FBK-NLP at NEEL-IT: Active Learning for Domain Adaptation

Anne-Lyse Minard[1,2], Mohammed R. H. Qwaider[1], Bernardo Magnini[1]
[1] Fondazione Bruno Kessler, Trento, Italy
[2] Dept. of Information Engineering, University of Brescia, Italy
{minard,qwaider,magnini}@fbk.eu

Abstract

English. In this paper we present the FBK-NLP system which participated to the NEEL-IT task at Evalita 2016. We concentrated our work on domain adaptation of an existed Named Entity Recognition tool. Particularly, we created a new annotated corpus for the NEEL-IT task using an Active Learning method. Our system obtained the best results for the task of Named Entity Recognition, with an F1 of 0.516.

Italiano. *In questo articolo descriviamo il sistema FBK-NLP con il quale abbiamo partecipato al task NEEL-IT a Evalita 2016. Ci siamo concentrati sull'adattamento di un sistema per il riconoscimento di entità al dominio dei tweets. In particolare, abbiamo creato un nuovo corpus usando una metodologia basata su Active Learning. Il sistema ha ottenuto i risultati migliori sul sottotask di riconoscimento delle entità, con una F1 di 0,516.*

1 Introduction

This paper describes the FBK-NLP system which participated to the NEEL-IT task at EVALITA 2016 (Basile et al., 2016). The NEEL-IT task focuses on Named Entity Linking in tweets in Italian. It consists in three steps: Named Entity Recognition and Classification (NER) in 7 classes (person, location, organization, product, event, thing and character); the linking of each entity to an entry of DBpedia; the clustering of the entities. Our participation to the task was mainly motivated by our interest in experimenting on the application of Active Learning (AL) for domain adaptation, in particular to adapt a general purpose Named Entity Recognition system to a specific domain (tweets) by creating new annotated data.

The system follows 3 steps: entity recognition and classification, entity linking to DBpedia and clustering. Entity recognition and classification is performed by the EntityPro module (Pianta and Zanoli, 2007), which is based on machine learning and uses the SVM algorithm. Entity linking is performed using the named entity disambiguation module developed within the NewsReader project for several languages including Italian. In addition we used the Alignments dataset (Nechaev et al., 2016), a resource which provides links between Twitter profiles and DBpedia. Clustering step is string-based, i.e. two entities are part of the same cluster if they are equal.

The paper is organized as follows. In Section 2 we present the domain adaptation of the Named Entity Recognition tool using Active Learning. Then in Section 3 we describe the system with which we participated to the task and in Section 4 the results we obtained as well as some further experiments. Finally we conclude the paper with a discussion in Section 5.

2 Domain Adaptation for NER

We have at our disposal a system for Named Entity Recognition and Classification, a module of the TextPro pipeline (Pianta et al., 2008) called EntityPro (Pianta and Zanoli, 2007), which works for 4 named entity categories in the news domain. It is trained on the publicly available Italian corpus I-CAB (Magnini et al., 2006). I-CAB is composed of news articles from the regional newspaper "L'Adige", is annotated with person, organization, location and geo-political entities, and was used for the Named Entity Recognition task at Evalita 2007 and 2009.[1] However, no annotated data are available for the task of NER in tweets for

[1] www.evalita.it/

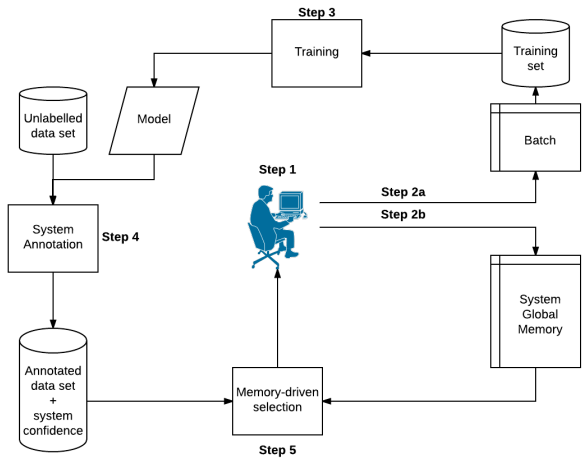

Figure 1: Architecture of the TextPro-AL platform

Italian.

