UNIMIB@NEEL-IT : Named Entity Recognition and Linking of Italian Tweets

Flavio Massimiliano Cecchini, Elisabetta Fersini, Pikakshi Manchanda, Enza Messina, Debora Nozza, Matteo Palmonari, Cezar Sas
Department of Informatics, Systems and Communication (DISCo)
University of Milano-Bicocca, Milan, Italy
{flavio.cecchini, fersini, pikakshi.manchanda, messina, debora.nozza, palmonari}@disco.unimib.it
c.sas@campus.unimib.it

Abstract

English. This paper describes the framework proposed by the UNIMIB Team for the task of Named Entity Recognition and Linking of Italian tweets (NEEL-IT). The proposed pipeline, which represents an entry level system, is composed of three main steps: (1) Named Entity Recognition using Conditional Random Fields, (2) Named Entity Linking by considering both Supervised and Neural-Network Language models, and (3) NIL clustering by using a graph-based approach.

Italiano.

Questo articolo descrive il sistema proposto dal gruppo UNIMIB per il task di Named Entity Recognition and Linking applicato a tweet in lingua italiana (NEEL-IT). Il sistema, che rappresenta un approccio iniziale al problema, è costituito da tre passaggi fondamentali: (1) Named Entity Recognition tramite l'utilizzo di Conditional Random Fields, (2) Named Entity Linking considerando sia approcci supervisionati sia modelli di linguaggio basati su reti neurali, e (3) NIL clustering tramite un approccio basato su grafi.

1 Introduction

Named Entity Recognition (NER) and Linking (NEL) have gained significant attention over the last years. While dealing with short textual formats, researchers face difficulties in such tasks due to the increasing use of informal, concise and idiosyncratic language expressions (Derczynski et al., 2015). In this paper, we introduce a system that tackles the aforementioned issues for **Italian language** tweets. A detailed description of these tasks is provided in the next sections.

2 Systems Description

The proposed system (Figure 1) comprises of three stages: Named Entity Recognition, Named Entity Linking and NIL Clustering. In this section,

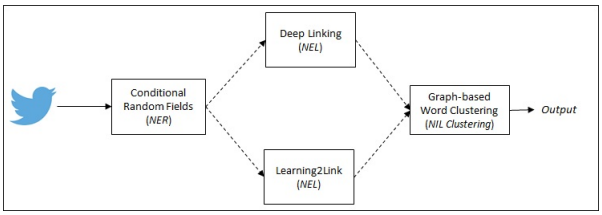

Figure 1: UNIMIB system. Dotted paths are related to optional paths.

we provide a detailed explanation of the different methods used to address these tasks.

2.1 Named Entity Recognition

In order to identify named entities from microblog text, we used Conditional Random Fields (CRF), i.e. a probabilistic undirected graphical model that defines the joint distribution $P(y|x)$ of the predicted labels (hidden states) $y = y_1, ..., y_n$ given the corresponding tokens (observations) $x = x_1, ..., x_n$. The probability of a sequence of label y given the sequence of observations x can be rewritten as:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{N} \sum_{k=1}^{K} \omega_k f_k(y_t, y_{t-1} x, t)\right) \quad (1)$$

where $f_k(y_t, y_{t-1} x, t)$ is an arbitrary feature function over its arguments and ω_k is a feature weight that is a free parameter in the model. Feature

functions are fixed in advance and are used to verify some properties of the input text, while the weights ω_k have to be learned from data and are used to tune the discriminative power of each feature function. In our runs, two configurations of CRF have been trained using the training data available for the challenge: (1) CRF and (2) CRF+Gazetteers. In particular, in the last configuration the model has been induced enclosing several gazetteers, i.e. products, organizations, persons, events and characters. The output of CRF is a set candidate entities $e_1, e_2, ..., e_m$ in each given tweet t.

