Word Embeddings for Unsupervised Named Entity Linking

Debora Nozza(✉), Cezar Sas, Elisabetta Fersini, and Enza Messina

University of Milano - Bicocca, Milan, Italy
{debora.nozza,elisabetta.fersini,enza.messina}@unimib.it,
c.sas@campus.unimib.it

Abstract. The huge amount of textual user-generated content on the Web has incredibly grown in the last decade, creating new relevant opportunities for different real-world applications and domains. In particular, microblogging platforms enables the collection of continuously and instantly updated information. The organization and extraction of valuable knowledge from these contents are fundamental for ensuring profitability and efficiency to companies and institutions. This paper presents an unsupervised model for the task of Named Entity Linking in microblogging environments. The aim is to link the named entity mentions in a text with their corresponding knowledge-base entries exploiting a novel heterogeneous representation space characterized by more meaningful similarity measures between words and named entities, obtained by Word Embeddings. The proposed model has been evaluated on different benchmark datasets proposed for Named Entity Linking challenges for English and Italian language. It obtains very promising performance given the highly challenging environment of user-generated content over microblogging platforms.

Keywords: Word Embeddings · Named Entity Linking · Social media

1 Introduction

With the continuous and fast evolution of the Internet and the advent of Social-Web, or Web 2.0, the amount of unstructured textual data produced by the social interactions between people has become an immense hidden treasure of knowledge [26]. Organizing and extracting valuable information from these data has become an important issue both for companies and institutions to ensure maximum profits and efficiency. In this context, the task of Information Extraction, and in particular Named Entity Linking, can provide a crucial advantage on automatically derive structured meaningful information from large collection of textual data.

Named-Entity Linking (NEL) is the task of determining the identity of entities mentioned in a textual document, that are usually extracted in the Named Entity Recognition (NER) task phase. NEL can be of great importance

C. Douligeris et al. (Eds.): KSEM 2019, LNAI 11776, pp. 115–132, 2019.
https://doi.org/10.1007/978-3-030-29563-9_13

in many fields: it can be used by search engines for disambiguating multiple-meanings entities in indexed documents or for improving queries precision, as named entities are averagely present in 70% of cases [29]. NEL systems can also be used in combination with other Natural Language Processing systems, such as Sentiment Analysis, for the generation of additional knowledge to describe users preferences towards companies, politicians, and so on.

The common NEL process typically requires annotating a potentially ambiguous entity mention with a link to global identifiers with unambiguous denotation, such as Uniform Resource Identifier (URI) in Knowledge Bases (KBs), describing the entity. A popular choice for the KB is Wikipedia, in which each page is considered as a named entity, or DBpedia, which is used as structured background knowledge in many NEL systems [2,8,10,16–18,21,25,43,44]. An example of a sentence processed for the Named Entity Linking task is shown in Fig. 1. In this example, it is possible to notice that the mention *@EmmaWatson* is correctly linked to the actress. A more difficult case regards the word *hermione*, since it can assume very different meanings, e.g. the name of an autobiographical novel, a common given name or the character of the movie Harry Potter.

Traditionally, Information Extraction studies has been successfully investigated on well-formed text, such as news or scientific publications [13,19,30,31]. Recently, a large number of studies focused on user-generated content as source data, in particular messages originated from users in micro-blogging platforms such as Twitter [2,5,12,22,33]. Due to its dynamic and informal nature, Twitter provides its users an easy way to express themselves and communicate thoughts and opinions in a highly colloquial way. This, in addition to the limitation of characters, induces the users to use abbreviations, slangs, and made-up words increasing the difficulty in recognizing and disambiguating the involved named entities. The achievement of obtaining results for social media content equally accurate to the ones on well-written text is still a long way off [9].

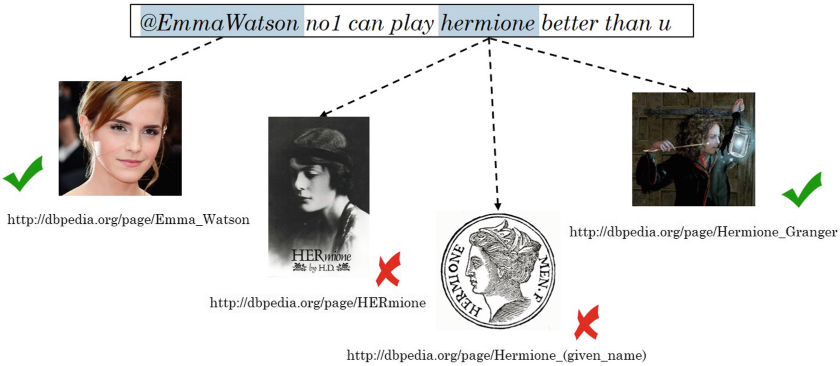
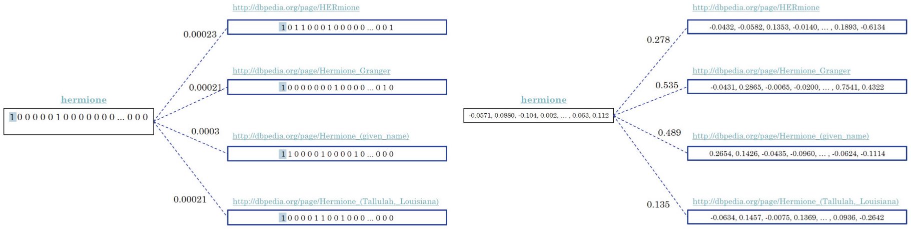

