Leveraging Entity Linking to Enhance Entity Recognition in Microblogs

Pikakshi Manchanda, Elisabetta Fersini and Matteo Palmonari
DISCo,University of Milano-Bicocca, Viale Sarca 336, 20126, Milan, Italy

Keywords: Web of Data, Information Extraction, Named Entity Recognition, Named Entity Linking, Knowledge Base, Microblogs.

Abstract: The Web of Data provides abundant knowledge wherein objects or entities are described by means of properties and their relationships with other objects or entities. This knowledge is used extensively by the research community for Information Extraction tasks such as Named Entity Recognition (NER) and Linking (NEL) to make sense of data. Named entities can be identified from a variety of textual formats which are further linked to corresponding resources in the Web of Data. These tasks of entity recognition and linking are, however, cast as distinct problems in the state-of-the-art, thereby, overlooking the fact that performance of entity recognition affects the performance of entity linking. The focus of this paper is to improve the performance of entity recognition on a particular textual format, viz, microblog posts by disambiguating the named entities with resources in a Knowledge Base (KB). We propose an unsupervised learning approach to jointly improve the performance of entity recognition and, thus, the whole system by leveraging the results of disambiguated entities.

1 INTRODUCTION

Recent years have witnessed an increasing interest amongst the scientific community towards making sense of data extracted from a variety of sources such as news articles, blog posts, and web pages. In particular, social media platforms, such as Twitter[1], have gained popularity amongst general masses and have become a choice of interest for Information Extraction (IE). Social media platforms provide a steady stream of fresh information in real-time, which may unveil new valuable pieces of knowledge in the form of new or suddenly relevant entities (e.g., a novel movie or a previously unknown singer), new relations (e.g., a new business partnership), or classification updates (e.g., a politician becomes President of a country). Intrinsic incompleteness of Knowledge Bases (KBs) w.r.t. new knowledge (Rula et al., 2012; Rula et al., 2014) is one of the significant reasons to perform IE from Unstructured Web (social media platforms, web pages, online health records..) (Guo et al., 2013).

IE from any form of text is associated with, first, identifying the information to be extracted, e.g., mentions of entities or relations between entities (Ritter et al., 2011; Liu et al., 2011; Hoffart et al., 2014), and then disambiguating its meaning, using, for example, a large target KB. If we focus on entities, the essential tasks in an IE framework are Named Entity Recognition (NER) and Named Entity Linking (NEL). NER has been studied under two subtasks in state-of-the-art: entity identification, i.e., identifying text fragments as **surface forms** that refer to entities in the real world, and entity classification, i.e., classifying the surface forms into a set of classes/entity types such as person, location, product, organization. Whereas the NER task returns a set of entity mentions and their predicted types, these entities may denote real world objects that are also represented in KBs such as Wikipedia, YAGO, DBpedia, or Freebase. A KB describes real-world objects by specifying their types and their relations with other entities. In the NEL task, a surface form is linked to a resource, i.e., a KB instance, which is used to represent the real-world object referred by the surface form.

Due to polysemy of natural language expressions, a key step in NEL is entity disambiguation, i.e., the task of finding the correct match for the surface form in the KB. For instance, surface forms '*J.Lo, Jennifer Lopez, Jennifer Lopez Muñiz*' refer to resource **dbpedia:Jennifer_Lopez**, whereas a surface form '*paris*' can refer to resources **dbpedia:Paris** (the capital city of France) or **dbpedia:Paris_Hilton** (the American Socialite).

[1]https://twitter.com/

Manchanda, P., Fersini, E. and Palmonari, M..
Leveraging Entity Linking to Enhance Entity Recognition in Microblogs.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 147-155
ISBN: 978-989-758-158-8

NER and NEL form the basis of a typical end-to-end entity linking pipeline and are widely-studied phases of this pipeline. These phases have been often investigated separately as well as two independent tasks performed sequentially (Ritter et al., 2011; Cucerzan, 2007; Damljanovic and Bontcheva, 2012; Usbeck et al., 2014; Liu et al., 2013). The performance of NER systems on well formed texts has significantly improved over the years and is now sufficient for many real world applications.