As we were interested in applying Active Learning (AL) methods to the production of training data, we decided to annotate manually our own set of domain specific training data using AL method.[2] Active Learning is used in order to select the most informative examples to be annotated, instead of selecting random examples.

We exploited TextPro-AL (Magnini et al., 2016), a platform which integrates a NLP pipeline, i.e. TextPro (Pianta et al., 2008), with a system of Active Learning and an annotation interface based on MTEqual (Girardi et al., 2014). TextPro-AL enables for a more efficient use of the time of the annotators.

2.1 The TextPro-AL platform

The architecture of the TextPro-AL platform is represented in Figure 1. The AL cycle starts with an annotator providing supervision on a tweet automatically tagged by the system (step 1): the annotator is asked to revise the annotation in case the system made a wrong classification. At step 2a the annotated tweet is stored in a batch, where it is accumulated with other tweets for re-training, and, as a result, a new model (step 3) is produced. This model is then used to automatically annotate a set of unlabeled tweets (step 4) and to assign a confidence score[3] to each annotated tweet. At step 2b the manually annotated tweet is stored in the Global Memory of the system with the information about the manual revision. At step 5 a single tweet is selected from the unlabeled dataset through a specific selection strategy (see Algorithm 1). The selected tweet is removed from the unlabeled set and is given for revision to the annotator.

The Global Memory contains the revision done by the annotator for each tweet. In particular we are interested in the entities wrongly annotated by the system, which are used to select new tweets to be annotated. Each entity (or error) saved in the memory is used up to 6 times in order to select new tweets. From the unlabeled dataset, the system selects the most informative instance (i.e. with the lowest confidence score) that contains one of the errors saved in the Global Memory (GM). The selection strategy is detailed in Algorithm 1. In a first step the system annotates the tweets of the unlabeled dataset. Then the tweets are sorted from the most informative to the less informative and browsed. The first tweet in the list that contains an error saved in the GM is selected to be revised by the annotator. If no tweets are selected through this process, the system picks one tweet randomly.

Algorithm 1: Algorithm of the selection strategy

```
Data: NESet = {NE_1 ... NE_n}
begin
    NESortedList ←
      getMostInformativeInstances(NESet);
    repeat
        instance, sample ←
          NESortedList.next();
        if inMemory(instance) and
          revised(instance) then
            return sample;
    until NESortedList.hasNext();
    return getRandomSample(NESet);
```

2.2 Available Data

As unlabeled database of tweets in the AL process we used around 8,000 tweets taken from the development set of Sentipolc 2014[4] (Basile et al., 2014) and the Twita corpus[5] (Basile and Nissim, 2013).

[2]The annotated data made available by the organizers of the task were used partly as test data and partly as a reference for the annotators (see Section 2.2).

[3]The confidence score is computed as the average of the margin estimated by the SVM classifier for each entity.

[4]http://www.di.unito.it/~tutreeb/sentipolc-evalita14/tweet.html

[5]http://valeriobasile.github.io/twita/about.html

class	AL tweets	NEEL-IT dev			news corpus
		test 70%	dev 30%	total	
# sent/tweets	2,654	700	300	1,000	458
# tokens	49,819	13,283	5,707	18,990	8,304
Person	1628	225	90	315	293
Location	343	89	43	132	115
Organization	723	185	63	248	224
Product	478	67	41	108	-
Event	133	12	3	15	-
Thing	15	15	4	19	-
Character	50	15	1	16	-

Table 1: Statistics about the used datasets. The numbers of tokens for the tweets are computed after the tokenizaion, i.e. the hashtags and aliases can be split in more than one token and the emoji are composed by several tokens (see Section 3.1).

The development data provided by the NEEL-IT organizers is composed by 1000 annotated tweets. We split it in two parts: 30% for development (used mainly as a reference for the annotators) and 70% for evaluation (referred to as *test 70%*).