2.2 Named Entity Linking

The task of Named Entity Linking (NEL) is defined as associating an entity mention e_j (identified from a tweet t) to an appropriate KB candidate resource c_j^i from a set $C_j = \{c_j^1, c_j^2, \cdots, c_j^k\}$ of candidate resources. We explored two different linking approaches: Learning2Link and Neural-Network Language Model (NNLM) Linking.

2.2.1 Learning2Link

For this phase, we used the Italian version of DBpedia as our KB. To this end, we extract Titles of all Wikipedia articles (i.e., the *labels* dataset) from Italian DBpedia and index them using LuceneAPI. For each entity mention e_j, we retrieve a list of top-k ($k = 10$) candidate resources from the KB. We compute the scores as described below (Caliano et al., 2016), which are used to create the input space for the Learning2Link (L2L) phase for each candidate resource for an entity mention:

- $lcs(e_j, l_{c_j^i})$ which denotes a normalized Lucene Conceptual Score between an entity e_j and the label of a candidate resource $l_{c_j^i}$;
- $cos(e_j^*, a_{c_j^i}))$ which represents a discounted cosine similarity between an entity context e_j^* (modeled as a vector composed of an identified entity e_j and non stop-words in a tweet t) and a candidate KB abstract description $a_{c_j^i}$;
- *Jaro-Winkler distance* (Jaro, 1995) between an entity e_j and the label of a resource $l_{c_j^i}$;
- $R(c_j^i)$ which is a popularity measure of a given candidate resource c_j^i in the KB.

This input space is used for training various learning algorithms such as *Decision Trees (DT), Multi-Layer Perceptron (MLP), Support Vector Machines (SVM)* with Linear-, Polynomial- and Radial-kernels, *Bayesian Networks (BN), Voted Perceptron (VP), Logistic Regression (LR)* and *Naïve Bayes* (NB). The target class is a boolean variable which indicates whether or not a candidate resource URI is a suitable link in the KB for the entity mention e_j. An important point to note here is that the models are learning by similarity, i.e., they learn the target class for a candidate resource by using the afore-mentioned similarity scores.

A **Decision Criteria** is further created based on the target class so as to predict the most suitable candidate resource URI from amongst a list of URIs of candidate resources $\{c_j^1, c_j^2, \cdots, c_j^k\}$ of an entity mention e_j (or detect the NIL mentions) in the test set. This criteria is described as follows:

if candidate resource c_j^i is predicted to be a suitable match for e_j **then**
  Map the entity mention e_j to the candidate resource c_j^i
else if more than one candidate resources have been predicted to be suitable matches for e_j **then**
  Map the entity mention e_j to the candidate resource c_j^i with the highest probability score
else if no candidate resource is predicted as a suitable match by the algorithm **then**
  Map the entity mention e_j to a NIL mention
end if

Finally, the entity type of a mention is determined by the DBpedia type of the selected candidate resource, which is finally mapped to a type in the Evalita Ontology based on an Ontology mapping that we developed between the Evalita Ontology and the DBpedia Ontology, as per the guidelines of the Challenge. In case, a mention has been mapped to a NIL mention, the entity type is determined by the CRF type obtained in the entity recognition phase.

2.2.2 Neural-Network Language Model (NNLM) Linking

The process of generating the candidate resource set C_j for the entity mention e_j is a crucial part for the NEL task. To obtain C_j, most of the state-of-the-art approaches (Dredze et al., 2010; McNamee, 2010) make use of exact or partial matching (e.g. Hamming distance, character Dice score, etc.) between the entity mention e_j and the labels

of all the resources in the KB. However, these approaches can be error-prone, especially when dealing with microblog posts rich of misspellings, abbreviations, nicknames and other noisy forms of text.