@EmmaWatson no1 can play hermione better than u
http://dbpedia.org/page/Emma_Watson
HERmione by H.D.
http://dbpedia.org/page/HERmione
HERMIONE MEN.F
http://dbpedia.org/page/Hermione_(given_name)
http://dbpedia.org/page/Hermione_Granger

Fig. 1. Example of a sentence processed by Named Entity Linking.

In order to address the issues on dealing with a micro-blogging environment, we propose a model for the exploitation of Named Entity Linking task in unsupervised settings for noisy social media texts, called *UNIMIB-WE*. The proposed model first investigates the contribution of ad-hoc preprocessing techniques for noisy text from social media, then it makes use of Word Embeddings to better deal with new emerging named entities and commonly used slangs and abbreviations. Moreover, it is expected that the use of Word Embeddings will improve the semantic similarity of the words comprising the named entities and the corresponding entries in the KB. By using the joint representation obtained with Word Embeddings models, the similarity measure will gain on semantic expressiveness resulting in a more accurate discrimination of the entities and coverage as it has been preliminary shown in [6].



(a) Bag-of-words representation.

(b) Word Embeddings representation.

Fig. 2. Example of Named Entity Linking similarity computation.

In order to demonstrate the advantage of employing Word Embeddings as word representation models, let us consider the tweet "@EmmaWatson no1 can play hermione better than u" and in particular the case of linking the entity mention "hermione". This ambiguous named entity must be disambiguated and consequently associated with several possible unambiguous entity candidates (e.g. with respect to DBpedia), comprising the correct one, i.e. the character of Hermione Granger. Figure 2 reports two possible scheme representations, one using the popular bag-of-words textual representation and the other using the more meaningful textual distributional representation of Word Embeddings. The numbers in the boxes represent the numerical vector representation associated with the text, i.e. the tweet text on the left and the textual description of the candidate KB resources on the right. The use of bag-of-words representation has been reported in Fig. 2(a), highlighting in light blue the presence (1) of the word "hermione" in each box. It is possible to note that the bag-of-words representation is very sparse, resulting in a low representative similarity measure which corresponds to 0.00021 with respect to the correct KB resource. Otherwise, the representation derived from Word Embeddings (Fig. 2(b)) permits to correctly rank as first the correct KB resource (*Hermione Granger*) with a similarity score of 0.535, as it provides a metric that better expresses the semantic properties for words and entities and consequently the similarity between them.

Following, Sect. 2 provides an overview of the state of the art approaches. Then, the proposed model for the exploitation of Word Embeddings representation is described in Sect. 3. Section 4 describes the developed framework, giving an overview of all the module components. The evaluation results on three benchmark datasets on different languages are reported in Sect. 5.

2 Related Works

The state of the art approaches in NEL can be mainly distinguished considering the specific task that they are addressing [36]:

- *Candidate Entity Generation*, which is aimed at extracting for each entity mention a set of candidates resources;
- *Candidate Entity Ranking*, focused on finding the most likely link among the candidate resources for the entity mention.
- *Unlinkable Mention Prediction*, which has the goal of predicting those mentions that cannot be associated with any resource in the KB. This step corresponds to what has been called so far as NIL prediction.

Candidate Entity Generation. The candidate generation step is a critical subprocess for the success of any NEL system. According to experiments conducted by Hachey et al. [17], a more precise candidate generator can also imply improved linking performance.

In the literature, candidate generation techniques can be mainly distinguished in Name Dictionary and Search Engine based methods. The former consists in constructing a dictionary-based structure where one or more KB resources are associated with a given named entity (dictionary key) based on some useful features available in the KB, such as redirect pages, disambiguation pages, bold phrases, etc. [10,43,44]. Given an entity mention extracted from text, the set of its candidate entities is obtained by using exact matching or partial matching with the corresponding dictionary keys [40]. An alternative solution for Candidate Entity Generation is represented by Search Engine based techniques, which make use of Web search engines for retrieving the list of candidate resources associated with an entity mention [18,21,25].

Candidate Entity Ranking. After the candidates' extraction, the list of candidates should be ranked in order to extract the most probable one. Most of the approaches are mainly based on Machine Learning algorithms for learning how to rank the candidate entities [2,8,16,19]. These approaches usually consider several features related to the named entity or the KB entry, such as *entity popularity*, the *ontology type* extracted by NER systems and vector-based representation of the context surrounding the named entity. Beyond Machine Learning models, it has also been proved that the combination of multiple features can be useful for ranking the mention candidates [5].