We decided to retrain EntityPro using a smaller training set to be able to change the behavior of the model more quickly. In particular we used a sub-part of the training data used by EntityPro, i.e. 6.25% of the training set of the NER task at Evalita 2007,[6] for a total of 8,304 tokens (referred to as *news corpus* in the remainder of the paper).

In order to determine the portion to be used, we tested the performance of EntityPro using as training data different portions of the corpus (50%, 25%, 12.5% and 6.25%) on *test 70%*. The best results were obtained using 6.25% of the corpus (statistics about this corpus is given in Table 1).

2.3 Manual Annotation of Training Data with TextPro-AL

In our experimentation with TextPro-AL for domain adaptation we built the first model using the *news corpus* only. Evaluated on *test 70%*, it reached an F1 of 41.62 with a precision of 54.91 and a recall of 33.51. It has to be noted that with this model only 3 categories of entities can be recognized: person, location and organization. Then every time that 50 new tweets were annotated, the system was retrained and evaluated on the *test 70%* corpus. The learning curves of the system are presented in Figure 2. In total we were able to manually annotate 2,654 tweets for a total of 3,370 entities (we will refer to this corpus as *AL tweets*), which allowed us to obtain an F1 of 53.22 on *test 70%*. Statistics about the corpus are presented in Table 1.

3 Description of the system

3.1 Entity Recognition and Classification

The preprocessing of the tweets is done using the TextPro tool suite[7] (Pianta et al., 2008), in particular using the tokenizer, the PoS tagger and the lemmatizer. The rules used by the tokenizer have been lightly adapted for the processing of tweets, for example to be able to split Twitter profile names and hashtags in small units. The PoS tagger and the lemmatizer have been used as they are, without any adaptation.

In order to avoid some encoding problems we replaced all the emoji by their Emoji codes (e.g. :confused_face:) using the python package emoji 0.3.9.[8]

The task of entity recognition and classification is performed using an adapted version of the EntityPro module (Pianta and Zanoli, 2007). EntityPro performs named entity recognition based on machine learning, using an SVM algorithm and the Yamcha tool (Kudo and Matsumoto, 2003). It exploits a rich set of linguistic features, as well as gazetteers. We added to the features an orthographic feature (capitalized word, digits, etc.) and bigrams (the first two characters and the last two).

The classifier is used in a one-vs-rest multi-classification strategy. The format used for the

[6] http://www.evalita.it/2007/tasks/ner

[7] http://textpro.fbk.eu/

[8] http://pypi.python.org/pypi/emoji/

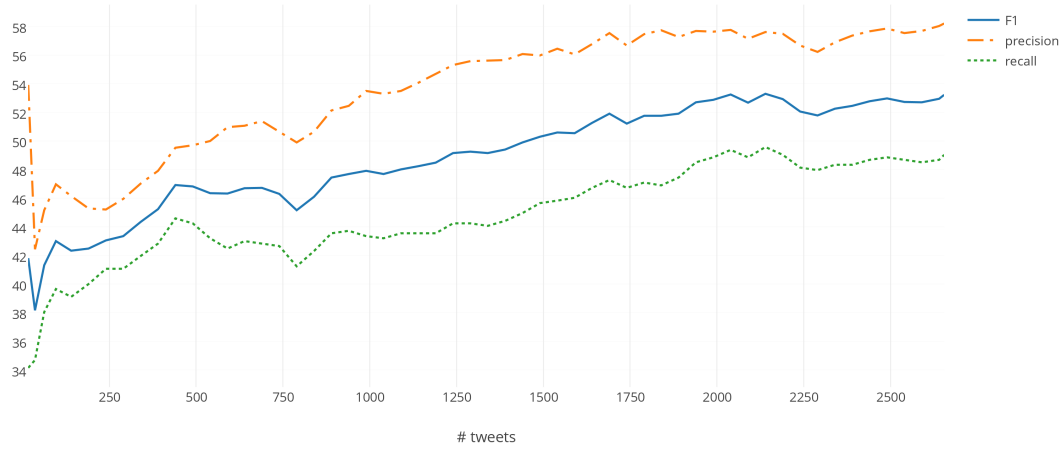

Figure 2: Learning curves of the system (recall, precision and F1)

annotation is the classic IOB2 format. Each token is labeled either as B- followed by the entity class (person, location, organization, product, event, thing or character) for the first token of an entity, I- followed by the entity class for the tokens inside an entity or O if the token is not part of an entity.