The idea behind the proposed NNLM Linking approach is to exploit a high-level similarity measure between the entity mentions e_j and the KB resources, in order to deal with the afore-mentioned issues. Instead of focusing on the similarity measure definition, we focus on the word representation. The need of a meaningful and dense representation of words, where words and entities are represented in a different way, and an efficient algorithm to compute this representation, lead us to the most used Neural-Network Language model, i.e. Word Embeddings (Mikolov et al., 2013).

A Word Embedding, $WE : words \rightarrow \mathbb{R}^n$, is a function which maps words in some language to high-dimensional vectors. Embeddings have been trained on the Italian Wikipedia and they have been generated for all the words in the Wikipedia texts, adding a specific tag if the words corresponded to a KB entry, i.e. a Wikipedia article.

Given an entity e_j and a word w belonging to the word's dictionary D of the Wikipedia text, we can define the similarity function s as:

$$s(e_j, w) = sim(\, WE(e_j),\, WE(w)), \qquad (2)$$

where sim is the cosine similarity.

Given an entity e_j, the candidate resource set C_j is created by taking the top-k words w for the similarity score $s(e_j, w)$. Then, the predicted resource c^* is related to the word with the highest similarity score such that the word corresponds to a KB entry and its type is coherent with the type resulting from the NER system. If C_j does not contain words correspondent to a KB entry, e_j is considered as a NIL entity.

2.3 NIL Clustering

We tackled the subtask of NIL clustering with a graph-based approach. We build a weighted, undirected co-occurrence graph where an edge represents the co-occurrence of two terms in a tweet. Edge weights are the frequencies of such co-occurrences. We did not use measures such as log likelihood ratio or mutual information, as frequencies might be too low to yield significant scores. In the word graph we just retained lemmatized nouns, verbs, adjectives and proper nouns, along with abbreviations and foreign words. More precisely, we used TreeTagger (Schmid, 1994) with Achim Stein's parameters for Italian part-of-speech tagging, keeping only tokens tagged as `VER`, `NOM`, `NPR`, `ADJ`, `ABR`, `FW` and `LS`. We made the tagger treat multi-word named entities (be they linked or NIL) as single tokens. The ensuing word graph was then clustered using the MaxMax algorithm (Hope and Keller, 2013) to separate it into rough topical clusters. We notice that tweets with no words in common always lie in different connected components of the word graph and thus in different clusters.

Subsequently, we reduced the clusters considering only tokens that were classified as NILs. Within each cluster, we measure the string overlap between each pair of NIL tokens s_1, s_2, assigning it a score in $[0, 1]$. We computed the length λ of the longest prefix[1] of the shorter string that is also contained in the longer string and assigned it the score $\frac{\lambda^2}{|s_1| \cdot |s_2|}$. Similar overlaps of two or less letters, i.e. when $\lambda \leq 2$, are not considered meaningful, so they automatically receive a score of 0; on the contrary, when two meaningfully long strings coincide, i.e. $\lambda = |s_1| = |s_2|$ and $|s_1| > 2$, the pair will receive a score of 1.

A token is considered to possibly represent the same entity as another token if 1) their named entity type is the same and 2a) their overlap score is greater than an experimentally determined threshold) or 2b) they co-occur in any tweet and their overlap score is greater than 0. For each token s, we consider the set of other tokens that satisfy 1) and 2a) or 2b) for s. However, this still does not define an equivalence relation, so that we have to perform intersection and union operations on these sets to obtain the final partition of the NIL tokens. Finally, each NIL named entity will be labelled according to its cluster.

3 Results and Discussion

We first evaluate our approach on the training set consisting of 1000 tweets made available by the EVALITA 2016 NEEL-IT challenge. The results have been obtained by performing a 10-folds cross-validation. For each stage, we report the performance measures computed independently from the precedent phases.

[1] A prefix of length n is defined here as the first n letters of a string.

In the last subsection we report the results obtained on the test set for the three run submitted:

- *run 01*: *CRF* as NER approach and *NNLM Linking* as NEL system;
- *run 02*: *CRF+Gazetteers* as NER approach and *NNLM Linking* as NEL system;
- *run 03*: *CRF+Gazetteers* as NER approach and *Learning2Link* with Decision Tree (DT) as NEL system.

