Entity Linking and Knowledge Discovery in Microblogs

Pikakshi Manchanda

Department of Computer Science, Systems and Communication,
Università di Milano-Bicocca, Milano, Italy
pikakshi.manchanda@disco.unimib.it

Abstract. Social media platforms have become significantly popular and are widely used for various customer services and communication. As a result, they experience a real-time emergence of new entities, ranging from product launches to trending mentions of celebrities. On the other hand, a Knowledge Base (KB) is used to represent entities of interest/relevance for general public, however, unlikely to cover all entities appearing on social media. One of the key tasks towards bridging the gap between Web of Unstructured Data and Web of Data is identifying such entities from social media streams which are important and haven't been yet represented in a KB. The main focus of this PhD work is discovery of new knowledge from social media streams in the form of new entities and/or new mentions of existing entities while enriching KBs as well as lexically extending them for existing entities. Based on the discovery of new entities or new mentions, structured data in the form of RDF (Resource Description Framework) can be extracted from the Web.

Key words: Social Media, Knowledge Base, Web of Data, RDF

1 Scene Setting

1.1 Objective of the Research

Microblogging platforms have become an indispensable resource for users by providing services such as sales and marketing, news and communication, trend detection and a variety of customer services. Due to their dynamic nature, they experience a steady emergence of new knowledge in the form of new entities (such as product launches), new relations between existing entities (such as a football player playing for *FC Barcelona* and *Real Madrid*), as well as new/popular mentions of existing entities (such as trending colloquial names for celebrities). Knowledge bases provide a broad (yet intrinsically non exhaustive) coverage of a variety of entities found on the Web and social media streams. However, it is unlikely that a KB can provide coverage of all new entities that emerge constantly on the Web. As a result, tasks such as Named Entity Recognition (NER), Disambiguation (NED) and Linking (NEL) have gained significant attention of NLP practitioners. Named entity recognition is the task of identifying a piece of

text as a named entity and classifying into types such as person, location, organization etc. whereas a named entity disambiguation task is to disambiguate a named entity with a resource in a KB and finally link it with the said resource.

In order to enrich a KB for new/relevant entities emerging on social media in real-time, it is necessary to identify those entities and gather contextual information from the Web and social media. The objective of this work is not only to identify and extract new knowledge, but also being able to use it in order to enrich and lexically extend KBs. In the process, we will be able to improve the accuracy of named entity recognition as well as disambiguation tasks.

1.2 Research Questions

The proposed research work aims to address the following research questions:
RQ1: Can we perform NER and NEL in microposts as a joint task and link the named entities to resources in a KB?
RQ2: Is it possible to use the results of an Information Extraction (IE) task to identify new entities?
RQ3: Can we use an enriched/lexically extended KB to improve the IE process of new entities from microblogging platforms?

1.3 Motivation and Relevance

Significant gain in momentum for IE (achieved mainly through NER and NEL), from news archives, blogs and Web pages, is attributed to need for bridging the gap between Document Web and Web of Data. The main motivation for carrying out a research on discovery of new knowledge by means of IE tasks is primarily because new entities emerge frequently over social media. Another motivating factor is being able to perform entity recognition and disambiguation on short textual formats, such as microblogs, as a joint task in an end-to-end entity linking pipeline. This is also important from the point of view of KB enrichment and its lexical extension for existing entities.

KB Enrichment can be performed automatically to some extent (by identifying a new entity, and collecting contextual information from the Web) or can even be performed interactively, for instance, driven by social content creation communities. The output of my research work combined with these techniques can be used to enrich KBs periodically. Furthermore, a lot of research (Semantic Search, Recommendation Systems, Disaster Discovery, Sentiment Analysis) is dependent on entity disambiguation and discovery of new knowledge.

1.4 Challenges and Opportunities

Challenges: The task of identification and disambiguation of entities from microblogs is challenging due to the following reasons:

- *Short, noisy nature*: An informal microblogging style, coupled with use of Internet slang and misspellings [4, 7, 8] renders it difficult to identify new entities, affecting the *accuracy* of entity recognition and disambiguation.

– *Occurrence of Out Of Vocabulary (OOV) mentions*: We define an OOV mention as an existing resource in a KB, being referred by an alternate entity mention in social media which is not present in KB. As a result, OOV mentions can't be disambiguated, causing the performance accuracy of an end-to-end entity linking system to suffer.
– *Occurrence of Out of Knowledge base (OOKB) entities*: We define an OOKB entity as one which is not covered by a KB and, thus, can be considered as newly emerging.