3.2 Entity Linking

Entity Linking is performed using the Named Entity Disambiguation (NED) module developed within the NewsReader Project[9] supplemented with the use of a resource for Twitter profiles linking. The NED module is a wrapper around DBpedia spotlight developed within NewsReader and part of the ixa-pipeline.[10] Each entity recognized by the NER module is sent to DBpedia Spotlight which returns the most probable URI if the entity exists in DBpedia.

The tweets often contain aliases, i.e. user profile names, which enable the author of the tweet to refer to other Twitter users. For example @edoardofasoli and @senatoremonti in the following tweet: *@edoardofasoli @senatoremonti Tutti e due.* In order to identify the DBpedia links of the aliases in the tweets we used the Alignments dataset (Nechaev et al., 2016). The Alignments dataset is built from the 2015-10 edition of English DBpedia, which contains DBpedia links aligned with Twitter profiles. It has 920,625 mapped DBpedia entries to their corresponding user profile(s) with a confidence score.

A procedure is built to query Twitter to get the Twitter profile id from the alias of a user, then query the Alignments dataset to get the corresponding DBpedia link if it exists.

3.3 Clustering

The clustering task aims at gathering the entities referring to the same instance and at assigning to them an identifier, either a DBpedia link or a corpus based identifier. We performed this task applying a basic string matched method, i.e. we consider that two entities are part of the same cluster if their strings are the same.

4 Results

We submitted 3 runs to the NEEL-IT task; they differ from the data included in the training dataset of EntityPro:

- Run 1: *news corpus* and *AL tweets*
- Run 2: *news corpus*, *AL tweets* and NEEL-IT devset
- Run 3: *AL tweets* and NEEL-IT devset

The official results are presented in the first part of Table 2. Our best performance is obtained with the run 3, with a final score of 0.49.

[9] http://www.newsreader-project.eu/

[10] https://github.com/ixa-ehu/ixa-pipe-ned

runs	training set	tagging	linking	clustering	final score
run 1	*news corpus* + *AL tweets*	0.509	0.333	0.574	0.4822
run 2	*news corpus* + *AL tweets* + NEEL-IT devset	0.508	0.346	0.583	0.4894
run 3	*AL tweets* + NEEL-IT devset	0.516	0.348	0.585	**0.4932**
run 4*	*AL tweets* + NEEL-IT devset	0.517	0.355	0.590	**0.4976**
run 5*	*news corpus*	0.378	0.298	0.473	0.3920
run 6*	NEEL-IT devset	0.438	0.318	0.515	0.4328
run 7*	*news corpus* + NEEL-IT devset	0.459	0.334	0.541	0.4543

Table 2: Results of the submitted runs (runs 1 to 3) and of some further experiments (runs 4 to 8). The official task metrics are "strong_typed_mention_match", "strong_link_match" and "mention_ceaf", and refer to "tagging", "linking" and "clustering" respectively.

After the evaluation period, we have run further experiments, which are marked with an asterisk in Table 2. The run 4 is a version of run 3 in which we have removed the wrong links to the Italian DBpedia (URIs of type `http://it.dbpedia.org/`). For runs 5, 6 and 7, EntityPro is trained using the *news corpus* alone, the NEEL-IT devset, and both respectively.