On the other hand, short texts such as microblog posts make extensive use of informal language, emoticons, Internet slangs and abbreviations leading to lesser contextual information and more noise (Ritter et al., 2011; Derczynski et al., 2015). Moreover, infrequent occurrence of a variety of named entity types makes it difficult to classify named entities correctly. These difficulties result in lower performance of NER systems on short textual formats such as microblog posts. To overcome this limitation, one of the approaches proposed in the state-of-the-art has been to consider NER and NEL as two co-dependent tasks (Yamada et al., 2015), thus achieving a significant improvement in NER accuracy.

We believe that the latter work opened an interesting research direction, i.e., the orchestration of NER and NEL in an end-to-end IE framework for microblog posts. This research direction is still unexplored to a large extent. For instance, several recognized named entities may not be covered by the target KB, giving rise to the problem of dealing with several unlinkable entities. In addition, the difficulty of NEL task on microblog posts may suggest to keep a local base of extracted named entities, together with their predicted links to KB resources. Even when the link is uncertain, information collected in the the NEL step can be useful to improve the output of the NER task. For example, even if it is not possible to disambiguate a surface form such as *Paris* with one of the eight locations named *Paris* described in DBpedia, this information can reinforce the classification of the recognized entity named *Paris* as a Location.

In this paper we provide several contributions to the problem of orchestrating NER and NEL for end-to-end IE on microblog posts:

- As a main contribution, we present an end-to-end IE approach where the output of the NEL task is fed back as input to the NER task in order to improve it. In particular, differently from (Guo et al., 2013), we use the feedback to improve the classification of named entities into entity types. Our approach integrates a state-of-the-art NER system and improves its classification using a greedy entity linking method.
- We provide a new version of a gold standard used in previous work on NER for microblogs where named entities in the gold standard have been re-annotated.
- We provide an experimental evaluation of our work which also provides interesting insights on the problem of orchestrating the NER and NEL tasks.

With respect to the latter item, we found out that the gold standard contains a significant number of non linkable entities, which motivates the preservation of the NER output as independent from the NEL output. In addition, classification errors occur with a higher percentage than identification errors on microblog posts, which makes it valuable to focus on better classification models. Finally, a small percentage of linkable entities have surface forms that do not get any match in the KB, which may suggest that applying NER over tweets can help to extend the vocabulary of knowledge bases. Overall, we believe that our work provides one of the first steps towards automatic strategies to make KBs keeping up the pace with changes in the world, using microblog posts as information sources.

The rest of this paper is structured as follows: Section 2 provides details about the related work. Our proposed end-to-end entity linking pipeline is described in Section 3. Section 4 presents our experimental set-up and results of the afore-mentioned contributions, followed by Conclusion in Section 5.

2 RELATED WORK

Named Entity Recognition and Named Entity Linking have gained a significant interest within the research community over the years. This can be attributed to the need of bridging the gap between unstructured data on the Document Web and structured data on the Web of Data. As a result, a variety of annotation frameworks, entity recognition as well as entity linking systems have been proposed in the state-of-the-art to address these tasks. We provide a brief description of these systems in the following subsections.

Named Entity Recognition

This section briefly describes related work in the field of NER. Research in the field of entity recognition and linking for different genres, such as microposts, blogs, news archives, has been quite recent and is fast becoming a largely employed

method of making sense of user-generated content on the Web.

(Ritter et al., 2011) have proposed a tweet-based NLP framework to perform tasks such as POS tagging, shallow parsing and named entity recognition in tweets by using Conditional Random Field (CRF) Model with the help of conventional features such as orthographic, contextual and dictionary features as well as tweet-specific features such as using retweets, @usernames, #hashtags, URLs, outperforming various state-of-the-art entity recognition systems for short textual formats, as reported by (Derczynski et al., 2015).

(Liu et al., 2011) have proposed an entity recognition framework by using a K-Nearest Neighbours (KNN) Classifier with a linear CRF Model for identifying named entities in tweets. On the other hand, crowd-sourcing services (such as CrowdFlower and Amazon's Mechanical Turk) have also been used (Finin et al., 2010) for annotating named entities in microposts.

