Using Embeddings for Both Entity Recognition and Linking in Tweets

Giuseppe Attardi, Daniele Sartiano, Maria Simi, Irene Sucameli
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
I-56127 Pisa, Italy
{attardi, sartiano, simi}@di.unipi.it
irenesucameli@gmail.com

Abstract

English. The paper describes our submissions to the task on Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) at Evalita 2016. Our approach relies on a technique of Named Entity tagging that exploits both character-level and word-level embeddings. Character-based embeddings allow learning the idiosyncrasies of the language used in tweets. Using a full-blown Named Entity tagger allows recognizing a wider range of entities than those well known by their presence in a Knowledge Base or gazetteer. Our submissions achieved first, second and fourth top official scores.

Italiano. *L'articolo descrive la nostra partecipazione al task di Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) a Evalita 2016. Il nostro approccio si basa sull'utilizzo di un Named Entity tagger che sfrutta embeddings sia character-level che word-level. I primi consentono di apprendere le idiosincrasie della scrittura nei tweet. L'uso di un tagger completo consente di riconoscere uno spettro più ampio di entità rispetto a quelle conosciute per la loro presenza in Knowledge Base o gazetteer. Le prove sottomesse hanno ottenuto il primo, secondo e quarto dei punteggi ufficiali.*

1 Introduction

Most approaches to entity linking in the current literature split the task into two equally important but distinct problems: *mention detection* is the task of extracting surface form candidates that correspond to entities in the domain of interest; *entity disambiguation* is the task of linking an extracted mention to a specific instance of an entity in a knowledge base.

Most approaches to mention detection rely on some sort of fuzzy matching between n-grams in the source text and a list of known entities (Rizzo et al., 2015). These solutions suffer severe limitations when dealing with Twitter posts, since the posts' vocabulary is quite varied, the writing is irregular, with variants and misspellings and entities are often not present in official resources like DBpedia or Wikipedia.

Detecting the correct entity mention is however crucial: Ritter et al. (2011) for example report a 0.67 F1 score on named entity segmentation, but an 85% accuracy, once the correct entity mention is detected, just by a trivial disambiguation that maps to the most popular entity.

We explored an innovative approach to mention detection, which relies on a technique of Named Entity tagging that exploits both character-level and word-level embeddings. Character-level embeddings allow learning the idiosyncrasies of the language used in tweets. Using a full-blown Named Entity tagger allows recognizing a wider range of entities than those well known by their presence in a Knowledge Base or gazetteer.

Another advantage of the approach is that no pre-built resource is required in order to perform the task, minimal preprocessing is required on the input text and no manual feature extraction nor feature engineering is required.

We exploit embeddings also for disambiguation and entity linking, proposing the first approach that, to the best of our knowledge, uses only embeddings for both entity recognition and linking.

We report the results of our experiments with this approach on the task Evalita 2016 NEEL-IT. Our submissions achieved first, second and fourth top official scores.

2 Task Description

The NEEL-IT task consists of annotating named entity mentions in tweets and disambiguating them by linking them to their corresponding entry in a knowledge base (DBpedia).

According to the task Annotation Guidelines (NEEL-IT Guidelines, 2016), a mention is a string in the tweet representing a proper noun or an acronym that represents an entity belonging to one of seven given categories (Thing, Event, Character, Location, Organization, Person and Product). Concepts that belong to one of the categories but miss from DBpedia are to be tagged as NIL. Moreover "The extent of an entity is the entire string representing the name, excluding the preceding definite article".

The Knowledge Base onto which to link entities is the Italian DBpedia 2015-10, however the concepts must be annotated with the canonicalized dataset of DBpedia 2015, which is an English one. Therefore, despite the tweets are in Italian, for unexplained reasons the links must refer to English entities.

3 Building a larger resource

The training set provided by the organizers consists of just 1629 tweets, which are insufficient for properly training a NER on the 7 given categories.

We thus decided to exploit also the training set of the Evalita 2016 PoSTWITA task, which consists of 6439 Italian tweets tokenized and gold annotated with PoS tags. This allowed us to concentrate on proper nouns and well defined entity boundaries in the manual annotation process of named entities.

We used the combination of these two sets to train a first version of the NER.

We then performed a sort of active learning step, applying the trained NER tagger to a set of over 10 thousands tweets and manually correcting 7100 of these by a team of two annotators.

These tweets were then added to the training set of the task and to the PoSTWITA annotated training set, obtaining our final training corpus of 13,945 tweets.

