ENCYCLOPEDIA WITH SEMANTIC COMPUTING
Vol. 1, No. 1 (2017) 1630020 (9 pages)
© The Authors

Entity Linking for Tweets

Pierpaolo Basile[†] and Annalina Caputo[‡]
[†]*Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, Bari, 70125, Italy*
[‡] *ADAPT Centre, Trinity College Dublin, Dublin, Ireland*
[†]pierpaolo.basile@uniba.it,[‡]annalina.caputo@adaptcentre.ie

Named Entity Linking (NEL) is the task of semantically annotating entity mentions in a portion of text with links to a knowledge base. The automatic annotation, which requires the recognition and disambiguation of the entity mention, usually exploits contextual clues like the context of usage and the coherence with respect to other entities. In Twitter, the limits of 140 characters originates very short and noisy text messages that pose new challenges to the entity linking task. We propose an overview of NEL methods focusing on approaches specifically developed to deal with short messages, like tweets. NEL is a fundamental task for the extraction and annotation of concepts in tweets, which is necessary for making the Twitter's huge amount of interconnected user-generated contents machine readable and enable the intelligent information access.

Keywords: Entity Linking, Twitter

1. Introduction

An average of 500 billion messages is being posted every day on Twitter making this social networking highly valuable for this huge amount of interconnected user-generated content. This information comes as unstructured short messages often characterised by noise, like misspelling, grammatical errors, jargon, implicit references to other messages, etc. In order to make such information machine readable and enable the intelligent information access, tools for the extraction and annotation of concepts in tweets are required.

Named Entity Linking (NEL) is the task of semantically annotating entity mentions in a portion of text with links to a knowledge base (e.g Wikipedia or DBpedia). This task comprises two steps. The former spots in the text all possible mentions to named entities, while the latter links each mention to the proper knowledge base. This last phase often implies the disambiguation of named entities, i.e. selecting the proper concept from a restricted set of candidates (e.g. Java <*programming language*> or Java <*place*>), since more than one concept can be referred to by the same textual form. Figure 1 shows a typical example of an ambiguous mention that can refer to different named entities. Here, from the context (*package*, *MineCraft*, *#gameDev*) it is possible to infer that the right named entity is *Java programming language*.

NEL, together with Word Sense Disambiguation, i.e. the task of associating each word occurrence with its proper meaning given a sense inventory, is critical to enable automatic systems to make sense of this unstructured text. Usually, the disambiguation of named entities is harder than the general word sense disambiguation due to their high ambiguity. Indeed, Hoffart et al. [1] report an average of 27 possible candidates per mention on CoNLL-YAGO dataset and an impressive average of 631 candidates per mention on KORE50. This is a remarkable figure if compared, for example, to 2.79, which is the average number of synsets associated to nouns in WordNet[2]. The mention context and the coherence with respect to other entity mentions play then a key role in the named entity disambiguation process since they provide useful evidence to discriminate among the many different concepts that a mention can take on.

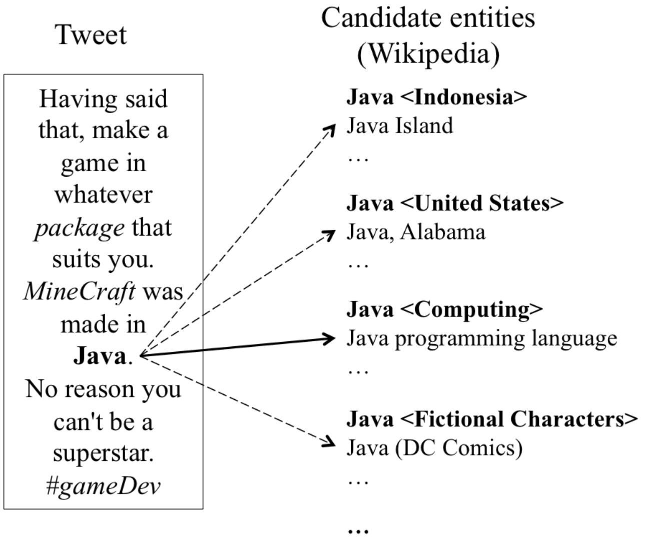
Tweet
Candidate entities (Wikipedia)
Having said that, make a game in whatever *package* that suits you. *MineCraft* was made in **Java**. No reason you can't be a superstar. *#gameDev*
Java <Indonesia>
Java Island
…
Java <United States>
Java, Alabama
…
Java <Computing>
Java programming language
…
Java <Fictional Characters>
Java (DC Comics)
…
…

Fig. 1. An example of ambiguous mention (Java) that can refer to different named entities (dashed arrows). The correct named entity for the given mention is pointed by a continuous arrow.

NEL techniques were initially developed for textual doc-

Accepted author manuscript.
Electronic version of an article published as ENCYCLOPEDIA WITH SEMANTIC COMPUTING AND ROBOTIC INTELLIGENCE, Vol. 1, No. 1 (2017) (9 pages), DOI: 10.1142/S2425038416300202 ©copyright World Scientific Publishing Company `http://www.worldscientific.com/doi/abs/10.1142/S2425038416300202`

uments, such as news articles [1; 3], where the usually lengthy and well-curated text provides enough context, in terms of surrounding words and co-occurring entities, in order to successfully disambiguate an ambiguous entity. The noise, shortness and poor language that characterise messages on microblogs like Twitter, however, severely hinder the performances of NEL techniques [4; 5].