3.1 Named Entity Recognition

We report the results of CRF, in terms of Precision (P), Recall (R) and F1-Measure (F1) in Table 1, according to the two investigated configurations: CRF and CRF+Gazetteers. First of all, we can note the poor recognition performances obtained in both configurations, which are mainly due to the limited amount of training data. These poor performances are highlighted even more by looking at the entity types *Thing* (20), *Event* (15) and *Character* (18), whose limited number of instances do not allow CRF to learn any linguistic pattern to recognize them. For the remaining types, CRF+Gazetteers is able to improve Precision but at some expenses of Recall.

Table 1: Entity Recognition Results

	CRF			**CRF+Gazetteers**		
Label	P	R	F1	P	R	F1
Thing	0	0	0	0	0	0
Event	0	0	0	0	0	0
Character	0	0	0	0	0	0
Location	0.56	0.40	0.47	0.64	0.40	0.5
Organization	0.43	0.24	0.31	0.60	0.20	0.30
Person	0.50	0.30	0.37	0.69	0.21	0.33
Product	0.12	0.11	0.11	0.31	0.10	0.16
Overall	0.37	0.24	0.29	0.57	0.20	0.30

The low recognition performance have a great impact on the subsequent steps of the pipeline. To this purpose, we will report the result of Entity Linking and NIL clustering by considering an oracle NER (i.e. a perfect named entity recognition system) in the following subsections.

3.2 Named Entity Linking

We report the Precision (P), Recall (R) and F-measure (F1) of the Strong Link Match (SLM) measure for each addressed approach for NEL in Table 2. The results have been computed assuming the NER system as an oracle, i.e., every entity mention is correctly recognized and classified.

Table 2: Strong Link Match measure.

	P	**R**	**F1**
NNLM Linking	0.619	**0.635**	**0.627**
L2L DT	0.733	0.371	0.492
L2L MLP	0.684	0.333	0.448
L2L NB	0.614	0.312	0.414
L2L LR	0.709	0.278	0.399
L2L SVM-Polynomial	0.721	0.27	0.393
L2L VP	0.696	0.274	0.393
L2L BN	**0.741**	0.266	0.392
L2L SVM-Radial	0.724	0.264	0.387
L2L SVM-Linear	0.686	0.266	0.384

Regarding the Learning2Link approach, we evaluate the results for each machine learning model considered. Although the low performances in terms of F-measure, we can highlight that Decision Tree (DT) is a leaner algorithm with the highest Strong Link Match F-measure. On the other hand, low recall scores could be attributed to the inability of the retrieval system to find the "correct" link in the top-10 candidate list. A list of irrelevant candidate resources results in uninformative similarity scores, which causes the learning models to predict a target class where none of the candidate resources is a suitable match for an entity mention.

NNLM Linking shows significant results, proving the importance of not considering an entity mention as a mere string but instead use a representation that is able to capture a deeper meaning of the word/entity.

3.3 NIL Clustering

Assuming every non-NIL entity has been correctly classified, our system for NIL clustering achieves a CEAF score of 0.994. We remark that NILs in the data set are very fragmented and a baseline system of one cluster per entity is capable of reaching a score of 0.975. Our algorithm however puts NILs represented in the tweets by the same string or sharing a significant portion of their strings in the same cluster; the reason why it does not get a perfect score is that either the same entity appears in tweets not sharing common words, and thus belonging to different components of the word graph (same NIL, different clusters), or that two entities are too similar and there is not enough context to distinguish them (different NILs, same cluster). As the data set is very sparse, these phenomena are

Table 3: Experimental results on the test set

run ID	MC	STMM	SLM	Score
run 01	0.193	0.166	0.218	0.192
run 02	**0.208**	**0.194**	**0.270**	**0.222**
run 03	0.207	0.188	0.213	0.203

very likely to occur. Finally, we notice that the NIL clustering performance strongly depends on the Named Entity Recognition and Linking output: if two occurrences of the same NIL are mistakenly assigned to different types, they will never end up in the same cluster.