Opportunities: If we are able to identify an OOV mention, we can *lexically enrich* the KB for said existing entity. Similarly, if we are able to detect an OOKB entity, we can *extensionally update* the KB for the new entity by collecting contextual information about it from the Web. On the other hand, by addressing the above challenges, we will also be able to improve the accuracy of the end-to-end entity linking pipeline.

2 Proposed Approach

2.1 Formal Definition and Properties of the Approach

Given a tweet t, the goal of the system is to identify named entities in t. Further, the system maps every identified entity e to a referent resource r in knowledge base K. More formally, we define a Named Entity Recognition task as a function which identifies and maps a set of words W in tweet t to a tuple of entity name, e_i, and a corresponding entity type, $type_{e_i}$, i.e.,

$$f_{NER} : W \rightarrow < e_i^t, type_{e_i}^t > \tag{1}$$

Next, we define a universe U consisting of entities present in unstructured/semi-structured data on social media and the Web as well as resources covered by KBs. Further, we define a Named Entity Linking task as a function which maps an identified entity e_i^t, as in equation (1), to a resource r_j in K, i.e.,

$$f_{NEL} : e_i^t \rightarrow r_j^K \tag{2}$$

Here every resource r_j in K can be associated with one or more resource types and is represented as c_{r_j}. f_{NEL} is defined for entities which are covered by resources in K. OOV mentions also have referent resources in K, however, the said mention has been referred in social media by an alternate name while K isn't lexically updated to provide coverage for it. It is to note here, that OOV mention, its original entity as well as the corresponding resource are present in U, however, f_{NEL} is unable to link the OOV mention to the corresponding resource. On the other hand, OOKB entities are new entities, present in U, which have not yet been covered by K and so f_{NEL} is unable to link them as well.

2.2 Relationship between your approach and state-of-art approaches

Various existing approaches [1, 5, 9] as well as a variety of commercial tools, such as Zemanta[1], Alchemy API[2], and DBpedia Spotlight[3] are used for entity recognition in text. However, these conventional tools perform poorly on short textual data [1], mainly due to lack of context and informal language. [8] propose a tweet-based NLP framework for entity recognition in tweets using a CRF model with the help of contextual, dictionary and orthographic features. In [5], Liu et al. (2011) propose an entity recognition framework using K-Nearest Neighbour (KNN) Classifier with a linear CRF Model.

State-of-the-art approaches provide a variety of methods for entity disambiguation. However, few existing approaches target the detection of new entities using existing knowledge provided by KBs. Liu et al (2013) use similarity measures to detect OOV mentions of existing entities [4], however, OOKB entities are not dealt with. [2] propose an end-to-end tweet-level identification and disambiguation system while using structural learning techniques to jointly optimize identification as well as disambiguation. However, their approach is not able to recognize or deal with OOKB entities. An approach for discovery of emerging OOKB entities with ambiguous names from documents has been proposed in [3]. This work is, in principle, a foundation for our research work, however, their approach doesn't consider the entities emerging in social media streams.

Based on the literature review, we observe that most state-of-the-art systems treat entity identification and disambiguation as separate tasks. In this research work, we propose an end-to-end entity linking pipeline where we study entity recognition and disambiguation as a joint problem for microposts. To the best of our knowledge, our work provides a novel contribution, in the sense, that we not only aim to improve the disambiguation of entities using linked datasets, but also we address the task of discovery of new entities from tweets, thus improving the overall accuracy of the system. We distinguish between OOV mentions and OOKB entities and also propose distinctive measures to deal with both types of entities. We use the discovered information for KB enrichment.

3 Implementation of the Proposed Approach

3.1 The Big Picture and Current Implementation

In this section, we present a brief overview of the system, as shown in Fig. 1. The system performs tweet-wise evaluation using Ritter et al's (2011) state-of-the-art T-NER system [8] for entity recognition and classification. Further, for NED, we have constructed an inverted index of the data property **rdfs:label** from DBpedia[4] which we currently consider for disambiguation and knowledge discovery.