Unlinkable Mention Prediction. An entity mention does not always have a corresponding entity in the KB, therefore systems have to deal with the problem of predicting NIL entities (unlinkable mentions). Some approaches [8] use a simple heuristic to predict unlinkable entity mentions: if it is not possible to retrieve any candidate for an entity mention, then the entity mention is unlinkable. Many NEL systems are based on a threshold method to predict the unlinkable entity mention [4,11,14,28,37,38]. In these systems, each ranked candidate is associated to a confidence score and if the score is lower than a given threshold, then the entity mention is considered a NIL. The NIL prediction can be also accomplished using approaches based on supervised Machine Learning, such as binary classification techniques [32,44].

As stated above, the candidate generation is a crucial part for any NEL task. The process of generating the candidate resource set for a given entity mention is usually obtained by exact or partial matching between the entity mention and the labels of all the resources in the KB. However, these approaches can be error-prone, especially when dealing with microblog posts that are rich in misspellings, abbreviations, nicknames and other noisy forms of text. In order to deal with these issues, the proposed NEL approach has been defined for taking into account specific preprocessing techniques for this data and subsequently exploit a similarity measure between the high-level representation of entity mentions and KB resources. These meaningful and dense representation of entity mentions and KB resources has been obtained by taking advantage of one of the most widely used neural network language models, i.e. Word Embeddings [23].

3 Representation and Linking Model

The task of Named Entity Linking (NEL) is defined as associating an entity mention $t_i \in T$, with an appropriate KB candidate resource $k_j \in K \subset \Omega$, where $K = \{k_1, k_2, \cdots, k_{n_k}\}$ is a set of candidate resources selected from the complete set of KB resources Ω.

The main contribution consists in creating a Word Embeddings model that is able to learn a heterogeneous representation space where similarities between KB resources and named entities can be compared. In particular, given a Word Embeddings training set composed of a large but finite set of words denoting a vocabulary V and the set Ω of KB resources, the Word Embeddings model can be expressed as a mapping (or embedding) function $C : \Gamma \rightarrow \mathbb{R}^m$ with $\Gamma = V \cup \Omega$. Therefore, the embedding function is trained on a heterogeneous space of KB resources and words, ensuring that the embedded representation will be inferred from the same Word Embeddings model. More details about the training process of this heterogeneous space Word Embeddings are given in Sect. 5.

Given an entity mention t_i and a KB resource k_j, the similarity function s_C can be written as:

$$s_C(t_i, k_j) = sim(C(t_i), C(k_j)), \tag{1}$$

where *sim* is a similarity function, e.g. cosine similarity. The candidate resource set K for t_i is then obtained by taking the top-n_k KB resources ranked by the similarity score $s_C(t_i, k_j)$. The predicted KB resource k^* is then the k_j with the highest similarity score. If K is an empty set, t_i is considered as a NIL entity.

This can be generalized in the case of a multi-word entity, i.e. entities composed by two or more words in a vocabulary $w \in V$, defined as $t_i = \{w_1^i, \ldots, w_n^i\}$. Since words can be considered as point in an m-dimensional feature space, the top-n_k similar KB resources will be the set K that maximizes the sum of the similarities between k_j and all the entity mention words.

4 Experimental Settings

For performing the experiments, the *UNIMIB-WE* system proposed in Fig. 3 has been implemented, starting from the input named entities (extracted by a NER system from user-generated content) to the output (KB resources). Following, each module of the pipeline is described for a broader understanding.

4.1 Model Training

In order to obtain a Word Embeddings model able to map both words and KB resources in the same representation space, its training process has been performed over a corpus that comprises both of them. For this reason, a dump of Wikipedia has been considered as the training set. The structure of a Wikipedia article fits well the model's needs since a named entity can be directly associated with the corresponding Wikipedia article title. The following snippet reports a sentence from the Wikipedia page of Harry Potter and the Philosopher's Stone related to the character of Hermione Granger.

> "... he quickly befriends fellow first-year Ronald Weasley and Hermione Granger ..."

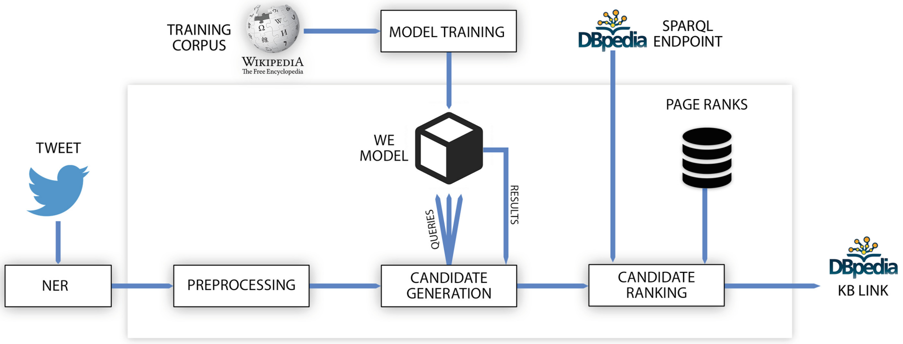


Fig. 3. Pipeline of the proposed *UNIMIB-WE* Named Entity Linking framework.