The lack of context is one of the main factors that hampers Twitter-based NEL algorithms. Since the length of a Twitter message cannot exceed the 140 characters, these messages are often hard to disambiguate even for a human reader who does not know the context (background information on the user, tweet history, similar tweets with same mentions/hashtags, etc.) in which the tweet has been posted.

For example, in the following tweet, the lack of knowledge about the author or previously posted tweets makes impossible to assign the mention *Demi* with one of its possible six named entities listed in Wikipedia (Demi <*author*>, Demi <*singer*>, Demi Lovato, Demi Moore, Demi Orimoloye):

> *[Demi]'s been through enough and the fact that you would wish someone else's struggles on someone is sick.*

However, by expanding the analysis of just a few tweets, other mentions appear, like *@ddlovato*, *@nickjonas*, *Selena Gomez*, that can help disambiguating *Demi* as *Demi Lovato*.

Our contribution aims to provide an overview of the NEL task in Twitter and it is organized as follows: Section 2 describes the historical background, while Section 3 reports more details about the methodologies and techniques adopted to solve the problem of entity linking in tweets, with an emphasis on the techniques and methods adopted to overcome the noisy and short nature of this kind of messages. Section 4 describes the evaluation methodologies adopted for entity linking (dataset, protocols, and main outcomes) in the specific context of Twitter. Finally, Section 5 provides some scenarios of key applications and future directions.

2. Historical Background

Historically, the NEL task has been performed on regular documents, such as news, and has its roots in the Information Extraction (IE) research area. IE aims to automatically extract structured information from unstructured documents. Generally, IE involves the processing of texts by means of Natural Language Processing (NLP) techniques. One of the most relevant IE task related to NEL is the Named Entity Recognition (NER) [6; 7; 8]. NER concerns the identification of names of entities, such as *organizations*, *locations*, *peoples*. Generally, it consists of two steps: 1) the recognition of entity span in the text and 2) the identification of the type of entity. NEL task can include NER, with the addition of the linking phase. For example[a], in the following text three are the text spans that should be annotated as named entities:

> *[U.N.]$_{ORG}$ official [Ekeus]$_{PER}$ heads for [Baghdad]$_{LOC}$*

More recently, many research efforts have focused on IE for microblogs [9; 7; 8; 10] showing how extremely challenging for state-of-the-art methods is to achieve good performance in this context. Ritter et al. [8] and Liu et al. [10] report that the NER accuracy on Twitter is about 30-50%, while NER methods achieve about 85-90% on regular text; these figures point out the extent of such a challenge on IE tasks applied at Twitter. The reasons behind such a low performance are to ascribe to the shortness of text, with its implication in terms of lack of context, in addition to phenomena like the use of slang, unusual spelling, irregular capitalization, emoticons and idiosyncratic abbreviation as reported in[5]. Notwithstanding the difficulties, the upsurge of interest in this domain has its roots in the vast amount of user generated content that is published on these kinds of platforms, such as Twitter, and which allows the access to the learning and investigation of user and social behaviour studies [11].

Before delving deeply into discussing NEL methods for tweets in Section 3, we provide here background details about the NEL task. The NEL task is composed of four main stages:

(1) **Named Entity Identification.** During this step, sequences of words that could refer to a named entity are identified in the text. The beginning and the end offset of each named entity are automatically extracted from the text. The portion of text identified as entity is usually referred to as entity mention, spot or surface form. This step is very close to NER, in fact, NER systems could be exploited during this stage.
(2) **Candidate Entity Generation.** For each entity mention, a list of candidate entities in the knowledge base is retrieved. In this step, it is possible to filter some nonrelevant entities or expand entities using dictionaries, surface form expansion from the document or other methods based on external search engines.
(3) **Candidate Entity Ranking.** The list of candidate entities for each mention usually contains more than one element. This means that the entity mention is ambiguous and a method to rank candidate entities in order to find the most likely link is needed. This step is similar to what happens in Word Sense Disambiguation when a meaning is assigned to each word occurrence by selecting it from a predefined set of meanings coming from a sense inventory.
(4) **Unlinkable Mention Prediction.** In some cases, it is not possible to link the mention or there is not enough evidence for choosing the correct link. In this case, a NIL value is assigned to the entity mention.

A key component in NEL is the knowledge base (KB) where entity mentions are linked. During the past years, sev-

[a] http://www.cnts.ua.ac.be/conll2003/ner/

eral KBs have been built and many of them are related to Wikipedia. The most popular KB is certainly Wikipedia[b], which is a free multilingual encyclopedia available on-line and developed by volunteers in a collaborative way. Each article in Wikipedia describes an entity and it is referenced by a unique identifier. Moreover, Wikipedia provides additional information such as categories, redirect pages, disambiguation pages and hyperlinks between Wikipedia articles. YAGO [12] is an open-domain KB built by combining Wikipedia and WordNet [13]. YAGO combines the large number of entities in Wikipedia with the clean and clear taxonomy of concepts proposed by WordNet. Similar to YAGO, BabelNet [14], is a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms obtained by combining WordNet with several other resources such as Wikipedia, Open Multilingual WordNet, Wikidata, Wiktionary. Moreover, BabelNet is also a semantic network where concepts and named entities are connected in a very large network of semantic relations. Recently, one of the most used KB is DBpedia [15], i.e. the structured version of Wikipedia. DBpedia is a multilingual KB that contains millions of RDF statements obtained by extracting structured information form Wikipedia info boxes, templates, categories and hyperlinks. Freebase [16] is a structured KB collaboratively built by its own community. It provides an interface in which the structured data represented in the KB can be either edited by non expert users or harvested from several other sources. Freebase, which was acquired by Google in 2010, is part of the Google's Knowledge Graph. Nowadays, Freebase is no publicly available and its content has been moved to Wikidata[c].