Commercial tools such as DBpedia Spotlight[2], Lupedia[3], and TextRazor[4] as well as NER systems such as ANNIE (Cunningham et al., 2002) and Stanford NER (Finkel et al., 2005) are available for performing entity recognition. However, the performance of these commercial tools as well as conventional NER systems on microblog textual formats is relatively poorer than their performance on longer textual formats such as news and blogs as reported by (Derczynski et al., 2015). One of the main reasons for such performance is lack of context, as well as an infrequent occurrence of a variety of entity types causing supervised learning based NER systems to exhibit poor accuracy in entity recognition, leading to misclassification errors.

Named Entity Linking

While dealing with entity linking, one of the major concerns of NEL systems is being able to disambiguate and link surface forms of entity mentions extracted from a piece of text to canonical resources in a KB. A surface form's contextual information as well as the coherence between a surface form (and its context) and descriptions of candidate KB resources is highly useful for entity disambiguation (Milne and Witten, 2008) and, thus, efficient linking. However, microposts are more challenging for entity linking than other textual formats, mainly due to insufficient or ambiguous contextual information, presence of polysemous, multilingual entities (Usbeck et al., 2014), and continuous emergence of new entities and entity types (Hoffart et al., 2014).

Several approaches have been proposed in the state-of-the-art to address entity linking. (Mendes et al., 2011) present DBpedia Spotlight for annotating and disambiguating Linked Data Resources in any textual format making use of DBpedia[5] ontology. (Ferragina and Scaiella, 2010) present an annotation and disambiguation system called TAGME[6] for annotating short texts with links to Wikipedia articles using Wikipedia *anchor texts* and the *pages* linked to those anchor texts. YODIE[7] is another entity linking system (Damljanovic and Bontcheva, 2012) for linking entities to DBpedia URIs using an amalgamation of similarity metrics which include string similarity, semantic similarity between entities in tweets, contextual similarity and URI frequency of Wikipedia articles. (Meij et al., 2012) propose a machine learning based approach using n-gram features, concept features, and tweet features in order to identify concepts that are semantically related to a tweet, thereafter, generating links to Wikipedia articles for every entity mention in a tweet. (Ibrahim et al., 2014) propose a framework called AIDA-Social for accommodating microblogs. They employ the techniques of mention normalization, contextual enrichment, and temporal entity importance in order to disambiguate and link entity mentions to a knowledge base.

The main limitation of the above mentioned approaches relates to the assumption that named entities are provided by an oracle, and therefore, always correct. However, in a real context, entities can be wrongly identified and/or wrongly classified, leading to poor performance of the linking systems. To the best of our knowledge, only one recent approach has been proposed to deal with NER and NEL as dependent tasks (Yamada et al., 2015). The authors propose to find entity mentions in a tweet by exploiting a "dictionary" derived by Wikipedia. Once a candidate entity mention has been found, and its possible referent entities have been identified by string matching with Wikipedia resources, a random forest is adopted to learn the patterns underlying the correct linking. The prediction provided by random forest are then exploited in a further random forest that, together with other features such as number of in-bound links and average page view, predicts the named entity type.

Our approach differs from (Yamada et al., 2015) from several points of view. First of all, our goal is to define a loop-based system that combines NER and

[2]http://dbpedia.org/spotlight
[3]http://lupedia.ontotext.com
[4]http://www.textrazor.com
[5]http://wiki.dbpedia.org/
[6]http://tagme.di.unipi.it/
[7]https://gate.ac.uk/applications/yodie.html

NEL predictions while (Yamada et al., 2015) use NEL to learn how to perform NER. Secondly, our system is mostly unsupervised while their approach requires a strong effort for labeling data to train the supervised models enclosed in the system. Finally, our system is able to deal with the emerging entity mentions that are likely to be out-of-vocabulary of KB.

In this paper, we followed the paradigm introduced by (Yamada et al., 2015) and propose an end-to-end entity linking system, i.e., we study entity recognition and linking as a joint problem for microposts (short textual formats) by re-casting and merging the processes of recognition and linking. We aim to improve the overall accuracy of the system by improving the performance of entity recognition (in particular, entity classification) by leveraging the results of disambiguated entities. In the next section, the proposed approach will be detailed.