In this sentence it is possible to identify two named entities (Ronald Weasley and Hermione Granger) that, thanks to the favorable article structure, are represented as a link to their Wikipedia articles which corresponds to https://en.wikipedia.org/wiki/Ron_Weasley and https://en.wikipedia.org/wiki/Hermione_Granger respectively. The training corpus is then obtained by merging and processing all the Wikipedia articles by specifically identifying each KB resources with the tag "KB_ID/" that corresponds to the article links (e.g. "KB_ID/Hermione_Granger"). After this process, the previous sentence will result as:

"... he quickly befriends fellow first-year *KB_ID/Ron_Weasley* and *KB_ID/Hermione_Granger* ..."

Then, the Skip-Gram model [23] has been used as effective Word Embeddings model for learning the function $C : \Gamma \rightarrow \mathbb{R}^m$.

Given a sequence $s_1, \ldots, s_{n_T}$ such that $s_i \in \Gamma$, containing words and KB resources, the objective function of the Skip-gram model is defined as:

$$\mathcal{L}_{Skip-gram} = \frac{1}{n_T} \sum_{i=1}^{n_T} \sum_{-k \leq j \leq k, j \neq 0} \log P(s_{i+j} \mid s_i) \tag{2}$$

where $s_i \in \{s_1, \ldots, s_{n_T}\}$ and n_T is the sequence length.

Given the large amount of data comprised in the Wikipedia dump, the processing and the learning process of the Skip-Gram model have been performed using the efficient Wiki2Vec tool [41], a software developed using Scala, Hadoop, and Spark that processes a large amount of text and makes it usable for our requirements. It is important to mention that, given the large amount of publicly available information particularly suitable for the proposed model, Wikipedia has been used for training the Word Embeddings model. Analogously to what has been done in numerous studies and organized challenges in the state of the art [34,35], DBpedia has been used as structured background knowledge for the the Named Entity Linking process. Nevertheless, since DBpedia is a a large-scale Knowledge Base built by extracting structured data from Wikipedia [20], there is a correspondence between the entities included in these two KBs.

4.2 Preprocessing

Since the input entity mention is originated from a microblog post, it is expected to increase the number of correctly linked named entities by performing textual preprocessing because of its noisy nature. Common preprocessing involves capitalization solving and typographical error correction, such as missing spaces or wrong word separators. Moreover, for improving the retrieval performance, some query expansion techniques have been adopted, i.e. appending the "KB_ID/" token before the named entity.

4.3 Candidates Generation, Ranking and NIL Prediction

As presented in Sect. 3, the candidate generation process is performed by computing a similarity measure between the entity mention t_i and all the words or resources present in the Word Embeddings training set. As for any NNLM, the fundamental property is that semantically similar words have a similar vector representation. Given an entity mention t_i, the model returns the candidate resource set K composed of the top-n_k similar KB resources or words ranked by the similarity measure s_C, from which only the KB resources k_j are extracted. The candidate resource set can be further reduced by considering the ontology type initially inferred by a given NER model. In particular, this reduction has been performed by considering only the KB resources k_j that have the same ontology type of the named entity t_i. While the ontology type of k_j can be obtained by querying the Knowledge Base, the one of t_i can be inferred with a given NER model[1]. Finally, the candidate k^* that has the highest similarity score compared to the entity t_i is selected as the predicted KB resource.

In the proposed system, an entity mention t_i has been considered as a NIL entity, if either the similarity between t_i and the predicted resource k^* is lower than a threshold or t_i is not present in the Word Embeddings training set.

5 Experimental Results

This section discusses the datasets and the performance measures involved in the evaluation of the proposed NEL system.

5.1 Datasets

The datasets adopted for the system evaluation are three benchmark datasets of microblog posts that have been made available for different Named Entity Recognition and Linking Challenges. The #Microposts2015 and #Microposts2016 datasets have been divulgated by the Making Sense of Microposts challenge [34,35]. Moreover, a dataset of Twitter posts in the Italian language, as promoted by the NEEL-IT challenge organized by EVALITA [1,3], has been considered. In this study, all the datasets provided by the challenges (i.e. Training, Test, Dev) have been used to perform the evaluation.

In Table 1, several statistics for both English and Italian micropost challenge datasets are reported. The tables contain the total number of entities, the number of linkable entities, and the number of NIL entities.

[1] In the experimental investigation, the considered NER model is the one proposed by Ritter et al. [33], which has been specifically designed for dealing with user-generated content.