For each step involved in NEL, we now provide a brief overview, while a wider analysis can be found in [17].

The named entity identification can be performed by using several NER tools such as Stanford NER[d], OpenNLP[e], LingPipe[f] and GATE [g]. Generally, when NER tools perform poorly (e.g. in tweets) the recognition step is jointly performed with the linking. We explain better this kind of approaches in Section 3.

The candidate entity generation is mainly performed by exploiting a dictionary. This dictionary is built by using Wikipedia or other KBs. Each entry in the dictionary corresponds to an entity mention (name) and a list of possible entities (links) is assigned to it. For example, the mention *Paris* has two possible entities: *Paris <city>* or *Paris Hilton*. When the dictionary is based on Wikipedia, it is usually built by leveraging some of its features like: to treat each page in Wikipedia as an entity and consider the page title as one of the possible mentions; to exploit redirect pages as alternative mentions that refer to the same entity; to compute the probability of an entity given a mention by using the disambiguation pages, which provide the list of entities that share the same mention; to define the list of all the possible alternative mentions for an entity by collecting the anchor text associated to the hyperlinks to a given entity page. Other techniques try to expand the surface forms in order to find new possible mentions. These approaches are based on rules or supervised methods [18], while other heuristics try to expand text in parenthesis (e.g. *Hewlett-Packard (HP)*) or the other way round (e.g *UNIBA (University of Bari)*). Finally, some approaches to the candidate generation use search engines, for example by querying a Web search engine in order to retrieve alternative mentions [19; 20]; while in [21] the Wikipedia search engine is used to build infrequently mentions.

The key component of an entity linking system is the candidate ranking module. This module ranks the list of candidate entities for each mention and selects the most appropriate entity. We can identify two main approaches: 1) **supervised methods**, which require annotated examples to learn how to rank entities by using machine learning methods, such as binary classification, learning to rank, probabilistic and graph-based approaches; 2) **unsupervised methods**, which rely on unlabelled data and generally use approaches based on Vector Space Model or other information retrieval strategies. Moreover, it is possible to categorize the entity ranking strategies:

(1) **Independent:** these approaches consider each mention independently. They do not exploit the relations between entities occurring in the same document, but they try to rank the entities by exploiting the context in which the mention occur, for example the surrounding text;
(2) **Collective:** these approaches are based on the idea that a document is focused on few related topics and then each entity is related to the other entities occurring in the same text;
(3) **Collaborative:** the idea behind these approaches is to exploit cross-document contextual information. For each entity mention these approaches try to find similar mentions occurring in similar contexts in other documents.

Several features can be taken into account in order to collect pieces of evidence useful for ranking candidate entities. Some of these features are context-independent, such as *name string comparison* between the mention and the entity name; *entity popularity*, which provides a prior probability of the candidate entity given the entity mention; *entity type*, which measures the consistency between the type of the entity mention and that of the candidate in the KB. The Entity popularity is one of the most used feature and can be easily

[b] `www.wikipedia.org`

[c] `https://www.wikidata.org`

[d] `http://nlp.stanford.edu/software/CRF-NER.shtml`

[e] `https://opennlp.apache.org/`

[f] `http://alias-i.com/lingpipe/`

[g] `https://gate.ac.uk/`

computed by exploiting the anchor text of Wikipedia links. Other approaches [22; 23] exploit Wikipedia page view statistics in order to estimate the popularity of each candidate entity or ranked entities through the Wikipedia graph structure [20].

Another kind of features is called context-dependent: in this case the context is not only the text around the mention but also the other entities in the documents. Some of these features are based on the textual content and try to measure the similarity between the description of each candidate entity and the text around the mention or the whole document. Both the description and the context can be represented using a simple bag-of-word approach or other conceptual vectors containing information like key-phrases automatically extracted from the text, categories, tags, anchor texts or other Wikipedia concepts. An important, and widely exploited, ranking feature is the *coherence* between entities that occur in the same document. Many state-of-the-art NEL approaches are based on this idea. For example, Cucerzan et al. [24] exploits the agreement between categories of two candidate entities, while Milne and Witten[25; 26] define the Wikipedia Link-based Measure (WLM), a measure based on the Google Distance [27] that computes the coherence between two entities according to the number of Wikipedia articles that link to both. This idea is used also by other authors to derive new coherence measures. For example, Ratinov et al.[28] propose a variation of the Google Distance based on the Point-wise Mutual Information, while Guo and Barbosa[22] adopt the Jaccard index. These measures work well for popular entities but they provide poor results for newly emerging entities that have few associated links. The system described in [1] tries to overcame this problem by exploiting a measure of semantic relatedness between entities which are represented as sets of weighted (multi-word) key-phrases, this similarity takes into consideration also partially overlapping phrases. Ceccarelli et al. [29] combine 27 different measures through a learning to rank algorithm. Following this approach, the method in [30] extends the set of 27 features adding further features based on word-embeddings.

The large number of different features prove the existence of several aspects that should be taken into account during the development of a NEL system. It is impossible to absolutely identify the best set of features since it depends on the applicative context and the nature of documents.