In Table 3, we present the performances of our systems in terms of precision, recall and F1 for the subtask of named entity recognition and classification. We observed that using the NEEL-IT devset the precision of our system increased, instead using the news corpus the recall increased.

	precision	recall	F1
run 1	0.571	0.459	0.509
run 2	0.581	0.451	0.508
run 3	0.598	0.454	0.516

Table 3: Results for the task of named entity recognition and classification

5 Discussion

We have described our participation to the NEEL-IT task at Evalita 2016. Our work focused on the task of named entity recognition, for which we get the best results. We were interested in the topic of domain adaptation. The domain adaptation includes two aspects: the type of the documents and the named entity classes of interest. Using EntityPro, an existing NER tool, and the TextPro-AL platform, we created a training dataset for NER in tweets, for the 7 classes identified in the task.[11] With this new resource our system obtained an F1 of 0.516 for named entity recognition.

Our work has been concentrated on the use of Active Learning for the domain adaptation of a NER system. On the other hand, the Micro-NEEL team (Corcoglioniti et al., 2016) focuses on the task of Entity Linking, using The Wiki Machine (Palmero Aprosio and Giuliano, 2016). We have combined our NER system with the Micro-NEEL system. For the tagging subtask we used the same configuration than run 4 (*AL tweets* + NEEL-IT devset). The results obtained with combination of the two systems are 0.517 for tagging, 0.465 for linking and 0.586 for clustering. The final score is 0.5290, surpassing all the runs submitted to the task.

One of the main difficulty in identifying named entities in tweets is the problem of the splitting of hashtags and aliases (e.g. the identification of *Monti* in *@senatoremonti*). We adapted the TextPro tokenizer to split in small units those sequences of characters, but it works only if the different words are capitalized or separated by some punctuation signs (e.g. _ or -). A more complex approach should be used, using a dictionary to improve the splitting.

Named entity categories covered in this task are seven: person, location, organization, product, event, thing and character. The first three categories are the classical ones and cover the highest number of named entities in several corpora. Table 1 gives us an evidence of the prominence of these three classes. With the AL method we used, we were able to annotate new tweets containing entities of the less represented classes, in particular for product, event and character. However the class thing is still not well represented in our corpus and the classes unbalanced. In the future we plan to add in the TextPro-AL platform the pos-

[11] We will soon make available the new training set from the website of the HLT-NLP group at FBK (`http://hlt-nlp.fbk.eu/`).

sibility for the annotators to monitor the Global Memory used in the AL process in order to give precedence to examples containing entities of not well represented classes.

Acknowledgments

This work has been partially supported by the EUCLIP (EUregio Cross LInguistic Project) project, under a collaboration between FBK and Euregio.[12]

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.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy.

Pierpaolo Basile, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Francesco Corcoglioniti, Alessio Palmero Aprosio, Yaroslav Nechaev, and Claudio Giuliano. 2016. MicroNeel: Combining NLP Tools to Perform Named Entity Detection and Linking on Microposts. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Christian Girardi, Luisa Bentivogli, Mohammad Amin Farajian, and Marcello Federico. 2014. Mt-equal: a toolkit for human assessment of machine translation output. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, August 23-29, 2014, Dublin, Ireland*, pages 120–123.

Taku Kudo and Yuji Matsumoto. 2003. Fast Methods for Kernel-based Text Analysis. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 24–31, Stroudsburg, PA, USA.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-2006)*.

Bernardo Magnini, Anne-Lyse Minard, Mohammed R. H. Qwaider, and Manuela Speranza. 2016. TEXTPRO-AL: An Active Learning Platform for Flexible and Efficient Production of Training Data for NLP Tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*.

Yaroslav Nechaev, Francesco Corcoglioniti, and Claudio Giuliano. 2016. Linking knowledge bases to social media profiles.

Alessio Palmero Aprosio and Claudio Giuliano. 2016. The Wiki Machine: an open source software for entity linking and enrichment. *ArXiv e-prints*.

Emanuele Pianta and Roberto Zanoli. 2007. Entitypro: Exploiting svm for italian named entity recognition. *Intelligenza Artificiale numero speciale su Strumenti per lelaborazione del linguaggio naturale per litaliano EVALITA 2007*, 4(2):69–70.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

[12] http://www.euregio.it