposed Candidate entity ranking algorithm. As we explained in Section 1.2. This step is used to select the most appropriate candidate entity that best matches each mention. As mentioned earlier, this is a key component of any entity linking system. Here, we utilize the context-independent feature called entity popularity to rank the generated candidate entities in the previous step (Shen, Wang, and Han 2015) states that entity popularity feature is significantly important and effective for the entity linking task. Because each candidate entity of a mention has different popularity. For example, entity mention “Barack Obama”, the candidate entity Barack Obama (Kenyan economist) is less likely than the candidate entity Barack Obama (44th President of the United States of America).

Entity popularity was measured by counting the number of statements and site links of each entity in wikidata item page.

[6] https://marcobonzanini.com/2015/02/25/fuzzy-string-matching-in-python/

Algorithm 2: Candidate Entity Ranking Algorithm

```
Input: List of candidate entities for each mention
Output: Wikidata Q number (best matched entity for the entity mention)

For entity mention candidate entities
      Compute sum of count of statements and count of sitelinks
      Return entity with largest sum
```

Unlinkable Mention Prediction:

However, some entity mention does not have its corresponding record in a knowledge base. In this case, we link the mention to NIL identifier.

Datasets

Wikidata dump

Wikidata was utilized as the knowledge base for our entity linking task. We downloaded the Wikidata dump on 06-Aug-2018, which contains 49,717,457 entities. List of wikidata unique numbers that denote any human in wikidata was retrieved from Wikidata query service[7] using python sparql wrapper.

There were 4,523,207 humans in wikidata at the time of running the query. The result was the collection of URI's that refer to each human. According to Wikidata glossary page, each entity has a dereferenceable URI that follows the pattern http://www.wikidata.org/entity/ID where ID is its entity ID. For example, URI of the entity Douglas Adams is http://www.wikidata.org/entity/ Q42. The Q number (entity id) that appear in URI's were extracted, which were then used to get the entities' English attributes such as labels, aliases, descriptions, claims, sitelinks from the downloaded JSON dump file (as of 06-Aug-2018).

Table 1: Summary of Wikidata Human entities.

Total number of humans	*4,523,207*
Num of humans with item label	*3,803,206*
Num of humans with alias	*561,073*

AIDA CoNLL-YAGO Dataset

Only a few datasets are publicly available for the evaluation of entity linking algorithms, One of the most studied datasets is AIDA CoNLL- YAGO Dataset (Hoffart et al. 2011) and consisting of

[7] https://query.wikidata.org/

annotated English articles (Lewis et al. 2004). We use the AIDA CoNLL-YAGO English dataset to evaluate the performance of our entity linking system. This data set is the biggest data set which has been labeled for both NER and linking tasks (Luo et al. 2015). This was used in the experiments of (Hoffart et al. 2011) and freely available to download[8]. (Hoffart et al. 2011) created their own dataset based on CoNLL 2003 data.

The shared task of Conference on Natural Language Learning 2003 (CoNLL-2003)[9] concerns language-independent NER. The English data was taken from 1,393 Reuters (Reuters Corpus, Volume 1) news articles published between 20th August 1966 and 19th August 1997 (Lewis et al. 2004). To build the complete CoNLL-2003 English dataset, we obtained the access to Reuters Corpus for research purposes without any charge from NIST[10].

CoNLL 2003 dataset consists of proper noun annotations for all 1,393 Reuters news-wire articles. They focused on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups. (Hoffart et al. 2011) hand-annotated all these proper nouns with corresponding entities in YAGO2.

Table 2: Overview of CoNLL 2003 / AIDA dataset.

Articles	*1,393*
Number of mentions	*34,922*
Number of mentions with no entity	*7,112*

This table slightly differs from the table provided in the (Hoffart et al. 2011). Because, the dataset has been updated on 2013-11-21, adding all but 7 Freebase MIDs, as well as Wikipedia IDs[11].

The dataset also provides Wikipedia URL of each entity mention for the convenience of evaluating against a Wikipedia based method. To make it suitable to our Wikidata based approach, we found the corresponding unique Wikidata identifier for all Wikipedia URLs. Wikidata supports sitelinks for Wikipedia, the site id for English Wikipedia is enwiki. (To be able to evaluate on the AIDA CoNLL-YAGO Dataset we need to convert the entities in them to Wikidata QID identifiers). Since, our system has been designed to disambiguate only person entity mentions in text, for evaluation we only consider the mentions which maps to human entities in wikidata.

AIDA CoNLL- YAGO data is split into 3 parts: TRAIN, TESTA, TESTB. Importantly, the experimental results are given for TESTB.

Table 3: Person entity annotation in AIDA CoNLL- YAGO dataset.