4 Description of the system

Our approach to Named Entity Extraction and Linking consists of the following steps:

- Train word embeddings on a large corpus of Italian tweets
- Train a bidirectional LSTM character-level Named Entity tagger, using the pre-trained word embeddings
- Build a dictionary mapping titles of the Italian DBpedia to pairs consisting of the corresponding title in the English DBpedia 2011 release and its NEEL-IT category. This helps translating the Italian titles into the requested English titles. An example of the entries in this dictionary are:

  Cristoforo_Colombo (http://dbpedia.org/resource/Christopher_Columbus, Person)
  Milano (http://dbpedia.org/resource/Milan, Location)

- From all the anchor texts from articles of the Italian Wikipedia, select those that link to a page that is present in the above dictionary. For example, this dictionary contains:

  Person Cristoforo_Colombo Colombo

- Create word embeddings from the Italian Wikipedia
- For each page whose title is present in the above dictionary, we extract its abstract and compute the average of the word embeddings of its tokens and store it into a table that associates it to the URL of the same dictionary
- Perform Named Entity tagging on the test set
- For each extracted entity, compute the average of the word embeddings for a context of words of size *c* before and after the entity.
- Annotate the mention with the DBpedia entity whose lc2 distance is smallest among those of the abstracts computed before.
- For the Twitter mentions, invoke the Twitter API to obtain the real name from the screen name, and set the category to Person if the real name is present in a gazetteer of names.

The last step is somewhat in contrast with the task guidelines (NEEL-IT Guidelines, 2016), which only contemplate annotating a Twitter mention as `Person` if it is recognizable on the spot as a known person name. More precisely "If the mention contains the name and surname of a person, the name of a place, or an event, etc., it

should be considered as a named entity", but "Should not be considered as named entity those aliases not universally recognizable or traceable back to a named entity, but should be tagged as entity those mentions that contains well known aliases. Then, `@ValeYellow46` should not be tagged as is not an alias for `Valentino Rossi`".

We decided instead that it would have been more useful, for practical uses of the tool, to produce a more general tagger, capable of detecting mentions recognizable not only from syntactic features. Since this has affected our final score, we will present a comparison with results obtained by skipping this last step.

4.1 Word Embeddings

The word embeddings for tweets have been created using the `fastText` utility[1] by Bojanowski et al. (2016) on a collection of 141 million Italian tweets retrieved over the period from May to September 2016 using the Twitter API. Selection of Italian tweets was achieved by using a query containing a list of the 200 most common Italian words.

The text of tweets was split into sentences and tokenized using the sentence splitter and the tweet tokenizer from the linguistic pipeline Tanl (Attardi et al., 2010), replacing emoticons and emojis with a symbolic name starting with EMO_ and normalizing URLs. This preprocessing was performed by MapReduce on the large source corpora.

We produced two versions of the embeddings, one with dimension 100 and a second of dimension 200. Both used a window of 5 and retained words with a minimum count of 100, for a total of 245 thousands words.

Word embeddings for the Italian Wikipedia were created from text extracted from the Wikipedia dump of August 2016, using the `WikiExtractor` utility by Attardi (2009). The vectors were produced by means of the `word2vec` utility[2], using the skipgram model, a dimension of 100, a window of 5, and a minimum occurrence of 50, retaining a total of 214,000 words.

4.2 Bi-LSTM Character-level NER

Lample et al. (2016) propose a Named Entity Recognizer that obtains state-of-the-art performance in NER on the 4 CoNLL 2003 datasets without resorting to any language-specific knowledge or resources such as gazetteers.

In order to take into account the fact that named entities often consist of multiple tokens, the algorithms exploits a bidirectional LSTM with a sequential conditional random layer above it.

Character-level features are learned while training, instead of hand-engineering prefix and suffix information about words. Learning character-level embeddings has the advantage of learning representations specific to the task and domain at hand. They have been found useful for morphologically rich languages and to handle the out-of-vocabulary problem for tasks like POS tagging and language modeling (Ling et al., 2015) or dependency parsing (Ballesteros et al., 2015).

The character-level embedding are given to bidirectional LSTMs and then concatenated with the embedding of the whole word to obtain the final word representation as described in Figure 1:

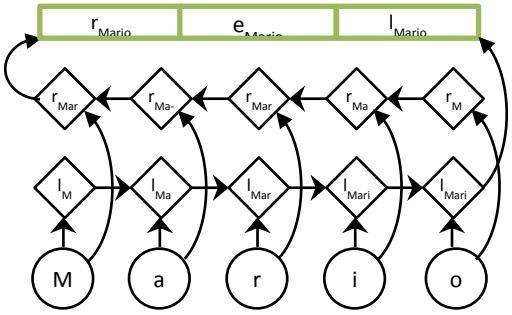

Figure 1. The embeddings for the word "Mario" are obtained by concatenating the two bidirectional LSTM character-level embeddings with the whole word embeddings.

The architecture of the NER tagger is described in Figure 2.