3 PROPOSED SYSTEM

In this section, we introduce our methodology for an end-to-end entity linking scenario. We define a three-step approach as depicted in Figure 1:

- **Step 1:** Entity Recognition and Classification: segmentation and annotation of named entities through Conditional Random Fields.
- **Step 2:** Candidate Match Retrieval and Disambiguation: Look-up of candidate KB resources for identified surface forms and disambiguating and linking each surface form with at most one suitable candidate KB resource (associated with a corresponding KB type).
- **Step 3:** Entity Recognition Enhancement: Reclassification of surface forms based on information extracted from selected resources.

NER systems often refer to smaller ontologies when classifying entities. On the other hand, KBs such as DBpedia[8], use larger ontologies. Our approach assumes that classes of the NER system are mapped to classes of the KB ontology. We summarize the ontologies in Table 1 as well as mappings between classes in the NER Ontology (in this case, T-NER Ontology) and classes in the KB Ontology (in this case, DBpedia Ontology).

Table 1: Mapping between T-NER and DBpedia Ontologies.

T-NER Ontology	DBpedia Ontology
Band	Band, MusicGroup
Company	Company, Business
Facility	Award, EducationalInstitution, WebSite, SportFacility, ...
Geo-Location	Place, Location, PopulatedPlace, Country, City, Locality, Region, Park, ...
Movie	Film
Other	MeansOfTransportation, Holiday, ArtWork, Cartoon, Species, Food, Event, ...
Person	Person
Product	VideoGame, MusicalWork, Software, Album, Device, ...
Sportsteam	Sportsteam, SportsClub
TVshow	TelevisionShow, TelevisionEpisode, TelevisionSeries

Step 1: Entity Recognition and Classification

In order to identify named entities in the first step, we use a state-of-the-art entity recognition system (Ritter et al., 2011) already trained on Twitter data. The system, named T-NER, has been grounded on Conditional Random Fields (CRF) (Sutton and McCallum, 2006) and performs segmentation of a tweet while subsequently classifying each token according to the classes of the ontology reported in Table 1. T-NER uses IOB encoding (i.e., each word is either inside/outside/beginning of a named entity) for named entity segmentation before classifying the named entities. As mentioned before, T-NER is based on a linear chain CRF model, which is an undirected probabilistic graphical model for segmenting and labeling text. Based on this model, every named entity e can belong to one or more entity type/class c, each associated with a probability score denoted as $\mathrm{P}_{CRF}(c,e)$, which illustrates the probability that a named entity e belongs to an entity type/class c. P_{CRF} is defined as follows:

$$P_{CRF}(e,c) = \exp(\sum_{k=1}^{K} w_k f_k(e,c)) \qquad (1)$$

where w_k are the weights learned from data and f_k are the feature functions encoded by CRF. The probability reported in equation (1) represents an *apriori* estimation of the entity type that will be exploited in the subsequent phases. Concerning the time complexity, this step requires $O(T \times |C|^2)$ for each tweet, where T denotes the length of the tweet and $|C|$ is the number of entity types.

Step 2: Candidate Match Retrieval and Disambiguation

The surface forms e identified in step 1 are disam-

[8]http://mappings.dbpedia.org/server/ontology/classes/

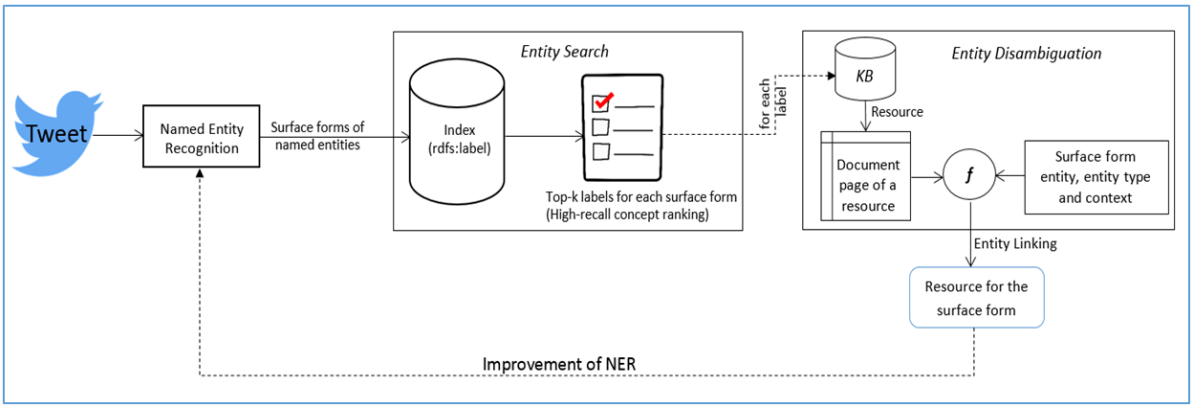

Figure 1: Framework of the proposed system.