Table 1. Datasets statistics.

	#Micropost2015			#Micropost2016		
	# Entities	# Linkable Entities	# NIL Entities	# Entities	# Linkable Entities	# NIL Entities
Training	4016	3565	451	8665	6374	2291
Dev	790	428	362	338	253	85
Test	3860	2382	1478	1022	738	284

	EVALITA NEEL-IT 2016		
	# Entities	# Linkable Entities	# NIL Entities
Training	787	520	267
Test	357	226	131

5.2 Performance Measures

NEL systems are commonly evaluated using Strong Link Match (SLM) [34,35]. Given the ground truth (GT), a pair $<t_i, k_j>$ can be considered as:

- True Positive (TP): if the system correctly recognizes the link of the entity.
- False Positive (FP): if the link recognized by the system is different by the one in the GT.
- True Negative (TN): if the link is not recognized by the system and in the GT. In this case the link is NIL.
- False Negative (FN): if the system recognizes the entity, but the entity is not recognized by the GT. In other words, the system returns NIL but the GT has a link.

Using these definitions, the traditional performance measure for the SLM score, i.e. *Precision*, *Recall* and *F-measure*, can be computed. In addition, NEL systems usually measure the *NIL Score*, as the equivalent to the Recall for the NIL labeled entities.

6 Experimental Evaluation

In this section, the results achieved by the proposed approach are introduced, analyzed and presented, showing the impact of the different pipeline components on the performance measures. In particular, the system has been investigated by considering three different configurations: without preprocessing, with preprocessing and by including the ontology type into the candidate generation process. Finally, a comparison with the available state of the art approaches is discussed.

In Tables 2, 3, and 4, the results of the proposed approach without preprocessing for both English and Italian challenges are shown. As it is possible to notice, the results are promising, achieving an overall *F-measure* of 40%.

Table 2. Results for #Micropost2015 without preprocessing.

	SLM scores for #Micropost2015			
	Precision	Recall	F-measure	NIL Score
Training	0.4604	0.5186	0.4877	0.7472
Dev	0.2265	0.4182	0.2939	0.8895
Test	0.3370	0.5417	0.4168	0.8748

Table 3. Results for #Micropost2016 without preprocessing.

	SLM scores for #Micropost2016			
	Precision	Recall	F-measure	NIL Score
Training	0.3840	0.5221	0.4425	0.8520
Dev	0.3461	0.4624	0.3959	0.8235
Test	0.2563	0.3550	0.2977	0.8380

Table 4. Results for EVALITA NEEL-IT 2016 without preprocessing.

	SLM scores for EVALITA NEEL-IT 2016			
	Precision	Recall	F-measure	NIL Score
Training	0.2477	0.3750	0.2983	0.6779
Test	0.2380	0.3761	0.2915	0.5954

In order to deal with the variety of problems related to the language register of Web 2.0, the **preprocessing** step has been performed. The results are reported in Tables 5 and 6. As expected, all the performance measures have been increased of 10%–15% with respect to previous experimental settings.

An example of correctly linked entity mention after preprocessing is "*f1*": in the baseline experiment it has been labeled with the wrong link "*dbpedia.org/resource/Family_1*", while, if properly capitalized in "*F1*", the result is the correct link "*dbpedia.org/resource/Formula_One*". Another example for the Italian dataset (Table 7) regards the entity "*FEDEZ*", an Italian singer, that has been linked to a NIL entity in the baseline. By performing the capitalization resolution, the model is able to correctly link the entity to "*dbpedia.org/resource/Fedez*".

In spite of the overall performance improvements, for some entities, the preprocessing module associates erroneous links that were correctly given by the baseline method. An example is the entity "*repubblicait*", from Italian tweets, which is the account of an Italian newspaper called "La Repubblica", that after the preprocessing step is determined as NIL.

Table 5. Results for #Micropost2015 with preprocessing.

	SLM scores for #Micropost2015			
	Precision	Recall	F-measure	NIL Score
Training	0.53	0.60	0.56	0.72
Dev	0.28	0.53	0.37	0.87
Test	0.40	0.65	0.50	0.86

Table 6. Results for #Micropost2016 with preprocessing.

	SLM scores for #Micropost2016			
	Precision	Recall	F-measure	NIL Score
Training	0.45	0.61	0.52	0.84
Dev	0.53	0.71	0.61	0.78
Test	0.43	0.59	0.50	0.82

Table 7. Results for EVALITA NEEL-IT 2016 with preprocessing.

	SLM scores for EVALITA NEEL-IT 2016			
	Precision	Recall	F-measure	NIL Score
Training	0.3100	0.4692	0.3733	0.6479
Test	0.2689	0.4247	0.3293	0.6183

Another investigated experimental setting consists of considering the **ontology type** of the entity mention in the candidate generation process, taking advantage of the preprocessing step. The ontology type can be obtained by performing a type classification with a Named Entity Recognition and Classification method. It is expected that considering the ontology type of entity will help the linking process. For instance, given the entity mention "*Paris*", the corresponding inferred type *Person* will contribute to link "*Paris*" to the celebrity Paris Hilton, instead of the France capital. The achieved results are shown in Tables 8, 9, and 10.

