Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015
3-4 December 2015, Trento

Accademia University Press

Entity Linking for Italian Tweets

Pierpaolo Basile, Annalina Caputo and Giovanni Semeraro

DOI: 10.4000/books.aaccademia.1304
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: 11 November 2016
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899200008

Entity Linking for Italian Tweets

Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro
Dept. of Computer Science - University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{firstname.lastname}@uniba.it

Abstract

English. Linking entity mentions in Italian tweets to concepts in a knowledge base is a challenging task, due to the short and noisy nature of these short messages and the lack of specific resources for Italian. This paper proposes an adaptation of a general purpose Named Entity Linking algorithm, which exploits the similarity measure computed over a Distributional Semantic Model, in the context of Italian tweets. In order to evaluate the proposed algorithm, we introduce a new dataset of tweets for entity linking that we developed specifically for the Italian language.

Italiano. *La creazione di collegamenti tra le menzioni di un'entità in tweet in italiano ed il corrispettivo concetto in una base di conoscenza è un compito problematico a causa del testo nei tweet, generalmente corto e rumoroso, e della mancanza di risorse specifiche per l'italiano. In questo studio proponiamo l'adattamento di un algoritmo generico di Named Entity Linking, che sfrutta la misura di similarità semantica calcolata su uno spazio distribuzionale, nel contesto dei tweet in Italiano. Al fine di valutare l'algoritmo proposto, inoltre, introduciamo un nuovo dataset di tweet per il task di entity linking specifico per la lingua italiana.*

1 Introduction

In this paper we address the problem of entity linking for Italian tweets. Named Entity Linking (NEL) is the task of annotating entity mentions in a portion of text with links to a knowledge base. This task usually requires as first step the recognition of portions of text that refer to named entities (*entity recognition*). The linking phase follows, which usually subsumes the entity disambiguation, i.e. selecting the proper concept from a restricted set of candidates (e.g. New York city or New York state). 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 unstructured text.

Initially developed for reasonably long and clean text, such as news articles, NEL techniques usually show unsatisfying performance on noisy, short and poorly written text constituted by microblogs such as Twitter. These difficulties notwithstanding, with an average of 500 billion posts being generated every day[1], tweets represent a rich source of information. Twitter-based tasks like user interest discovery, tweet recommendation, social/economical analysis, and so forth, could benefit from such a kind of semantic features represented by named entities linked to a knowledge base. Such tasks become even more problematic when the tweet analysis involves languages different from English. Specifically, in the context of Italian language, the lack of language-specific resources and annotated tweet datasets complicates the assessment of NEL algorithms for tweets.

Our main contributions to this problem are:

- An adaptation of a Twitter-based NEL algorithm based on a Distributional Semantic Model (DSM-TEL), which needs no specific Italian resources since it is completely unsupervised (Section 3).
- An Italian dataset of manually annotated tweets for NEL. To the best of our knowledge, this is the first Italian dataset of such a type. Section 2 reports details concerning the annotation phase and statistics about the

[1] http://www.internetlivestats.com/twitter-statistics/

dataset.

- An evaluation of well known NEL algorithms available for Italian language on this dataset, comparing their performance with our DSM-TEL algorithm in terms of both entity recognition and linking. Section 4 shows and analyses the results of that evaluation.

2 Dataset

One of the main limitations to the development of specific algorithms for tweet-based entity linking in Italian lies on the dearth of datasets for training and assessing the quality of such techniques. Hence, we built a new dataset by following the guidelines proposed in the #Microposts2015 Named Entity Linking (NEEL) challenge[2] (Rizzo et al., 2015). We randomly selected 1,000 tweets from the TWITA dataset (Basile and Nissim, 2013), the first corpus of Italian tweets. For each tweet, we first select the named entities (NE). A NE is a string in the tweet representing a proper noun, excluding the preceding article (like "il", "lo", "la", etc.) and any other prefix (e.g. "Dott.", "Prof.") or post-posed modifier. More specifically, an entity is any proper noun that: 1) belongs to one of the categories specified in a taxonomy and/or 2) can be linked to a DBpedia concept. This means that some concepts have a NIL DBpedia reference; these concepts belong to one of the categories but they have no corresponding concept in DBpedia. The taxonomy is defined by the following categories: Thing[3], Event, Character, Location, Organization, Person and Product.