biguated and linked in this step with suitable KB resources. Firstly, we retrieve a set of candidate KB resources to be matched with each input surface form, using a high-recall approach. We use the notation r_c to denote a resource r retrieved from the KB that is member of a class c. Secondly, to link a surface form e with an appropriate KB resource, we estimate a similarity score $P_{KB}(e, r_c)$ between the surface form e and every retrieved resource r_c by leveraging information about r_c extracted from the KB, namely its label l_{r_c} and its semantic description d_{r_c}. We use DBpedia as an external knowledge base for candidate match retrieval, entity disambiguation and linking. The similarity score $P_{KB}(e, r_c)$ comprises of a Lexical Similarity measure (lex) and a Coherence measure (coh) and is estimated as an average over these measures, as reported in equation (2).

$$P_{KB}(e, r_c) = \frac{lex(e, l_{r_c}) + coh(e^+, d_{r_c})}{2} \quad (2)$$

Lexical Similarity measure: denotes a similarity score between a surface form e and a label of a candidate match l_{r_c}. The scoring is established using Lucene's Scoring function[9] and is performed in order to filter out false positives obtained due to high recall.

Coherence measure: implements a cosine similarity between a surface form context and a resource description, making use of a Vector Space Model representation (Cohen et al., 2003). In particular, the contextual information of surface form e, denoted as e^+, comprises of the entity type c derived by the T-NER system, the surrounding content (i.e. nouns/verbs/adjectives) available in the tweet and the surface form itself [10].

[9]https://lucene.apache.org/core/4_6_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

[10]The entity type and surrounding contents are useful for entity resolution in case of polysemous entities.

The entire candidate matching and disambiguation step requires $O(B \times V)$ for each surface form e, where B denotes the number of KB entries and V represents the length of the Vector Space Model representation used for retrieving entries in the KB.

Step 3: Entity Recognition Enhancement

In this step, we determine the DBpedia Ontology classes of all top-k candidates collected in the previous step. The Ontology classes for the candidates are, further, mapped to entity types based on T-NER Ontology, as shown in Table 1. Using the DBpedia Ontology for each candidate obtained in step 2, we predict the most suitable entity type of each surface form by making use of the *apriori* estimation provided by the T-NER system smoothed by the similarity score derived through the KB. In particular, the most probable entity type c_e^* is determined according to the following decision rule:

$$c_e^* = \underset{c}{\text{argmax}}\{P_{CRF}(c, e) * P_{KB}(e, r_c)\} \quad (3)$$

where c_e^* denotes the new entity type/class of the reclassified surface form e which may or may not be the original class c as estimated by the T-NER system. We discuss various scenarios in Section 4 where we use this function in order to improve the performance of entity recognition, thus, impacting the performance of the whole entity linking pipeline.

4 EXPERIMENTAL SETUP AND RESULTS

In this section, we briefly discuss the experimental setup of the system, and the datasets used. Further, we discuss in detail the experimental results obtained for each step of the proposed framework.

Table 2: Comparative Analysis: T-NER and T-NER+.

	T-NER Performance Analysis			T-NER+ Performance Analysis		
Entity type	P	R	F1	P	R	F1
Band	0.26	0.88	0.40	0.39	0.90	0.54
Company	0.78	0.90	0.84	0.81	0.90	0.85
Facility	0.45	0.72	0.55	0.50	0.72	0.59
Geo-Location	0.80	0.95	0.87	0.80	0.95	0.87
Movie	0.24	0.88	0.38	0.34	0.88	0.49
Other	0.57	0.70	0.63	0.56	0.76	0.64
Person	0.72	0.92	0.81	0.77	0.92	0.84
Product	0.60	0.69	0.65	0.63	0.71	0.67
Sportsteam	0.52	0.83	0.64	0.63	0.85	0.72
TVshow	0.51	0.91	0.66	0.45	0.89	0.59
Overall	0.62	0.87	0.73	0.66	0.88	0.76