Regarding the supervised methods adopted to solve the candidate entity ranking problem, the most simple approach is based on the binary classification. Given a pair of entity mention and a candidate entity, the classifier predicts whether the entity mention refers to the candidate entity. Since the classifier can positively predict more than one entity for a single mention, a further strategy is used to select the best candidate, for example by exploiting the classifier confidence. These systems require a large pairs of annotated entities during the training phase. Other supervised approaches rely on learning to rank techniques, which are able to directly learn the rank of a list of items. In this case, for each mention the algorithm exploit the list of items represented by candidate entities.

Coherence measures can be used also to build a graph where the edges between the entities are weighted according to the coherence. For example, in [31], a *Referent Graph* is built by exploiting both the textual context similarity and the coherence between entities. A collective inference algorithm over the graph is used to infer the mapping between entities and mentions. This approach is similar to the topic-sensitive PageRank proposed in [32]. In other cases, the structure of the KB can be exploited. For example, Babelfy [33] proposes a graph-based method able to identify candidate meanings coupled with a sub-graph heuristic that selects semantic interpretations with high-coherence. The graph is built by exploiting semantic relations in Babelnet,

Recently, supervised methods based on deep learning have been developed [34; 35; 36]. Generally, these approaches do not rely on hand-crafted features but encode mention, context and entity in a continuous vector space. The method proposed in [36] tries to encode the KB structure by exploiting a deep learning approach. The idea is to map heterogeneous types of knowledge associated with an entity to numerical feature vectors in a latent space where the distance between semantically-related entities is minimized. In [37] a new measure of entities relatedness computed by a convolution network is proposed. The proposed network operates at multiple granularities to exploit different kinds of topic information.

3. Entity Linking in Tweets

In the previous section, we reported an overview of NEL methods, while in this section we describe how NEL approaches are modified or extended in order to deal with microblog texts, in particular tweets.

Two main challenges affect tweets: their noisy lexical nature and the lack of context. Tweets are very short (only 140 characters) and context-dependent features can reduce their effectiveness. The use of no regular language makes hard to identify mentions, for example the hashtag *#BarackObama* contains the valid mention *BarackObama* that should be linked to the entity *Barack Obama*. This means that a correct pre-processing of the text is needed in order to correctly identify mentions. On the other side, the social nature of Twitter can provide further sources of context, such us the user profile, tweets posted by the same authors, or other tweets in the same stream or topic that can fill the gap of the lack of context.

Some first attempts to NEL in tweets are adaptation of existing NEL tools to the context of Twitter. For example, in [38], the authors describe an adaptation of the AIDA [39] tool by improving the named entity recognition and the entity candidate lookup. While in [40] an evolution of TAGME [41] is described. The proposed approach maintains the core algorithm of TAGME, but adds functionality and several improvements in terms of pre-processing for cleaning the input text and identifying mentions. The system proposed in [42] adapts several existing tools to the context of Twitter by using supervised algorithms, such as Support Vector Machine (SVM) and Conditional Random Field (CRF) for the disambiguation; while REL-RW [43] is a RandomWalk approach based on the entity graph built from the knowledge base to compute semantic relatedness between entities and it is used in [44]. In [45] a supervised approach is boosted by the integration of a semantic search engine [46] in order to improve the mention detection. The system proposed in [47] is an adaptation of the method described in [48] that exploits several features, both context-independent and context-dependent, to which the authors added a new tokenizer and stop word removal specifically developed for tweets. Moreover, cut-off and threshold values are adapted to the context of tweets. This system achieves the best performance in the #Micropost 2016 NEEL challenge [49].

An unsupervised approach for linking is proposed in [50] where a distributional semantic model is used to compute the semantic relatedness between the entity description and the textual context in which the entity mention occurs. Conversely in [51] three different learning models that rely on different sets of features are used to perform the linking, the NIL detection and the type prediction.

An interesting approach is proposed in [52] where a step called "Candidates Filtering" is added after the entity recognition and linking. A SVM classifier is used to predict which candidates are true positives and which ones are not by relying on several features: shape features related to the mention, the entity popularity and other features related to the KBs (WordNet and DBpedia). Mentions are identified using the algorithm described in [53] specifically developed for tweets. Also, in [54] a specific algorithm [8] for tweets is used to extract mentions, while the linking step is performed by exploiting several feature in a learning to rank approach based on LambdaMART [55].

Differently form the previous approaches, the one proposed in [56] treats entity recognition and disambiguation as a single task by jointly optimizing them. The optimization is performed using a supervised approach and several features based on textual context and semantic cohesiveness between the entity-entity and entity-mention pairs. Since this approach generates overlapped linked mentions, a dynamic programming resolves these conflicts by choosing the best-scoring set of non-overlapping mention-entity mappings. This system achieves the best performance in the named entity extraction and linking challenge [57] organized within the workshop of Making Sense of Microposts (#Microposts) 2014. We will provide more details about this challenge in Section 4. Also in [58], a collective inference method able to simultaneously resolve a set of mentions is proposed. This system exploits three kinds of similarities: mention-entity similarity, entity-entity similarity, and mention-mention similarity, to enrich the context for entity linking. The system proposed in [22] focuses on the mention detection task, which the authors consider as a performance bottleneck. In this work, the authors describe a supervised algorithm based on SVM that jointly optimizes mention detection and entity disambiguation as a single end-to-end task. The learning step combines a variety of first-order, second-order, and context-sensitive features.