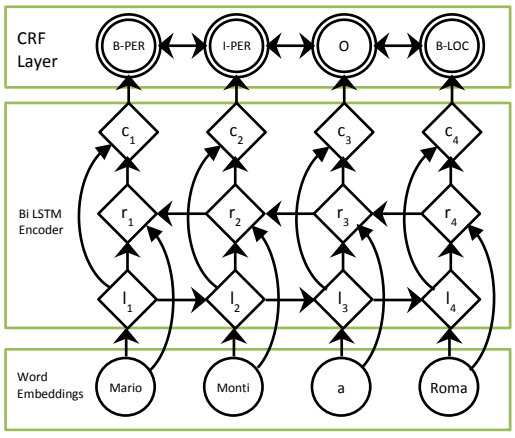

Figure 2. Architecture of the NER

[1] https://github.com/facebookresearch/fastText.git

[2] https://storage.googleapis.com/google-code-archive-source/v2/code.google.com/word2vec/source-archive.zip

5 Experiments

Since Named Entity tagging is the first step of our technique and hence its accuracy affects the overall results, we present separately the evaluation of the NER tagger.

Here are the results of the NER tagger on a development set of 1523 tweets, randomly extracted from the full training set.

Category	Precision	Recall	F1
Character	50.00	16.67	25.00
Event	92.48	87.45	89.89
Location	77.51	75.00	76.24
Organization	88.30	78.13	82.91
Person	73.71	88.26	88.33
Product	65.48	60.77	63.04
Thing	50.00	36.84	42.42

Table 1. NER accuracy on devel set.

On the subset of the test set used in the evaluation, which consists of 301 tweets, the NER performs as follows:

Category	Precision	Recall	F1
Character	0.00	0.00	0.00
Event	0.00	0.00	0.00
Location	72.73	61.54	66.67
Organization	63.46	44.59	52.38
Person	76.07	67.42	71.49
Product	32.26	27.78	29.85
Thing	0.00	0.00	0.00

Table 2. NER accuracy on gold test set.

In the disambiguation and linking process, we experimented with several values of the context size *c* of words around the mentions (4, 8 and 10) and eventually settled for a value of 8 in the submitted runs.

6 Results

We submitted three runs. The three runs have in common the following parameters for training the NER:

Character embeddings dimension	25
dropout	0.5
Learning rate	0.001
Training set size	12,188

Table 3. Training parameters for NER.

Specific parameters of the individual runs are:

- UniPI.1: twitter embeddings with dimension 100, disambiguation by frequency of mention in Wikipedia anchors
- UniPI.2: twitter embeddings with dimension 100, disambiguation with Wikipedia embeddings
- UniPI.3: twitter embeddings with dimension 200, disambiguation with Wikipedia embeddings, training set with geographical entities more properly annotated as Location (e.g. Italy).

The runs achieved the scores listed in the following table:

Run	Mention ceaf	Strong typed mention match	Strong link match	Final score
UniPI.3	0.561	0.474	0.456	**0.5034**
UniPI.1	0.561	0.466	0.443	0.4971
Team2.base	0.530	0.472	**0.477**	0.4967
UniPI.2	0.561	0.463	0.443	0.4962
Team3.3	**0.585**	**0.516**	0.348	0.4932

Table 4. Top official task results.

The final score is computed as follows:

$$0.4\ mention\_ceaf + 0.3\ strong\_typed\_mention\_match + 0.3\ strong\_link\_match$$

As mentioned, our tagger performs an extra effort in trying to determine whether Twitter mentions represent indeed Person or Organization entities. In order to check how this influences our result we evaluate also a version of the UniPI.3 run without the extra step of mention type identification. The results are reported in the following table:

Run	Mention ceaf	Strong typed mention match	Strong link match	Final score
UniPI.3 without mention check	**0.616**	0.531	0.451	**0.541**

Table 5. Results of run without Twitter mention analysis.

On the other hand, if we manually correct the test set annotating the Twitter mentions that indeed refer to Twitter users or organizations, the score for strong typed mentions match increases to 0.645.

7 Discussion

The effectiveness of the use of embeddings in disambiguation can be seen in the improvement in the strong link match score between run UniPI.2 and UniPI.3. Examples where embeddings lead to better disambiguation are:

Liverpool_F.C. vs Liverpool
Italy_national_football_team vs Italy
S.S._Lazio vs Lazio
Diego_Della_Valle vs Pietro_Della_Valle
Nobel_Prize vs Alfred_Nobel

There are many cases where the NER recognizes a Person, but the linker associates the name to a famous character, for example:

Maria_II_of_Portugal for Maria
Luke_the_Evangelist for Luca

The approach of using embeddings for disambiguation looks promising: the abstract of articles sometimes does not provide appropriate evidence, since the style of Wikipedia involves providing typically meta-level information, such as the category of the concept. For example disambiguation for "Coppa Italia" leads to "Italian_Basketball_Cup" rather than to "Italian_Football_Cup", since both are described as sport competitions. Selecting or collecting phrases that mention the concept, rather than define it, might lead to improved accuracy.