[1] `http://www.zemanta.com/`
[2] `http://www.alchemyapi.com/`
[3] `http://dbpedia.org/spotlight/`
[4] `http://wiki.dbpedia.org/`

For every identified entity in a tweet, an ad-hoc index lookup is performed to obtain a list of candidate resources which we rank using a high-recall lookup approach. We use contextual and orthographic features, identified entity-type,

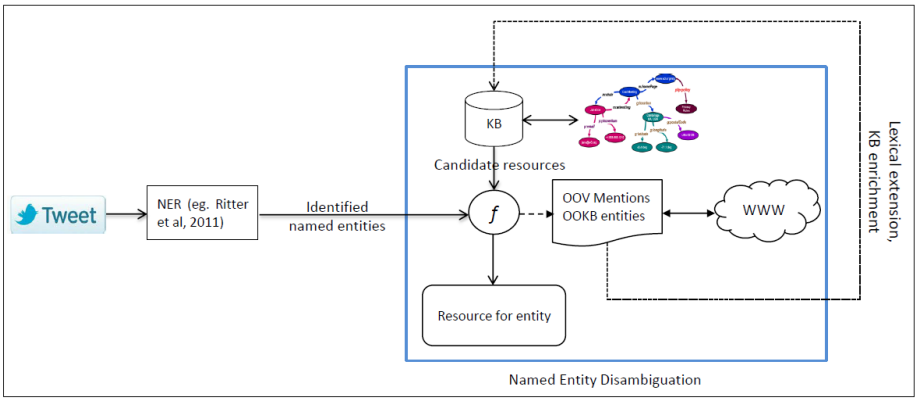

Fig. 1. Framework of the proposed model

as well as tweet-specific features such as use of @usernames, #hashtags and URLs which aid in disambiguation amongst the candidate resources. This is accomplished using a probabilistic matching function, being presented in this work, which takes into consideration the following factors:

1. Lexical Similarity between an entity in a tweet and candidate resource
2. Coherence between an entity and (structured) document page of candidate resource in KB
3. Relatedness between entities in a tweet, in case where there is more than one entity in a tweet

Currently, we have implemented measures to calculate similarity between entity in a tweet and candidate resources, as well as relatedness between entities (from KB perspective, i.e., how frequently entities mentioned in a tweet co-occur in a KB). In the future, we plan to take into consideration the relatedness between entities from real-world perspective.

The probabilistic matching function helps to disambiguate named entities with resources in KB. Subsequently, we will obtain a pool consisting of entities which can't be disambiguated. This pool will consist of OOV mentions, OOKB entities as well as noise (text wrongly identified as an entity). Information, such as usage patterns, frequency of usage, as well as contextual patterns from social media streams and the Web, will be collected for entities in pool. This information can be used to enrich a KB, either automatically or manually by content creation communities, with the help of which disambiguation is performed again to improve the overall accuracy of disambiguation process of the proposed system.

4 Empirical Evaluation Methodology

4.1 General Strategy

The research questions described above are related to a few hypotheses:
H1: If an entity is a word that appears in the lexicon of a resource, the system links it with the resource with a certain degree of accuracy. For this, we use entity information from tweet and resource in KB. In order to accomplish this task, we perform NER and NEL jointly (explained in detail in section 4.2). A NER system exhibits segmentation errors (such as *St. Mary's* identified as 2 distinct entities), identification erros (such as *justten* being identified as an entity) and classification errors (such as *Hawaii* being identified as Person). We use Ritter et al.'s (2011) gold standard corpus of 2400 tweets for NER. Additionally, we created a manually annotated gold standard collection of named entities for NEL from gold standard corpus used for NER.

H2: If there is a pool of unknown entities, we collect additional knowledge, from the Web and social media, in order to classify them as new (OOKB) entities or use that knowledge to resolve (OOV) entity mentions and link them with resources in KB. A gold standard corpus of such unknown entities needs to be created for this step. We can also use the pool of entities from NEL's gold standard which aren't disambiguated. We plan to expand this gold standard in the future.

4.2 Current State of the Evaluation

In this section, we present the evaluation results achieved so far for hypotheses H1. We plan to start creating a gold standard for H2 by December 2015.

H1-Task1: Entity Recognition (Experimental Analysis of T-NER)
Using Ritter et al.'s (2011) gold standard corpus of 2400 tweets, T-NER identifies a total of 1496 named entities classified into 10 distinct entity types (person, location, organization,..), in contrast to 1612 named entities as found in the ground truth. T-NER exhibits an identification error rate of 9.62%, whereas segmentation error rate is negligible. We summarize the classification error rate of every entity type in Table 1 below. As is evident, the classification error rate is quite high for entity types *Movie* and *Band*. A significant reason for this could be attributed to out-of-date knowledge utilized by T-NER for entity recognition.