The supervised method proposed in [59] exploits a non-linear learning model based on trees. Non-linear models are able to capture the relationships between features, this is useful when dense features such as statistical and embedding features are used. The Structured Multiple Additive Regression Trees (S-MART) proposed by the authors is able to capture high order relationships between features by exploiting non-linear regression trees.

One of the main challenge that affects tweet is the lack of context, in [60] the authors propose a graph-based framework to collectively link all the named entity mentions in all tweets posted by modelling the user's topics of interest. The main idea behind this approach is that each user has an underlying topic interest distribution over various named entities. The method integrates the intra-tweet local information with the inter-tweet user interest information into a unified graph-based framework. Another approach proposed in [23] collects several "social signals" to improve the NEL accuracy in the context of social media such as Twitter. For example, the content of a cited URL is retrieved, the most recent tweets about any hashtags are included, and the last tweets posted by a mentioned user are extracted. The approach proposed in [61] tries to extend the context by collecting additional tweets that are similar to the target one. The authors indexed a collection of tweets that is subsequently exploited to retrieve similar tweets by using the target as a query. The system described in [62] extends the context by analysing both the user interest and the content of news. In particular, the system exploits other tweets containing the mention under analysis published by the user and the content of news that cite the mention.

An interesting approach proposed in [63] tries to include spatio-temporal information during the linking. The idea is that the prior probability of an entity of being assigned to the mention depends by spatio-temporal signals. As reported by the authors, for example, the mention "spurs" can refer to two distinct sport teams (San Antonio Spurs, which is a basketball team in the US, and Tottenham Hotspur F.C., which is a soccer team in the UK). In this case, the information about the location is crucial. The proposed method incorporates spatio-temporal signals through a weakly supervised process. In particular, the timestamps and the location (if available) are exploited as further features.

In conclusion, we can split the NEL methods for tweets in two macro categories: 1) systems that tries to improve existing NEL tools by adding specific pre-processing operations or typical features related to Twitter; 2) methods that extend the context by analysing the stream or the user profile. Some-time these two approaches can be combined to improve the performance.

4. Evaluation

Entity Linking algorithms are evaluated using typical metrics such as precision (P), recall (R) and F1-measure (F). However, the linking task can involve several steps: entity recognition, entity typing, linking and NIL instances identification. This opens the possibility to several evaluation metrics which combine some or all the involved aspects.

Following the guidelines of the #Microposts2015 NEEL challenge [64] and the Knowledge Base Population (KBP2014) Entity Linking Track[h], we can identify three measures used in the context of NEL in tweets:

(1) *strong_typed_mention_match*: it is the micro average F1 for all annotations considering the mention boundaries and their types. An annotation is correct if both its boundaries and type correspond to those in the gold standard;
(2) *strong_link_match*: it is the micro average F1 for annotations obtained by considering the correct link for each mention;
(3) *mention_ceaf*: the Constrained Entity-Alignment F-measure (CEAF) [65] is a clustering metric that evaluates clusters of annotations. It measures the F1 score for both NIL and non-NIL annotations in a set of mentions. The CEAF measure was originally proposed to evaluate co-reference resolution systems. The metric is computed by aligning gold standard annotations and system annotations with the constraint that a gold annotation is aligned with at most one system annotation. Finding the best alignment is a maximum bipartite matching problem which can be solved by the Kuhn-Munkres algorithm. More details about this metric are reported in [65].

In order to produce a unique evaluation score, Rizzo et al. [64] propose a weighted linear combination of the three aforementioned measures. Moreover, for the first time in this kind of challenge, the organizers also consider a *latency* measure. The *latency* computes the time, in seconds, required to produce a tweet annotation. However, the *latency* is used only in case of a tie and it is not included in the final score.

Results reported in several Micrposts NEEL challenges show that some remarkable results can be achieved. For example, in #Microposts2015, the winning system [51] achieves a final score of 0.8067 during the challenge, while the other participants obtain very low performance. For instance, the second system has a final score of 0.4756. It is important to underline that this system [51] is an extension of the winner system during the #Microposts2014 challenge. This proves that an end-to-end approach for both candidate selection and mention typing is effective, moreover this approach exploits two supervised learning models for NIL and type prediction.

When the type of the entity is not involved, generally the micro average F1 is computed taking into account both the entity boundaries and the link. In this case, a pair is correct only if both the entity mention and the link match the corresponding set in the gold standard. This is the case of the #Microposts2014 NEEL challenge [57] and the Entity Recognition and Disambiguation Challenge (ERD 2014) [66]. It is important to underline that the ERD challenge is not focused on tweets but the short text track is performed in the context of search engine queries.

Another resource for the evaluation of NEL in Twitter is the dataset[i] exploited by [59] and developed by Microsoft. In this case the evaluation schema is the same adopted in #Microposts2014, since entity type and NIL instances are not used.

All the previous datasets are developed for the English, however a first attempt to provide a dataset for the Italian language is reported in [67]. Moreover, a preliminary evaluation of NEL systems for the Italian is provided. The reported results prove that the task is quite difficult, as pointed out by the very low performance of all the systems employed.

In conclusion, a valuable effort to provide a standard framework for the evaluation of NEEL in tweets has been conducted during #Microposts NEEL challenges since the 2013. Additional data for training and testing can be found in the dataset developed by Microsoft. All these datasets are for English, and currently only one resource [67] is available for a language different from English, in this instance for the Italian language.