We use a gold-standard corpus of tweets made available by (Ritter et al., 2011) for our experiments. The dataset consists of $\approx$ 2400 tweets with 47k+ tokens. Further, for performing entity disambiguation and linking, we use DBpedia as an external KB. For performing a detailed comparative analysis, we also prepared a manually curated set of 1616 named entities identified from the given corpus of tweets, along with their entity types (in equivalence to T-NER entity types, as shown in Table 1). We have also indexed the DBpedia dataset of 'Lables' of $\approx$ 4.5M things[11] which serves as a look-up repository for candidate match retrieval for a surface form in step 2. In the following section, we report an analysis for all the steps detailed in the proposed framework in Section 3.

Step 1: Entity Recognition and Classification

We present the class-wise performance analysis of T-NER over the gold standard corpus of tweets, made available by (Ritter et al., 2011), in Table 2 where:

$$Precision(P) = \frac{|\{cor.cl\} \cap \{cl\}|}{|\{cl\}|} \quad (4)$$

$$Recall(R) = \frac{|\{cor.cl\} \cap \{cl\}|}{|\{cor.cl\}|} \quad (5)$$

$$F_1 Measure = \frac{2 \times P \times R}{P + R} \quad (6)$$

Here *cor.cl* denotes correctly classified entities, while *cl* denotes classified entities. T-NER identifies a total of 1496 named entities from microposts, in contrast to 1616 named entities present in the ground truth. 8% of entities are not even recognized and thus classified as non-entities (amongst other 44k tokens) . This is so because a large number of new entities emerge constantly on the Web as well as social media, before a KB can index them. This causes a supervised entity recognition system (such as T-NER) to fail to recognize such entities (8%).

T-NER segments tweets in the gold standard as text phrases where each text phrase is classified as an entity or non-entity. An entity is associated with an entity type with probability $P_{CRF}(c,e)$ as mentioned in equation (1), whereas a non-entity refers to the text phrase in a tweet that may not refer to any entity in the real world. We perform a detailed analysis of this entity classification aggregated at various levels, as summarized in Table 3. The classification performance reported in Table 3 depicts the percentage of text phrases that are classified as either entity or non-entity. As mentioned above, a total of 1496 text phrases are identified as entities, while the remaining 44k text phrases have been identified and classified as non-entities.

A text phrase that is an entity is correctly classified if the system recognizes it as a mention of an entity and classifies it under the same class as in the gold standard (such as *Justin Bieber* classified as *Person*). A text phrase that is an entity is incorrectly classified if the system recognizes it as a mention of an entity but classifies it under a different class as compared to gold standard (such as *Chicago* classified as *Person*, while it is of type *Geo-Location*). Text phrases that are entity mentions but have not been correctly segmented (and thus cannot be correctly recognized and classified) are denoted as segmentation errors (such as *Alpha-Omega*, identified as two distinct entities *Alpha* and *Omega*, thus classified incorrectly as *Geo-Location, Band* respectively, instead of being classified as *Movie*). A text phrase is considered as a missed entity if the text phrase refers to an entity in the real world, however, it has not been recognized and classi-

[11] http://wiki.dbpedia.org/services-resources/datasets/dataset-statistics

Table 3: T-NER Classification Performance.

Text Phrase	Classification Level	Example		Classification(%)
		Entity	Entity Type	
Entities (1496)	Correctly Classified	Justin Bieber	Person	61.57
	Incorrectly Classified	Chicago	Person	37.96
	Segmentation Error	Alpha, Omega	Geo-Location, Band	0.47
Non-Entities (44,792)	Correctly Classified	It	O	99.8
	Incorrectly Classified	justthen	Person	0.2

fied as an entity by the NER system (such as *London* classified as *O* where O is being used to denote a non-entity).

On the other hand, non-entities are either correctly classified, i.e., a text phrase that does not refer to any entity in the real world and has not been classified as an entity by the NER system (such as a text phrase *It* being classified as *O*) or incorrectly classified, i.e., a text phrase that does not refer to any entity in the real world but has been classified as an entity by the NER system (such as text phrase *justten* being classified as *Person*, instead of *O*). As evident from Table 3, 61% of entities are correctly recognized and classified whereas as much as 38% of the entities are correctly recognized but misclassified. Segmentation affects only a small subset of the entities to be recognized according to the gold standard.