We annotated concepts by using the canonicalized dataset of Italian DBpedia 2014[4]. For specific Italian concepts that are not linked to an English article, we adopt the localized version of DBpedia. Finally, some concepts have an Italian Wikipedia article but they are not in DBpedia; in that case we linked the entity by using the Wikipedia URL. Entities represented neither in DBpedia nor Wikipedia are linked to NIL.

The annotation process poses some challenges specific to the Twitter context. For example, entities can be part of a user mention or tag; all these strings are valid entities: #[Alemanno], and @[CarlottaFerlito]. The '#' and '@' characters are not considered as part of the annotation.

This first version of the dataset was annotated by only one annotator, and comprises 756 entity mentions, with a mean of about 0.75 entities for each tweet. The distribution of entities in categories is as follows: 301 Persons, 197 Organizations, 124 Locations, 96 Products, 18 Things, 11 Events and 9 Characters. 63% of tweets links to a DBpedia concept, about 30% of them is annotated as NIL, 6% links to an URL of a Wikipedia page, while only one entity links to an Italian DBpedia concept.

The dataset[5] is composed of two files: the data and the annotation file. The data file contains pairs of tweet id and text, each listed on a different line. The annotation file consists of a line for each tweet id, which is followed by the start and the end offset[6] of the annotation, the linked concept and the category. All values are separated by the TAB character. For example, for the tweet: *"290460612549545984 @CarlottaFerlito io non ho la forza di alzarmi e prendere il libro! Help me"*, we have the annotation: *"290460612549545984 1 16 http://dbpedia.org/resource/Carlotta˙Ferlito Person"*.

3 DSM-TEL algorithm

We propose an Entity Linking algorithm specific for Italian tweets that adapts the original method proposed during the NEEL challenge (Basile et al., 2015b). Our algorithms consists of two-steps: the initial identification of all possible entity mentions in a tweet followed by the linking of the entities through the disambiguation algorithm. We exploit DBpedia/Wikipedia twice in order to (1) extract all the possible surface forms related to entities, and (2) retrieve glosses used in the disambiguation process. In this case we use as gloss the extended abstract assigned to each DBpedia concept. To speed up the recognition of entities we build an index where each surface form (entity) is paired with the set of all its possible DBpedia concepts. The surface forms are collected by analysing all the internal links in the Italian Wikipedia dump. Each internal link reports the surface form and the linked Wikipedia page that corresponds to a DB-

[2] http://www.scc.lancs.ac.uk/research/workshops/microposts2015/challenge/

[3] Languages, ethnic groups, nationalities, religions, ...

[4] This dataset contains triples extracted from Italian Wikipedia articles whose resources have an equivalent English article.

[5] Available at: https://github.com/swapUniba/neel-it-twitter

[6] Starting from 0.

pedia resource. The index is built by exploiting the Lucene API[7]. Specifically for each possible surface form a document composed of two fields is created. The first field stores the surface form, while the second one contains the list of all possible DBpedia concepts that refer to the surface form in the first field. The entity recognition module exploits this index in order to find entities in a tweet. Given a tweet, the module performs the following steps:

1. Tokenization of the tweet using the Tweet NLP API[8]. We perform some pre-processing operations to manage hashtags and user mentions; for example we split tokens by exploiting upper-case characters: "CarlottaFerlito" $\Longrightarrow$ "Carlotta Ferlito";
2. Construction of a list of candidate entities by exploiting all n-grams up to six words;
3. Query of the index and retrieval of the top 100 matching surface forms for each candidate entity;
4. Scoring each surface form as the linear combination of: a) a string similarity function based on the Levenshtein Distance between the candidate entity and the surface form in the index; b) the Jaccard Index in terms of common words between the candidate entity and the surface form in the index;
5. Filtering the candidate entities recognized in the previous steps: entities are removed if the score computed in the previous step is below a given threshold. In this scenario we empirically set the threshold to 0.66;
6. Finally, we filter candidate entities according to the percentage of words that: (1) are stop words, (2) are common words[9]; and (3) do not contain at least one upper-case character. We remove the entity if one of the aforementioned criteria is above the 33%.

The output of the entity recognition module is a list of candidate entities with their set of candidate DBpedia concepts.

For the disambiguation, we exploit an adaptation of the distributional Lesk algorithm proposed by Basile et al. (Basile et al., 2015a; Basile et al., 2014) for disambiguating named entities. The algorithm replaces the concept of word overlap initially introduced by Lesk (1986) with the broader concept of semantic similarity computed in a distributional semantic space. Let $e_1, e_2, ...e_n$ be the sequence of entities extracted from the tweet, the algorithm disambiguates each target entity e_i by computing the semantic similarity between the glosses of concepts associated with the target entity and its context. The context and the gloss are represented as the vector sum of words they are composed of in a Distributional Semantic Model (DSM). The similarity between the two vectors, computed as the cosine of the angle between them, takes into account the word co-occurrence evidences previously collected through a corpus of documents. We exploit the word2vec tool[10] (Mikolov et al., 2013) in order to build a DSM, by analyzing all the pages in the last Italian Wikipedia dump[11]. The semantic similarity score is combined with a function which takes into account the frequency of the concept usage. More details are reported in (Basile et al., 2015a; Basile et al., 2014; Basile et al., 2015b).

4 Evaluation

The evaluation aims to compare several entity linking tools for Italian language exploiting the proposed dataset. We include in the evaluation our method that is an adaptation of the system that participated in the NEEL challenge for English tweets (Basile et al., 2015b).

We select three tools able to perform entity linking for Italian: TAGME, Babelfy, and TextRazor. TAGME (Ferragina and Scaiella, 2010) has a particular option that enables a special parser for Twitter messages. This parser has been designed to better handle entities in tweets like URL, user mentions and hash-tag. However, some other tools are not developed specifically for Twitter. For example, Babelfy (Moro et al., 2014) is an algorithm for entity linking and disambiguation developed for generic texts that uses BabelNet (Navigli and Ponzetto, 2012) as knowledge source. The third system is TextRazor[12], a commercial system able to recognize, disambiguate and link entities in ten languages, including Italian. Systems are compared using the typical metrics of precision, recall and F-measure. We compute the metrics in two settings: the **exact match** set requires that both

[7] http://lucene.apache.org/

[8] `http://www.ark.cs.cmu.edu/TweetNLP/`

[9] We exploit the list of 1,000 most frequent Italian words: `http://telelinea.free.fr/italien/1000_parole.html`

[10] https://code.google.com/p/word2vec/

[11] We use 400 dimensions for vectors analysing only terms that occur at least 25 times.

[12] `https://www.textrazor.com/`

start and end offsets match the gold standard annotation, while in **non exact match** the offsets provided by the systems can differ of two positions with respect to the gold standard.

Each algorithm provides a different output that needs some post-processing operations in order to make it comparable with our annotation standard. Most of the annotations are made with respect to the canonicalized version of DBpedia, while only 6% of the dataset is annotated using Wikipedia page URLs or the localized version (just one). Babelfy is able to directly provide canonicalized DBpedia URIs. When a BabelNet concept is not mapped to a DBpedia URIs, we return a NIL instance. TAGME returns Italian Wikipedia page titles that we easily translate into DBpedia URIs. We firstly try to map the page title in the canonicalized DBpedia, otherwise into the Italian one. TextRazor supplies Italian Wikipedia URLs or English Wikipedia URLs that we map to DBpedia URIs. Our algorithm provides Italian DBpedia URIs that we translate into canonicalized URIs when it is possible, otherwise we keep the Italian URIs. To recap: all algorithms are able to provide canonicalized and localized DBpedia URIs, only Babelfy is limited to canonicalized URIs.