5. Key Applications and Future Directions

NEL is a fundamental task for the extraction and annotation of concepts in tweets, which is necessary for making the Twitter's user-generated content machine readable and enable the intelligent information access. Linking mentions in tweets allows to connect these very short messages to the Linked Open Data (LOD) [68] cloud thorough DBpedia or other KBs published in the LOD. This allows data from different sources, tweets included, to be connected and queried.

From an applicative perspective, microposts comprise an invaluable wealth of data, ready to be mined for training predictive models. Analysing the sentiment conveyed by microposts can yield a competitive advantage for businesses [69] and mining opinions about specific aspects of entities [70] being discussed is of paramount importance in this sense. Beyond the pure commercial application domain, the analysis of microposts can serve to gain crucial insights about political sentiment and election results [71], political movements [72], and health issues [73]. Due to the pervasiveness of mobile devices and the ubiquitous diffusion of social media platforms, the information analysis of microposts is now being exploited to

[h] http://nlp.cs.rpi.edu/kbp/2014/

[i] https://www.microsoft.com/en-us/download/details.aspx?id=52530

forecast real-world market outcomes [74]. The availability of a constant flow of information makes possible to collect real time information about several events that are being written about. The attractiveness of this is evident, and so is its potential social utility. Entity Linking is a fundamental step to add semantics and this can boost and improve any tools for the analysis of microposts.

Many challenges still remain open: 1) it is necessary to improve the pre-processing steps in order to tackle the noisy language of tweets; 2) a deep analysis of the effectiveness of different methods for extending the context is needed, in particular about the user profile and the social relations between users; 3) how to deal with emerging and popular entities mentioned in tweets that are not in the KB.

Furthermore, many few supervised approaches exploit Deep Learning techniques, which have shown to improve the performance in different NLP tasks.

Another point to take into account is the computational complexity that is crucial in the case of social media. Currently, most work on NEL lacks an analysis of computational complexity, and they usually do not evaluate the scalability and efficiency of their systems. However, for real-time and large-scale applications such as social media analysis, efficiency and scalability are significantly important and essential. Therefore, a promising direction for future research is to design methods that can improve the efficiency and scalability while aiming at, or preserving, high accuracy.

Acknowledgments

This work is supported by the project "Multilingual Entity Liking" funded by the Apulia Region under the program FutureInResearch and by the ADAPT Centre for Digital Content Technology, which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

[1] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald and G. Weikum, *Kore: keyphrase overlap relatedness for entity disambiguation*, *Proceedings of the 21st ACM international conference on Information and knowledge management*, (2012). pp. 545–554.

[2] D. Weissenborn, L. Hennig, F. Xu and H. Uszkoreit, *Multi-objective optimization for the joint disambiguation of nouns and named entities*, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Association for Computational Linguistics, Beijing, China, July 2015). pp. 596–605.

[3] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater and G. Weikum, *Robust disambiguation of named entities in text*, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (2011). pp. 782–792.

[4] E. Meij, W. Weerkamp and M. de Rijke, *Adding semantics to microblog posts*, *Proceedings of the fifth ACM international conference on Web search and data mining*, (2012). pp. 563–572.

[5] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak and K. Bontcheva, Analysis of named entity recognition and linking for tweets, *Information Processing & Management* **51**, 32 (2015).

[6] E. F. Tjong Kim Sang and F. De Meulder, *Introduction to the conll-2003 shared task: Language-independent named entity recognition*, *Proceedings of CoNLL-2003*, eds. W. Daelemans and M. Osborne (Edmonton, Canada, 2003). pp. 142–147.

[7] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau and M. Dredze, *Annotating named entities in twitter data with crowdsourcing*, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2010). pp. 80–88.

[8] A. Ritter, S. Clark, O. Etzioni *et al.*, *Named entity recognition in tweets: an experimental study*, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (2011). pp. 1524–1534.

[9] A. E. Cano, M. Rowe, M. Stankovic and A. Dadzie (eds.), *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts', Rio de Janeiro, Brazil, May 13, 2013*, CEUR Workshop Proceedings Vol. 1019, (CEUR-WS.org, 2013).

[10] X. Liu, M. Zhou, F. Wei, Z. Fu and X. Zhou, *Joint inference of named entity recognition and normalization for tweets*, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, (2012). pp. 526–535.

[11] K. Bontcheva and D. Rout, Making sense of social media streams through semantics: a survey, *Semantic Web* **5**, 373 (2014).

[12] G. K. Fabian M. Suchanek and G. Weikum, *Yago: A core of semantic knowledge unifying wordnet and wikipedia*, *16th International World Wide Web Conference, (WWW 2007)*, (2007). pp. 697–706.

[13] G. A. Miller, Wordnet: a lexical database for english, *Communications of the ACM* **38**, 39 (1995).

[14] R. Navigli and S. P. Ponzetto, BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial Intelligence* **193**, 217 (2012).

[15] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, Dbpedia: A nucleus for a web of open data, *6th International Semantic Web Conference (ISWC 2007)*, (Springer, 2007) pp. 722–735.

[16] K. Bollacker, C. Evans, P. Paritosh, T. Sturge and J. Taylor, *Freebase: a collaboratively created graph database for structuring human knowledge*, *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, (2008). pp. 1247–1250.

[17] W. Shen, J. Wang and J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, *IEEE Transactions on Knowledge and Data Engineering* **27**, 443 (2015).