	TRAIN	***TESTA***	***TESTB***	***Total***
Num Articles / Documents	*946*	*216*	*231*	*1393*

[8] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/

[9] https://www.clips.uantwerpen.be/conll2003/ner/

[10] https://trec.nist.gov/data/reuters/reuters.html

[11] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/

Num Doc with at least one Person entity mention	*461*	*123*	*109*	*693*
Num Person entity Annotations	*4413*	*1286*	*998*	*6,697*

3. EVALUATION OF OUR SYSTEM

Our NEL system takes input text with mentions of human names and maps them to their proper wikidata unique identifier, thus giving a disambiguated meaning to entity mentions in the text. The performance of our entity linking system is measured in terms of precision, recall, F1-measure, and accuracy. We only consider mention entity pairs where the ground-truth gives a known entity, and ignore around 20% of Unlinkable mentions in the ground truth. Also, our system has been designed to disambiguate only person entity mentions in text, therefore for evaluation we only consider the mentions which maps to human entities in wikidata.

Most of the available NEL systems are evaluated on CoNLL'03 dataset. The performance of our system is compared with the results of AIDA – a state of art NED system (Hoffart et al. 2011) and AIDA-light (Nguyen et al. 2014). Their results were produced on the CoNLL'03 testb dataset. Hence, we used the same testb dataset for our experiment. This dataset contains 231 news articles. However, person entity annotations were available from 109 documents out of 231. Therefore, our results are produced for 998 person entity mentions from 109 documents in AIDA CoNLL-YAGO testb dataset.

As described in (Shen, Wang, and Han 2015), the following equations were used to compute the performance of the proposed system.

The precision of an entity linking system is a fraction of correctly linked entity mentions that are generated by the system. It determines how correct entity mentions linked by the proposed entity linking system.

$$precision = \frac{correctly\ linked\ entity\ mentions}{linked\ mentions\ generated\ by\ system}$$

The recall of an entity linking system is a fraction of correctly linked entity mentions that should be linked. It determines how correct linked entity mentions are with regard to total entity mentions that should be linked.

$$recall = \frac{correctly\ linked\ entity\ mentions}{entity\ mentions\ that\ should\ be\ linked}$$

The third measure is the F1 score. It defines the harmonic mean of precision and recall.

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

4. RESULTS

The proposed candidate entity generation algorithm achieves the precision of 93.11%. The output produced in this step was a limited set of candidate entities that best matches from the knowledge base for each mention.

Table 4: Evaluation of proposed candidate entity generation algorithm on CoNLL-YAGO testb.

Dataset	*Precision*
CoNLL-YAGO	*93.11%*

Our entity linking system achieves the top1 precision of 77.1% (where top-1 means that the correct entity for the mention is being ranked at first place), recall of 76.25% and accuracy of 76.68%.

Table 5: Performance measure of our system on CoNLL-YAGO testb.

Dataset	*Top-1 Precision*	*Recall*	*F1-score*	*Top-5 Precision*
CoNLL-YAGO	*77.1%*	*76.25%*	*76.68%*	*84.4%*

We compare the performance of our system with AIDA: a state-of-art method for entity linking task (Hoffart et al. 2011). The top-1 precision of 77.1% achieved by our system is much close to the precision of 81.91% achieved by AIDA on the same AIDA CoNLL-YAGO testb dataset.

Table 6: top-1 precision of our NED on CoNLL.

Dataset	*AIDA*	*Our NED*
CoNLL-YAGO	*81.91%*	*77.1%*

Another NED system called AIDA-light (Nguyen et al. 2014) achieved top-5 precision of 95.2% on the same dataset, whereas our system achieves top-5 precision of 84.4% . Here, top-5 precision is a ratio of mentions whose ground truth candidate was among the 5 best candidate entities generated.

Table 7: top-5 precision of our NED on CoNLL.

Dataset	*AIDA-light*	*Our NED*
CoNLL-YAGO	*95.2%*	*84.4%*

5. CONCLUSION

Previous works on NEL systems have argued that NEL systems often suffer because of the named entity mention detection phase. Any missed mention by the NER system is also a missed entity for

NEL (Sil and Yates 2013). Therefore, we did not focus on NER step in this paper. State-of-the-art methods for named entity disambiguation face significant trade-offs regarding efficiency/scalability vs. accuracy. Our proposed system is able to handle the scalability issue in NEL. Also, the system can map human mention in the text to the relevant wikidata entity with top-5 precision of 84.4%. In order to handle the issue of scalability in NEL, we only made use of few features in document and knowledge base. Our future work will focus on improving accuracy of our system without compromising the scalability.