Step 2: Candidate Match Retrieval and Disambiguation

Here, we present the detailed analysis of 1496 identified entities in step 1 that are linked to KB resources in step 2. In order to disambiguate these entities, we use our index of 'Labels' described above to retrieve top-k candidate resources for each identified entity. A total of 1442 entities out of 1496 entities are disambiguated with $\approx$ 4k candidate KB resources, while the rest do not produce any candidate resource on index look-up. The detailed analysis of this step is summarized in Table 4.

A linkable entity refers to a text phrase that has been identified as an entity by the NER system in step 1 and can be linked with an existing resource in a KB. Further, a correctly linked entity is one that has been identified and classified as an entity in step 1 and has been correctly linked to an existing KB resource using the approach in Section 3, step 2 (such as the identified and classified entity *Wisconsin* has been linked to a KB resource http://dbpedia.org/resource/Wisconsin of DBpedia type *Geo-Location*).

As evident from Table 4, 63% of such entities are discovered by our approach. Further, an incorrectly linked entity is one that has been identified and classified as an entity in step 1 but has been incorrectly linked to an existing KB resource which is not representative of the entity (such as the identified and classified entity *America* has been linked to a KB resource http://dbpedia.org/resource/America_America of DBpedia type *Movie*). Further, a small percentage (1.21%) of classifiable, and linkable entities (such as *Widro*) can not be linked to any resource in a KB, which can be attributed to the emergence of new entities not available in the KB (knowledge gaps). In some cases the information about the class that we extract from the linked entity is uninformative, regardless of the link being correct or not. This happens every time we establish a link to an entity typed only with the 'Thing' class, because 'Thing' is too generic. We refer to such entities as uninformative entities. For instance, the entity *N.J.* has been linked to a DBpedia resource which does not provide any information that can be used to improve the similarity scores or help in disambiguation.

Table 4: Entity Linking-Performance Analysis.

Classifiable Named Entity	Linking Level	Example		Linking(%)
		Entity	DBpedia Type	
Linkable	Correctly Linked	Wisconsin	Geo-Location	63.11
	Incorrectly Linked	America	Movie	3.05
	Uninformative	N.J.	Thing	16.15
Non-Linkable	Uninformative	Secrets	Thing	11.85
	Generic	Whitney	Other	5.83

On the other hand, a non-linkable entity refers to a text phrase that has been identified as an entity by the NER system in step 1, however, it can not be linked to any resource in DBpedia. A huge fraction of classifiable, and non-linkable entities are uninformative (such as *Secrets* has been linked to the parent DBpedia type 'Thing'). $\approx$ 6% of classifiable, non-linkable entities are too generic to be linked and can be either correctly linked or maybe incorrectly linked. For example, an entity *Steve* (a very common named entity) has been linked to a DBpedia resource of type *Person*, however, we are not certain if the entity has been linked to the correct KB resource. On the other

hand, the entity *Whitney* (another common named entity) has been linked to a DBpedia resource of type *Other* since the corresponding tweet does not provide sufficient evidence to look for the correct KB resource. Lastly, a percent of classifiable, non-linkable entities exist for which no entity link could be established (1.21%, also included in 8% missed entities of step 1).

Step 3: Entity Recognition Enhancement

Here, we present an analysis of the proposed system discussed in Section 3, step 3. By using equation (2), we re-classify the (classifiable, and linkable) named entities that have been classified and linked in step 1 and step 2, respectively, irrespective of the entity types discovered in step 1. By following this approach, we are able to improve the performance of entity recognition step of the entity linking pipeline of our system. We denote the improved entity recognition system as T-NER+.

Table 5: Example: Re-classification of entities.