Using character-based embeddings and a large training corpus requires significant computational resources. We exploited a server equipped with nVidia Tesla K 80 GPU accelerators.

Nevertheless training the LSTM NER tagger still required about 19 hours: without the GPU accelerator the training would have been impossible.

8 Related Work

Several papers discuss approaches to end-to-end entity linking (Cucerzan, 2007; Milne and Witten, 2008; Kulkarni et al., 2009; Ferragina and Scaiella, 2010; Han and Sun, 2011; Meij et al., 2012), but many heavily depend on Wikipedia text and might not work well in short and noisy tweets.

Most approaches to mention detection rely on some sort of fuzzy matching between n-grams in the source and the list of known entities (Rizzo et al., 2015).

Yamada et al. (2015) propose an end-to-end approach to entity linking that exploits word embeddings as features in a random-forest algorithm used for assigning a score to mention candidates, which are however identified by either exact or fuzzy matching on a mention-entity dictionary built from Wikipedia titles and anchor texts.

Guo et al. (2016) propose a structural SVM algorithm for entity linking that jointly optimizes mention detection and entity disambiguation as a single end-to-end task.

9 Conclusions

We presented an innovative approach to mention extraction and entity linking of tweets, that relies on Deep Learning techniques. In particular we use a Named Entity tagger for mention detection that exploits character-level embeddings in order to deal with the noise in the writing of tweet posts. We also exploit word embeddings as a measure of semantic relatedness for the task of entity linking.

As a side product we produced a new gold resource of 13,609 tweets (242,453 tokens) annotated with NE categories, leveraging on the resource distributed for the Evalita PoSTWITA task.

The approach achieved top score in the Evalita 2016 NEEL-IT Challenge and looks promising for further future enhancements.

Acknowledgments

We gratefully acknowledge the support by the University of Pisa through project PRA 2016 and by NVIDIA Corporation through the donation of a Tesla K40 GPU accelerator used in the experiments.

We thank Guillaume Lample for making available his implementation of Named Entity Recognizer.

References

Giuseppe Attardi. 2009. WikiExtractor: A tool for extracting plain text from Wikipedia dumps. https://github.com/attardi/wikiextractor

Giuseppe Attardi, Stefano Dei Rossi, and Maria Simi. 2010. The Tanl Pipeline. In *Proc. of LREC Workshop on WSPP*, Malta.

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based dependency parsing by modeling characters instead of words with LSTMs. In *Proceedings of EMNLP 2015*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vec-

tors with Subword Information. https://arxiv.org/abs/1607.04606

Silviu Cucerzan. 2007. Large-scale named entity disam- biguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference of EMNLP-CoNLL*, pages 708–716.

Evalita. 2016. NEEL-IT. http://neel-it.github.io

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia enti- ties). In *Proceedings of the 19th ACM international conference on Information and knowledge manage- ment*, CIKM '10, pages 1625–1628, New York, NY, USA. ACM.

Stephen Guo, Ming-Wei Chang and Emre Kıcıman. 2016. To Link or Not to Link? A Study on End-to-End Tweet Entity Linking.

Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Asso- ciation for Computational Linguistics: Human Lan- guage Technologies - Volume 1*, HLT '11, pages 945– 954, Stroudsburg, PA, USA. Association for Compu- tational Linguistics.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD), pages 457–466, New York, NY, USA. ACM.

Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition, In *Proceedings of NAACL-HLT (NAACL 2016).*

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 563–572, New York, NY, USA. ACM.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 509–518, New York, NY, USA. ACM.

NEEL-IT Annotation Guidelines. 2016. https://drive.google.com/open?id=1saUb2NSxml67pcrz3m_bMcibe1nST2CTedeKOBklKaI

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. arXiv:1310.4546

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Giuseppe Rizzo, Amparo Elizabeth Cano Basave, Bianca Pereira, and Andrea Varga. 2015. Making sense of microposts (#microposts2015) named entity recognition and linking (NEEL) challenge. In *Proceedings of the 5th Workshop on Making Sense of Microposts*, pages 44–53.

Ikuya Yamada, Hideaki Takeda, and Yoshiyasu TakefujiIkuya. 2015. An End-to-End Entity Linking Approach for Tweets. In *Proceedings of the 5th Workshop on Making Sense of Microposts*, Firenze, Italy.