[18] W. Zhang, Y. C. Sim, J. Su and C. L. Tan, *Entity linking with effective acronym expansion, instance selection, and topic modeling.*, *IJCAI*, (2011). pp. 1909–1914.

[19] X. Han and J. Zhao, *Nlpr_kbp in tac 2009 kbp track: a two-stage method to entity linking*, *Proceedings of Test Analysis Conference 2009 (TAC 09)*, (2009).

[20] M. Dredze, P. McNamee, D. Rao, A. Gerber and T. Finin, *Entity disambiguation for knowledge base population*, *Proceedings of the 23rd International Conference on Computational Linguistics*, (2010). pp. 277–285.

[21] W. Zhang, J. Su, C. L. Tan and W. T. Wang, *Entity linking leveraging: automatically generated annotation*, *Proceedings of the 23rd International Conference on Computational Linguistics*, (2010). pp. 1290–1298.

[22] S. Guo, M.-W. Chang and E. Kiciman, *To link or not to link? a study on end-to-end tweet entity linking.*, *HLT-NAACL*, (2013). pp. 1020–1030.

[23] A. Gattani, D. S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan and A. Doan, Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach, *Proceedings of the VLDB Endowment* **6**, 1126 (2013).

[24] S. Cucerzan, *Large-scale named entity disambiguation based on wikipedia data.*, *EMNLP-CoNLL*, (2007). pp. 708–716.

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

[26] D. Milne and I. H. Witten, An open-source toolkit for mining wikipedia, *Artificial Intelligence* **194**, 222 (2013).

[27] R. L. Cilibrasi and P. M. Vitanyi, The google similarity distance, *IEEE Transactions on knowledge and data engineering* **19**, 370 (2007).

[28] L. Ratinov, D. Roth, D. Downey and M. Anderson, *Local and global algorithms for disambiguation to wikipedia*, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, (2011). pp. 1375–1384.

[29] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego and S. Trani, *Learning relatedness measures for entity linking*, *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, (2013). pp. 139–148.

[30] P. Basile, A. Caputo, G. Rossiello and G. Semeraro, *Learning to rank entity relatedness through embedding-based features*, *International Conference on Applications of Natural Language to Information Systems*, (2016). pp. 471–477.

[31] X. Han, L. Sun and J. Zhao, *Collective entity linking in web text: a graph-based method*, *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, (2011). pp. 765–774.

[32] T. H. Haveliwala, *Topic-sensitive pagerank*, *Proceedings of the 11th international conference on World Wide Web*, (2002). pp. 517–526.

[33] A. Moro, A. Raganato and R. Navigli, Entity Linking meets Word Sense Disambiguation: a Unified Approach, *Transactions of the Association for Computational Linguistics (TACL)* **2**, 231 (2014).

[34] Y. Sun, L. Lin, D. Tang, N. Yang, Z. Ji and X. Wang, *Modeling mention, context and entity with neural networks for entity disambiguation*, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, (2015). pp. 1333–1339.

[35] Z. He, S. Liu, M. Li, M. Zhou, L. Zhang and H. Wang, *Learning entity representation for entity disambiguation.*, *51st Annual Meeting of the Association for Computational Linguistics*, (2013). pp. 30–34.

[36] H. Huang, L. Heck and H. Ji, Leveraging deep neural networks and knowledge graphs for entity disambiguation, *arXiv preprint arXiv:1504.07678* (2015).

[37] M. Francis-Landau, G. Durrett and D. Klein, Capturing semantic similarity for entity linking with convolutional neural networks, *arXiv preprint arXiv:1604.00734* (2016).

[38] M. A. Yosef, J. Hoffart, Y. Ibrahim, A. Boldyrev and G. Weikum, *Adapting aida for tweets*, *Making Sense of Microposts (# Microposts2014)*, (2014).

[39] M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol and G. Weikum, Aida: An online tool for accurate disambiguation of named entities in text and tables, *Proceedings of the VLDB Endowment* **4**, 1450 (2011).

[40] U. Scaiella, M. Barbera, S. Parmesan, G. Prestia, E. Del Tessandoro and M. Verı, *Datatxt at# microposts2014 challenge*, *Making Sense of Microposts (# Microposts2014)*, (2014). pp. 1–15.

[41] P. Ferragina and U. Scaiella, Fast and accurate annotation of short texts with wikipedia pages, *IEEE Software* **1**, 70 (2012).

[42] H. Barathi Ganesh, N. Abinaya, M. Anand Kumar, R. Vinayakumar and K. Soman, *Amrita-cen@ neel: Identification and linking of twitter entities*, *Making Sense of Microposts (# Microposts2015)*, (2015).

[43] Z. Guo and D. Barbosa, *Robust entity linking via random walks*, *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, (2014). pp. 499–508.

[44] Z. Guo and D. Barbosa, *Entity recognition and linking on tweets with random walks*, *Making Sense of Microposts (# Microposts2015)*, (2015).

[45] C. Gârbacea, D. Odijk, D. Graus, I. Sijaranamual and M. de Rijke, *Combining multiple signals for semanticizing tweets: University of amsterdam at# microposts2015*, *Making Sense of Microposts (# Microposts2015)*, (2015). pp. 59–60.

[46] D. Graus, D. Odijk, M. Tsagkias, W. Weerkamp and M. De Rijke, *Semanticizing search engine queries: the university of amsterdam at the erd 2014 challenge*, *Proceedings of the first international workshop on Entity recognition & disambiguation*, (2014). pp. 69–74.