Differently from the expectations, with the introduction of the entity types, the performance have barely improved with respect to the configuration with only preprocessing (Tables 5, 6 and 7). This behavior can be justified by the fact that the candidate resources k_j of an entity mention t_i are already mostly related to the same ontology class. Thus, do not providing any additional discriminative information, e.g. the most similar entities to a birth name will very likely be of type *Person*.

Table 8. Results for #Micropost2015 with preprocessing and considering entity types.

	SLM scores for #Micropost2015			
	Precision	Recall	F-measure	NIL Score
Training	0.5333	0.6008	0.5650	0.7184
Dev	0.2860	0.5280	0.3711	0.8729
Test	0.4015	0.6507	0.4966	0.8626

Table 9. Results for #Micropost2016 with preprocessing and considering entity types.

	SLM scores for #Micropost2016			
	Precision	Recall	F-measure	NIL Score
Training	0.4520	0.6145	0.5209	0.8358
Dev	0.5355	0.7154	0.6125	0.7764
Test	0.4015	0.6507	0.4966	0.8626

Table 10. Results for EVALITA NEEL-IT 2016 with preprocessing and considering entity types.

	SLM scores for EVALITA NEEL-IT 2016			
	Precision	Recall	F-measure	NIL Score
Training	0.3672	0.5557	0.4422	0.6479
Test	0.3165	0.5000	0.3876	0.6183

A small improvement in terms of F-measure can be observed when the candidate list is composed of resources with different ontology types. An example is the entity mention "*Interstellar*": the first match of the system based only on the preprocessing step is "*dbpedia.org/resource/Interstellar_travel*", while including the entity type *Product* gives the correct resource "*dbpedia.org/resource/Interstellar_(film)*".

6.1 State of the Art Comparison

This section presents a comparison between the proposed NEL system and the current state of the art solutions. Tables 11 and 12 report a comparison of the proposed approach with the state of the art (only those approaches providing individual results for the specific NEL task have been considered). From the results, it is possible to notice that the proposed system (**UNIMIB-WE**) has comparable performance to the top performant systems proposed at #Micropost challenges. In the #Micropost2015 challenge UNIMIB-WE places in the third position, close to the solution proposed by *Acubelab* [27] in second place. In the 2016 edition, UNIMIB-WE achieves the second place, with 60% of F-measure. The main reason why the proposed system is overcome by *KEA* [40] regards the specific optimization that this model has performed on the challenge dataset, in fact this domain-specific optimization process induced an increase of 40% in terms of F-measure compared to the not optimized version. Similarly, the *Ousia* model [42] is a supervised learning approach which exploits an ad-hoc acronym expansion dictionary.

In spite of the better-achieved results, these models have the main problem of limited generalization abilities and the need of a manually label dataset, which is very expensive in terms of human effort. Differently, the proposed NEL system does not need any supervision or labeled dataset and, given the wider range of named entities that can cover, it provides good generalization abilities to other domains.

Table 13 reports the results related to the participants of the EVALITA NEEL-IT challenge that provided the specific NEL performance. Regarding the comparison with the model proposed in [24], UNIMIB-WE obtains similar performance in terms of F-measure, but different in terms of Precision and Recall. UNIMIB-WE is less precise, but it has a higher Recall. The same performance gap occurs when comparing with the *sisinflab*'s solution [7], in this case, the higher Precision is due to the combined specific three different approaches that they used in the NEL system. They make use of DBpedia Spotlight for span and URI detection, DBpedia lookup for URI generation given a keyword, and a Word Embeddings model trained over tweets with a URI generator. Both of these solutions use an ensemble of state of the art techniques, that gives them the ability to overcome the problems of individual methods and achieve better overall performance.

Table 11. Comparison for #Micropost2015 sorted by F-measure.

#Micropost2015 Test set		
Team Name	Reference	F-measure
Ousia	[42]	0.7620
Acubelab	[27]	0.5230
UNIMIB-WE	[6]	0.5059
UNIBA	[2]	0.4640

Table 12. Comparison for #Micropost2016 sorted by F-measure.

#Micropost2016 Dev set				
Team Name	Reference	Precision	Recall	F-measure
KEA	[40]	0.6670	0.8620	0.7520
UNIMIB-WE	[6]	0.5295	0.7075	0.6057
MIT Lincoln Lab	[15]	0.7990	0.4180	0.5490
Kanopy4Tweets	[39]	0.4910	0.3240	0.3900

Table 13. Comparison for EVALITA NEEL-IT 2016.