Entity	Ground-Truth	T-NER	T-NER+
30stm	Band	Product	Band
Yahoo	Company	Band	Company
Southgate House	Facility	Band	Facility
Canada	Geo-Location	Person	Geo-Location
Camp rock 2	Movie	Person	Movie
Thanksgiving	Other	Person	Other
John Acuff	Person	Facility	Person
iphone	Product	Company	Product
Lions	Sportsteam	Person	Sportsteam
TMZ	TVshow	Band	TVshow

Table 2 summarizes the results of this step where we also present the comparative analysis with T-NER. As evident, we are able to improve the class-wise classification of a majority of entity types, except the entity type *TVshow* for which there is a decline in classification accuracy by almost 7%. Entity types *Geo-Location* and *Other* experience marginal decline in classification accuracy. Table 5 presents an example of re-classification of entities into correct entity types w.r.t ground truth.

5 CONCLUSIONS

In this paper, we have presented an end-to-end entity linking pipeline for short textual formats, in particular tweets. We also presented an approach to improve the entity recognition performance of a NER system by using re-classification. By our approach, we are able to enhance the classification performance of the NER system, however, the scale of this enhancement can be still improved. One outcome of our work is that newly emerging knowledge (new entities or new mentions of existing entities) on the Web, in particular social media platforms, can be extracted if not covered by an existing KB. During entity recognition and classification, we come across 8% entities that are not identified by the system. These entities comprise newly emerging entities as well as entities that have not been identified, and hence not classified. While, during entity linking, we came across $\approx 2.4\%$ entities for which a match could not be found with any resource in the DBpedia KB, owing to either non-existence of such entities in the KB or to non coverage of their surface form in the KB vocabulary.

Our next step in this field is to extract information from the Web as well as social media platforms for new entities that are discovered in the entity recognition and entity linking phase in order to, not only improve NER and NEL, but also work towards real-time lexical extensions of a KB. Concerning the future work, a possible contribution could be given by comparing the performance of the proposed approach with the most relevant related work (Yamada et al., 2015) on a common dataset, as well as using additional datasets (Rizzo et al., 2015).

REFERENCES

Cohen, W., Ravikumar, P., and Fienberg, S. (2003). A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation*, volume 3, pages 73–78.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). A framework and graphical development environment for robust nlp tools and applications. In *ACL*, pages 168–175.

Damljanovic, D. and Bontcheva, K. (2012). Named entity disambiguation using linked data. In *Proceedings of the 9th Extended Semantic Web Conference*.

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*.

Ferragina, P. and Scaiella, U. (2010). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Cre-*

ating Speech and Language Data with Amazon's Mechanical Turk, pages 80–88. Association for Computational Linguistics.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Guo, S., Chang, M.-W., and Kiciman, E. (2013). To link or not to link? a study on end-to-end tweet entity linking. In *HLT-NAACL*.

Hoffart, J., Altun, Y., and Weikum, G. (2014). Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee.

Ibrahim, Y., Amir Yosef, M., and Weikum, G. (2014). Aida-social: Entity linking on the social stream. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 17–19. ACM.

Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., and Lu, Y. (2013). Entity linking for tweets. In *ACL (1)*.

Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 359–367. Association for Computational Linguistics.

Meij, E., Weerkamp, W., and de Rijke, M. (2012). Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rizzo, G., Cano, A. E., Pereira, B., and Varga, A. (2015). Making sense of microposts (# microposts2015) named entity recognition & linking challenge. In *5th International Workshop on Making Sense of Microposts (# Microposts 15)*.

Rula, A., Palmonari, M., and Maurino, A. (2012). Capturing the age of linked open data: Towards a dataset-independent framework. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012*, pages 218–225. IEEE Computer Society.

Rula, A., Panziera, L., Palmonari, M., and Maurino, A. (2014). Capturing the currency of dbpedia descriptions and get insight into their validity. In Hartig, O., Hogan, A., and Sequeda, J., editors, *Proceedings of the 5th International Workshop on Consuming Linked Data (COLD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 20, 2014.*, volume 1264 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sutton, C. and McCallum, A. (2006). An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, pages 93–128.

Usbeck, R., Ngonga Ngomo, A.-C., Luo, W., and Wesemann, L. (2014). Multilingual disambiguation of named entities using linked data. In *International Semantic Web Conference (ISWC), Demos & Posters*.

Yamada, I., Takeda, H., and Takefuji, Y. (2015). Enhancing named entity recognition in twitter messages using entity linking. *ACL-IJCNLP 2015*, page 136.