Table 1: Results of the entity recognition evaluation with exact and non exact match.

	Exact match			Non exact match		
System	P	R	F	P	R	F
Babelfy	.431	.161	.235	.449	.168	.244
TAGME	.363	.458	.405	.391	.492	.436
TextRazor	.605	.310	.410	.605	.310	.410
DSMTEL	.470	.505	**.487**	.495	.532	**.513**

Table 1 reports the results about the entity recognition task with respect to exact and non exact match respectively. DSM-TEL provides the best results followed by TextRazor (exact match) and TAGME (non exact match), while the low performance of Babelfy proves that it is not able to tackle the irregular language used in tweets. In all the cases TextRazor achieves the best precision.

Entity linking performance are reported in Tables 2. It is important to underline that a correct linking requires the proper recognition of the entity involved. TextRazor achieves the best performance in the entity linking task with an F-measure in both exact and non exact match of 0.280.

Moreover, in order to understand if the recognition and linking tasks pose more challenges for Italian language, we evaluated all the systems on an English dataset. Although the two datasets are not directly comparable (due to the different sizes and the number of entities involved per tweet), we run an experiment over the Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge dataset (Rizzo et al., 2015) (Table 2). The evaluation results show a different behaviour of the systems for the English language. F-measure values are slightly lower than for Italian and TextRazor almost always outperforms other systems, with the only exception of TAGME for the linking with non exact match.

Table 2: Results of the entity linking evaluation with exact and non exact match.

	Exact match			Non exact match		
System	P	R	F	P	R	F
Babelfy	.318	.119	.173	.322	.120	.175
TAGME	.226	.284	.252	.235	.296	.262
TextRazor	.413	.212	**.280**	.413	.212	**.280**
DSM-TEL	.245	.263	.254	.254	.272	.263

Table 3: F-Measure results for English #Microposts2015 NEEL dataset.

	Recognition		Linking	
System	Exact	No Exact	Exact	No Exact
Babelfy	.134	.137	.102	.104
TAGME	.352	.381	.290	**.311**
TextRazor	**.460**	**.485**	**.294**	.295
DSMTEL	.442	.467	.284	.299

5 Conclusion

We tackled the problem of entity linking for Italian tweets. Our contribution is threefold: 1) we build a first Italian tweet dataset for entity linking, 2) we adapted a distributional-based NEL algorithm to the Italian language, and 3) we compared state-of-the-art systems on the built dataset. As for English, the entity linking task for Italian tweets turn out to be quite difficult, as pointed out by the very low performance of all systems employed. As future work we plan to extend the dataset in order to provide more examples for training and testing data.

References

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia, June. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015a. Uniba: Combining distributional semantic models and sense distribution for multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 360–364, Denver, Colorado, June. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Fedelucio Narducci. 2015b. UNIBA: Exploiting a Distributional Semantic Model for Disambiguating and Linking Entities in Tweets. In *Proceedings of the the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference (WWW 2015)*, volume 1395, pages 62–63. CEUR-WS.

Paolo Ferragina and Ugo Scaiella. 2010. Fast and accurate annotation of short texts with wikipedia pages. *IEEE Software*, 29(1):70–75.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proc. of SIGDOC '86*, SIGDOC '86, pages 24–26. ACM.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of ICLR Work.*

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Giuseppe Rizzo, Amparo Elizabeth Cano Basave, Bianca Pereira, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. 2015. 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 (WWW 2015)*, volume 1395, pages 44–53. CEUR-WS.