EVALITA NEEL-IT 2016				
Team Name	Reference	Precision	Recall	F-measure
FBK-NLP (train)	[24]	0.5980	0.4540	0.5160
UNIMIB-WE (train)	[6]	0.4231	0.6403	0.5095
UNIMIB-WE (test)	[6]	0.3529	0.5575	0.4322
sisinflab (test)	[7]	0.5770	0.2800	0.3800

As a conclusion, it is possible to state that the results obtained by the proposed model are very promising, given the highly challenging environment of user-generated content over microblogging platforms. This supports the evidence of Word Embeddings as providers of semantically meaningful word representation. The model would certainly gain with the addition of a supervision procedure able to learn which module should be used with respect to the similarity score. For instance, if the similarity score between "*dbpedia.org/resource/La_Repubblica_(quotidiano)*" and "*repubblicait*" is higher than the one to "*RepubblicaIT*", the capitalization module would not be activated.

7 Conclusion

This paper introduces a Named Entity Linking system based on Word Embeddings in unsupervised settings. We addressed different issues of noisy microblogging data with an ad-hoc preprocessing that experimentally demonstrates to be an important step for this task. The introduction of Word Embeddings permits to improve the semantic similarity of the words comprising the named entities and the corresponding entries in the KB and also to better capture new emerging named entities and commonly used slangs and abbreviations. Considering the difficulties of the investigated environment, the obtained results are very promising, proving the potential of the Word Embedding model as a high-level word representation.

One of the main problems in standard Word Embeddings representation is that each word must encode all of its possible meaning into a single vector. This causes some word representation to be placed into a position that is the average of all the possible meaning of that word. Future studies could explore this issue by conveying the representation of each word occurrence considering its specific meaning.

Acknowledgements. This work has been partially supported by PON I&R 2014-20, with the grant for research project "SmartCal", CUP B48I15000180008.

References

1. Basile, P., Caputo, A., Gentile, A.L., Rizzo, G.: Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task. In: Proceedings of 3rd Italian Conference on Computational Linguistics & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, vol. 1749 (2016)
2. Basile, P., Caputo, A., Semeraro, G., Narducci, F.: UNIBA: exploiting a distributional semantic model for disambiguating and linking entities in tweets. In: Proceedings of the 5th Workshop on Making Sense of Microposts Co-located with the 24th International World Wide Web Conference, vol. 1395, p. 62 (2015)
3. Basile, P., et al. (eds.): Proceedings of 3rd Italian Conference on Computational Linguistics & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, vol. 1749 (2016)
4. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
5. Caliano, D., Fersini, E., Manchanda, P., Palmonari, M., Messina, E.: UniMiB: entity linking in tweets using Jaro-Winkler distance, popularity and coherence. In: Proceedings of the 6th Workshop on Making Sense of Microposts Co-located with the 25th International World Wide Web Conference, vol. 1691, pp. 70–72 (2016)
6. Cecchini, F.M., et al.: UNIMIB@NEEL-IT: named entity recognition and linking of italian tweets. In: Proceedings of 3rd Italian Conference on Computational Linguistics & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, vol. 1749 (2016)
7. Cozza, V., Bruna, W.L., Noia, T.D.: sisinflab: an ensemble of supervised and unsupervised strategies for the NEEL-IT challenge at Evalita 2016. In: Proceedings of 3rd Italian Conference on Computational Linguistics & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, vol. 1749 (2016)
8. Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716 (2007)
9. Derczynski, L., Maynard, D., Aswani, N., Bontcheva, K.: Microblog-genre noise and impact on semantic annotation accuracy. In: Proceedimgs of the 24th ACM Conference on Hypertext and Social Media, pp. 21–30 (2013)
10. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 277–285 (2010)