[47] J. Waitelonis and H. Sack, *Named entity linking in# tweets with kea*.

[48] H. Sack, *The journey is the reward-towards new paradigms in web search*, *International Conference on Business Information Systems*, (2015). pp. 15–26.

[49] D. R. K. W. Amparo E. Cano, Daniel PreoÂÿtiuc-Pietro and A.-S. Dadzie, *6th workshop on making sense of microposts (# microposts2016)*, *Word Wide Web Conference (WWWâĂŹ16) Companion*, (ACM.

[50] P. Basile, A. Caputo, G. Semeraro and F. Narducci, *Uniba: Exploiting a distributional semantic model for disambiguating and linking entities in tweets*, *Making Sense of Microposts (# Microposts2015)*, (2015).

[51] I. Yamada, H. Takeda and Y. Takefuji, *An end-to-end entity linking approach for tweets*, *Making Sense of Microposts (# Microposts2015)*, (2015).

[52] M. B. Habib, M. Van Keulen and Z. Zhu, *Named entity extraction and linking challenge: University of twente at# microposts2014*, *Making Sense of Microposts (# Microposts2014)*, (2014).

[53] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun and B.-S. Lee, *Twiner: named entity recognition in targeted twitter stream*, *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, (2012). pp. 721–730.

[54] R. Bansal, S. Panem, P. Radhakrishnan, M. Gupta and V. Varma, *Linking entities in# microposts*, *Making Sense of Microposts (# Microposts2014)*, (2014).

[55] Q. Wu, C. J. Burges, K. M. Svore and J. Gao, Adapting boosting for information retrieval measures, *Information Retrieval* **13**, 254 (2010).

[56] M.-W. Chang, B.-J. Hsu, H. Ma, R. Loynd and K. Wang, *E2e: An*

end-to-end entity linking system for short and noisy text, *Making Sense of Microposts (# Microposts2014)*, (2014).

[57] A. E. Cano, G. Rizzo, A. Varga, M. Rowe, M. Stankovic and A.-S. Dadzie, *Making sense of microposts:(# microposts2014) named entity extraction & linking challenge*, *CEUR Workshop Proceedings*, (2014). pp. 54–60.

[58] X. Liu, Y. Li, H. Wu, M. Zhou, F. Wei and Y. Lu, *Entity linking for tweets.*, *51st Annual Meeting of the Association for Computational Linguistics*, (ACL, 2013). pp. 1304–1311.

[59] Y. Yang and M.-W. Chang, *S-mart: Novel tree-based structured learning algorithms applied to tweet entity linking*, *Proceedings of Association for Computational Linguistics*, (2015). pp. 504–513.

[60] W. Shen, J. Wang, P. Luo and M. Wang, *Linking named entities in tweets with knowledge base via user interest modeling*, *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, (2013). pp. 68–76.

[61] Y. Guo, B. Qin, T. Liu and S. Li, *Microblog entity linking by leveraging extra posts*, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (2013). pp. 863–868.

[62] S. Jeong, Y. Park, S. Kang and J. Seo, *Improved entity linking with user history and news articles*, *9th Pacific Asia Conference on Language, Information and Computation*, (2015). pp. 19–26.

[63] Y. Fang and M.-W. Chang, Entity linking on microblogs with spatial and temporal signals, *Transactions of the Association for Computational Linguistics* **2**, 259 (2014).

[64] G. Rizzo, A. C. Basave, B. Pereira, A. Varga, M. Rowe, M. Stankovic and A. Dadzie, *Making sense of microposts (# microposts2015) named entity recognition and linking (neel) challenge*, *5th Workshop on Making Sense of Microposts (# Microposts2015)*, (2015). pp. 44–53.

[65] X. Luo, *On coreference resolution performance metrics*, *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, (2005). pp. 25–32.

[66] D. Carmel, M.-W. Chang, E. Gabrilovich, B.-J. P. Hsu and K. Wang, *Erd'14: entity recognition and disambiguation challenge*, *ACM SIGIR Forum*, (2) (2014). pp. 63–77.

[67] P. Basile, A. Caputo and G. Semeraro, *Entity linking for italian tweets*, *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, eds. C. Bosco, S. Tonelli and F. M. Zanzotto (Accademia University Press, 2015). pp. 36–40.

[68] C. Bizer, T. Heath and T. Berners-Lee, Linked data-the story so far, *Semantic Services, Interoperability and Web Applications: Emerging Concepts* , 205 (2009).

[69] B. J. Jansen, M. Zhang, K. Sobel and A. Chowdury, Twitter power: Tweets as electronic word of mouth, *J. Am. Soc. Inf. Sci. Technol.* **60**, 2169 (2009).

[70] S. Batra and D. Rao, Entity based sentiment analysis on twitter, *Science* **9**, 1 (2010).

[71] A. Tumasjan, T. Sprenger, P. Sandner and I. Welpe, Predicting elections with twitter: What 140 characters reveal about political sentiment (2010).

[72] K. Starbird and L. Palen, *(how) will the revolution be retweeted?: Information diffusion and the 2011 egyptian uprising*, *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12 (ACM, New York, NY, USA, 2012). pp. 7–16.

[73] M. D. Michael J. Paul, *You are what you tweet: Analyzing twitter for public health*, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, (2011). pp. 265–272.

[74] S. Asur and B. A. Huberman, *Predicting the future with social media*, *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10 (IEEE Computer Society, Washington, DC, USA, 2010). pp. 492–499.