H1-Task2: Entity Disambiguation (Experimental Analysis of Lexical Similarity Measure and Relatedness)

In this step, we use the set of named entities identified in Task 1 and based on an *ad-hoc candidate match retrieval approach*, we obtain candidate resources for these named entities from our index of rdfs:label. A manually annotated gold

Table 1. Classification Error rate for T-NER

Entity Type	Error (%)
Band	73.83
Company	21.9
Facility	54.79
Geo-Location	19.75
Movie	75.83
Other	46.29
Person	28.18
Product	39.70
Sportsteam	48.27
TVshow	48.71

Table 2. Precision-Recall of named entities with candidate matches

Entity Query Representation	P(%)	R%)
Entity Mention	92	95
Entity Mention and type	87	99
Combined Entity Mentions	31	24

standard of 1455 named entities is created out of 1496 entities that were identified in Task 1 to aid in candidate match retrieval. The remaining entities serve as a pool of unknown entities and need to be further expanded in order to be used as a gold standard for Task 3 described below. We have experimented with varying forms of entity representations (only entity mention, entity mention with its entity type, and a combination of entity mentions) in order to obtain sufficient number of candidate matches for each named entity. Table 2 summarizes the precision-recall for varying entity representations.

The first representation produces a list of candidate resources (highest precision) for disambiguation. Second representation produces a list with the highest recall (fetching noisy results as well), however, there is a decrease in precision. The reason for such an output can be due to knowledge gaps in KB for specific entity types (thus, a justified need for KB enrichment). The third representation is for tweets which have more than one entity. This representation exhibits the lowest precision as well as recall amongst all three. This can be due to infrequent occurrence of various entities together in social media, thus making it difficult to find sufficient evidences of their co-existence in a KB.

We implement a lexical similarity measure using *Lucene's Vector Space Model of Information Retrieval* to estimate similarity between an entity and a candidate resource so as to choose the most suitable resource for an entity. We have also used a relatedness measure in order to estimate co-occurrence frequency between two entities in a tweet, using a method described in [6]. Currently, we have implemented this measure from KB perspective, i.e., how often entities in a tweet can co-occur in a KB.

We found a total of 399 tweets in Ritter et al's dataset which have more than one entity. A high relatedness score depicts presence of a strong evidence in the KB that said entities co-occur frequently. Use of relatedness measure is attributed towards the need of improving the accuracy of disambiguation for infrequent/long-tail entities found in social media streams. Another significant reason for the use of this measure is in identifying an OOV entity mention.

H2-Task3: OOV Mention/OOKB Entity discovery
Discovery of OOV mentions as well as OOKB entities depends to a great extent on the performance accuracy of entity recognition as well as disambiguation. Herein, we propose to improve entity recognition by improving entity disambiguation, which is currently under progress. In order to achieve this, we use features (contextual information, evidences from KB, relatedness of an entity with other real-world entities) for entity recognition that are conventionally being used for entity disambiguation in the state-of-the-art. By improving entity recognition, the overall accuracy of the system will be improved.

5 Lessons Learned, Open Issues, and Future Directions

It is essential to discover new entities for the enrichment of KBs. However, one of the important lessons that we have learned is that, not every entity that has been discovered can be updated in a KB. Its authenticity needs to be verified as well as its relation and relevance w.r.t other entities in the real world has to be taken into consideration to update concepts in KBs.

Enrichment of KBs, specifically enriching the lexicon of an entity in a KB using information extracted from social media is one of the most important open issues in the Semantic Web community. As of now, we have conducted a variety of experiments for improving disambiguation. While we continue to improve it, the next step in this research work is working towards enrichment of KBs in time and extending them with quality information extracted from Social Media and the Web.

References

1. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. Analysis of named entity recognition and linking for tweets. *Information Processing & Management* (2015).
2. Guo, S., Chang, M.-W., and Kiciman, E. To link or not to link? a study on end-to-end tweet entity linking. In *HLT-NAACL* (2013).
3. Hoffart, J., Altun, Y., and Weikum, G. Discovering emerging entities with ambiguous names. In *Proceedings of Conference on WWW* (2014).
4. Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., and Lu, Y. Entity linking for tweets. In *ACL (1)* (2013).
5. Liu, X., Zhang, S., Wei, F., and Zhou, M. Recognizing named entities in tweets. In *Proceedings of ACL: Human Language Technologies* (2011).
6. Medelyan, O., Witten, I. H., and Milne, D. Topic indexing with wikipedia. In *Proceedings of AAAI WikiAI workshop* (2008).
7. Meij, E., Weerkamp, W., and de Rijke, M. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining* (2012).
8. Ritter, A., Clark, S., Etzioni, O., et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on EMNLP* (2011).
9. Usbeck, R., Ngonga Ngomo, A.-C., Luo, W., and Wesemann, L. Multilingual disambiguation of named entities using linked data. In *International Semantic Web Conference* (2014).