11. Ferragina, P., Scaiella, U.: TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities). In: Proceedings of the 19th ACM Conference on Information and Knowledge Management, pp. 1625–1628 (2010)
12. Fersini, E., Manchanda, P., Messina, E., Nozza, D., Palmonari, M.: Adapting named entity types to new ontologies in a microblogging environment. In: Mouhoub, M., Sadaoui, S., Ait Mohamed, O., Ali, M. (eds.) IEA/AIE 2018. LNCS (LNAI), vol. 10868, pp. 783–795. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92058-0_76
13. Fersini, E., Messina, E., Felici, G., Roth, D.: Soft-constrained inference for Named Entity Recognition. Inf. Process. Manag. **50**(5), 807–819 (2014)
14. Gottipati, S., Jiang, J.: Linking entities to a knowledge base with query expansion. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 804–813 (2011)
15. Greenfield, K., et al.: A reverse approach to named entity extraction and linking in microposts. In: Proceedings of the 6th Workshop on Making Sense of Microposts Co-located with the 25th International World Wide Web Conference, vol. 1691, pp. 67–69 (2016)
16. Guo, S., Chang, M., Kiciman, E.: To link or not to link? A study on end-to-end tweet entity linking. In: Proceedings of the 2013 Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies, pp. 1020–1030 (2013)
17. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating Entity Linking with Wikipedia. Artif. Intell. **194**, 130–150 (2013)
18. Han, X., Zhao, J.: NLPR_KBP in TAC 2009 KBP track: a two-stage method to entity linking. In: Proceedings of the 2nd Text Analysis Conference (2009)
19. Hoffart, J., et al.: Robust disambiguation of named entities in text. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 782–792 (2011)
20. Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web **6**(2), 167–195 (2015)
21. Lehmann, J., Monahan, S., Nezda, L., Jung, A., Shi, Y.: LCC approaches to knowledge base population at TAC 2010. In: Proceedings of the 3rd Text Analysis Conference (2010)
22. Manchanda, P., Fersini, E., Palmonari, M., Nozza, D., Messina, E.: Towards adaptation of named entity classification. In: Proceedings of the Symposium on Applied Computing, pp. 155–157 (2017)
23. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the 1st International Conference on Learning Representations (2013)
24. Minard, A., Qwaider, M.R.H., Magnini, B.: FBK-NLP at NEEL-IT: active learning for domain adaptation. In: Proceedings of 3rd Italian Conference on Computational Linguistics & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, vol. 1749 (2016)
25. Monahan, S., Lehmann, J., Nyberg, T., Plymale, J., Jung, A.: Cross-lingual cross-document coreference with entity linking. In: Proceedings of the Fourth Text Analysis Conference (2011)
26. Nozza, D., Ristagno, F., Palmonari, M., Fersini, E., Manchanda, P., Messina, E.: TWINE: a real-time system for TWeet analysis via INformation Extraction. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 25–28 (2017)

27. Piccinno, F., Ferragina, P.: From TagME to WAT: a new Entity Annotator. In: Proceedings of the 1st ACM International Workshop on Entity Recognition & Disambiguation, pp. 55–62 (2014)
28. Pilz, A., Paaß, G.: From names to entities using thematic context distance. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management, pp. 857–866 (2011)
29. Pound, J., Mika, P., Zaragoza, H.: Ad-hoc object retrieval in the web of data. In: Proceedings of the 19th International Conference on World Wide Web, pp. 771–780 (2010)
30. Rao, D., McNamee, P., Dredze, M.: Entity linking: finding extracted entities in a knowledge base. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing, pp. 93–115. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-28569-1_5
31. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the 13th Conference on Computational Natural Language Learning, pp. 147–155 (2009)
32. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to wikipedia. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1375–1384 (2011)
33. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: an experimental study. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534 (2011)
34. Rizzo, G., Basave, A.E.C., Pereira, B., Varga, A.: Making sense of microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) challenge. In: Proceedings of the the 5th Workshop on Making Sense of Microposts Co-located with the 24th International World Wide Web Conference, vol. 1395, pp. 44–53 (2015)
35. Rizzo, G., van Erp, M., Plu, J., Troncy, R.: Making sense of microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) challenge. In: Proceedings of the 6th Workshop on 'Making Sense of Microposts' Co-located with the 25th International World Wide Web Conference, vol. 1691, pp. 50–59 (2016)
36. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2015)
37. Shen, W., Wang, J., Luo, P., Wang, M.: LINDEN: linking named entities with knowledge base via semantic knowledge. In: Proceedings of the 21st World Wide Web Conference 2012, pp. 449–458 (2012)
38. Shen, W., Wang, J., Luo, P., Wang, M.: Linking named entities in tweets with knowledge base via user interest modeling. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 68–76 (2013)
39. Torres-Tramón, P., Hromic, H., Walsh, B., Heravi, B.R., Hayes, C.: Kanopy4Tweets: entity extraction and linking for Twitter. In: Proceedings of the 6th Workshop on 'Making Sense of Microposts' Co-located with the 25th International World Wide Web Conference, vol. 1691, pp. 64–66 (2016)
40. Waitelonis, J., Sack, H.: Named Entity Linking in #Tweets with KEA. In: Proceedings of the 6th Workshop on 'Making Sense of Microposts' Co-located with the 25th International World Wide Web Conference, vol. 1691, pp. 61–63 (2016)

41. Yamada, I., Asai, A., Shindo, H., Takeda, H., Takefuji, Y.: Wikipedia2Vec: an optimized tool for learning embeddings of words and entities from Wikipedia. CoRR abs/1812.06280 (2018)
42. Yamada, I., Takeda, H., Takefuji, Y.: An end-to-end entity linking approach for tweets. In: Proceedings of the 5th Workshop on Making Sense of Microposts Co-located with the 24th International World Wide Web Conference, vol. 1395, pp. 55–56 (2015)
43. Zhang, W., Su, J., Tan, C.L., Wang, W.: Entity linking leveraging automatically generated annotation. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1290–1298 (2010)
44. Zheng, Z., Li, F., Huang, M., Zhu, X.: Learning to link entities with knowledge base. In: Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 483–491 (2010)