REFERENCES

Abel, Fabian, Qi Gao, Geert-jan Houben, and Ke Tao. 2011. "Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web." *Proceeding WebSci '11 Proceedings of the 3rd International Web Science Conference*: 1–8. http://journal.webscience.org/428/.

Abel, Fabian, Qi Gao, Geert Jan Houben, and Ke Tao. 2013. "Twitter-Based User Modeling for News Recommendations." *IJCAI International Joint Conference on Artificial Intelligence*: 2962–66.

Adomavicius, Gediminas, and Alexander Tuzhilin. 2005. "Towards the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions." *IEEE Transactions on Knowledge and Data Engineering* 17(6): 734–49.

Auer, Soren et al. 2007. "DBedpia: A Nucleus for a Web Od Open Data." *The emantic Web. Lecture Notes in Computer Science* 4825: 722.

Bizer, Christian, T Heath, and T Berners-Lee. 2009. "Linked Data-the Story so Far." *International journal on Semantic Web and Information Systems* 5(3): 1–22. http://eprints.soton.ac.uk/271285/.

Bollacker, Kurt et al. 2008. "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge." *SIGMOD 08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*: 1247–50. http://doi.acm.org/10.1145/1376616.1376746.

Cornolti, Marco, Paolo Ferragina, and Massimiliano Ciaramita. 2013. "A Framework for Benchmarking Entity-Annotation Systems." *Proceedings of the 22nd international conference on World Wide Web - WWW '13* (September 2014): 249–60. http://dl.acm.org/citation.cfm?doid=2488388.2488411.

Dimitrova, VG, A Piotrkowicz, J Otterbacher, and K Markert. 2017. "Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook." (Icwsm): 656–59. http://eprints.whiterose.ac.uk/115024/.

Hoffart, Johannes et al. 2011. "Robust Disambiguation of Named Entities in Text." *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*: 782–792. http://dl.acm.org/citation.cfm?id=2145432.2145521%5Cnhttp://www.aclweb.org/anthology/D11-1072.

Kanoje, Sumitkumar, Sheetal Girase, and Debajyoti Mukhopadhyay. 2014. "User Profiling Trends, Techniques and Applications." *International Journal of Advance Foundation and Research in Computer* 1(11): 2348–4853.

Lee, Won-Jo, Kyo-Joong Oh, Chae-Gyun Lim, and Ho-Jin Choi. 2014. "User Profile Extraction from Twitter for Personalized News Recommendation." In *16th International Conference on Advanced Communication Technology*, Global IT Research Institute (GIRI), 779–83. http://ieeexplore.ieee.org/document/6779068/.

Lewis, David D., Yiming Yang, Tony G. Rose, and Fan Li. 2004. "RCV1: A New Benchmark Collection for Text Categorization Research." *Journal of Machine Learning Research* 5: 361–97.

Liu, Jiahui, Peter Dolan, and Elin Rønby Pedersen. 2010. "Personalized News Recommendation Based on Click Behavior." (February).

Lu, Zhongqi et al. 2016. “Collaborative Evolution for User Profiling in Recommender Systems.” *IJCAI International Joint Conference on Artificial Intelligence* 2016–Janua: 3804–10.

Luo, Gang, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. “Joint Entity Recognition and Disambiguation.” *Proceedings of EMNLP* (September): 879–888. http://aclweb.org/anthology/D15-1104.

Nguyen, Dat Ba, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014. “AIDA-Light: High-Throughput Named-Entity Disambiguation.” *CEUR Workshop Proceedings* 1184.

Özgöbek, Özlem, Jon Atle Gulla, and R. Cenk Erdur. 2014. “A Survey on Challenges and Methods in News Recommendation.”

Plumbaum, Till. 2015. “User Modeling in the Social.”

Schiaffino, Silvia, and Analía Amandi. 2009. “Intelligent User Profiling.” *Artificial Intelligence LNAI* 5640: 193–216. https://pdfs.semanticscholar.org/d503/ee64901b3eb5edfb962041ff252b6545cd27.pdf (May 22, 2017).

Shen, Wei, Jianyong Wang, and Jiawei Han. 2015. “Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions.” *IEEE Transactions on Knowledge and Data Engineering* 27(2): 443–60.

Sil, Avirup, and Alexander Yates. 2013. “Re-Ranking for Joint Named-Entity Recognition and Linking.” *Conference on Information and Knowledge Management*: 2369–74. http://dl.acm.org/citation.cfm?id=2505601%5Cnhttp://www.cis.temple.edu/~yates/papers/2013-cikm-joint-nerd.pdf.

Suchanek, Fabian et al. 2017. “Yago : A Core of Semantic Knowledge Unifying WordNet and Wikipedia.”