3.4 Overall

The results of the submitted runs are reported in Table 3. The first column shows the given configuration, the other columns report respectively the F-measure of: Strong Link Match (SLM), Strong Typed Mention Match (STMM) and Mention Ceaf (MC).

As a first consideration we can highlight that involving CRF (*run 01*), instead of the CRF+Gazetteers configuration (*run 02* and *run 03*), has lead to a significant decrease of the performance, even more substantial than the one reported in Section 3.1.

Given the best NER configuration, the NNLM approach (*run 02*) is the one with better performances confirming the results presented in Section 3.2. As expected, the low recognition performance of the NER system strongly affected the NEL performance resulting in low results compared to the ones obtained considering an oracle NER.

The main limitation of the proposed pipeline emerged to be the Named Entity Recognition step. As mentioned before, one of the main problems is the availability of training data to induce the probabilistic model. A higher number of instances could improve the generalization abilities of Conditional Random Fields, resulting in a more reliable named entity recognizer. An additional improvement concerns the inclusion of information related to the Part-Of-Speech in the learning (and inference) phase of Conditional Random Fields. To this purpose, the Italian TreeTagger could be adopted to obtain the Part-Of-Speech for each token in tweets and to enclose this information into the feature functions of Conditional Random Fields. A further improvement relates to the use of extended gazetteers (not only related to the Italian language) especially related to the types *Event* and *Character* (which in most of the cases are English-based named entities). A final improvement could be achieved by introducing an additional step between the named entity recognition and the subsequent steps. To this purpose, the available Knowledge Base could be exploited as distant supervision to learn a "constrained" Topic Model (Blei et al., 2003) able to correct the type prediction given by Conditional Random Fields. This solution could not only help to overcome the limitation related to the reduced number of training instances, but could also have a good impact in terms of type corrections of named entities.

4 Conclusion

In this paper, we described a Named Entity Recognition and Linking framework for microposts that participated in EVALITA 2016 NEEL-IT challenge as UNIMIB team. We further provided an overview of our system for recognizing entity mentions from Italian tweets and introduced novel approach for linking them to suitable resources in an Italian knowledge base.

We observed a particularly poor performance of the Coditional Random Fields in the Named Entity Recognition phase, mainly due to lack of appropriate instances of entity types. Regarding the Named Entity Linking step, NNLM Linking shows significant results, proving the importance of a high-level representation able to capture deeper meanings of entities. Further, the Learning2Link phase turns out to be a promising approach, given the small amount of training instances, although, there is a considerable scope for improvement if more candidate resources are used. Other similarity measures can also be experimented with, while studying their impact on the feature space.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Davide Caliano, Elisabetta Fersini, Pikakshi Manchanda, Matteo Palmonari, and Enza Messina. 2016. Unimib: Entity linking in tweets using jaro-winkler distance, popularity and coherence. In *Proceedings of the 6th International Workshop on Making Sense of Microposts (# Microposts)*.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël

Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285. Association for Computational Linguistics.

David Hope and Bill Keller. 2013. Maxmax: a graph-based soft clustering algorithm applied to word sense induction. In *Computational Linguistics and Intelligent Text Processing*, pages 368–381. Springer.

Matthew A Jaro. 1995. Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7):491–498.

Paul McNamee. 2010. Hltcoe efforts in entity linking at tac kbp 2010. In *Proceedings of the 3rd Text Analysis Conference Workshop*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3:1–12, jan.

Helmut Schmid. 1994. Probabilistic part-of speech tagging using decision trees. In *New methods in language processing*, page 154. Routledge.