Bi-directional LSTM-CNNs-CRF for Italian Sequence Labeling

Pierpaolo Basile, Giovanni Semeraro, Pierluigi Cassotti
Department of Computer Science, University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{firstname.surname}@uniba.it, pierluigicassotti@gmail.com

Abstract

English. In this paper, we propose a Deep Learning architecture for sequence labeling based on a state of the art model that exploits both word- and character-level representations through the combination of bidirectional LSTM, CNN and CRF. We evaluate the proposed method on three Natural Language Processing tasks for Italian: PoS-tagging of tweets, Named Entity Recognition and Super-Sense Tagging. Results show that the system is able to achieve state of the art performance in all the tasks and in some cases overcomes the best systems previously developed for the Italian.

Italiano. *In questo lavoro viene descritta un'architettura di Deep Learning per l'etichettatura di sequenze basata su un modello allo stato dell'arte che utilizza rappresentazioni sia a livello di carattere che di parola attraverso la combinazione di LSTM, CNN e CRF. Il metodo è stato valutato in tre task di elaborazione del linguaggio naturale per la lingua italiana: il PoS-tagging di tweet, il riconoscimento di entità e il Super-Sense Tagging. I risultati ottenuti dimostrano che il sistema è in grado di raggiungere prestazioni allo stato dell'arte in tutti i task e in alcuni casi riesce a superare i sistemi precedentemente sviluppati per la lingua italiana.*

1 Background and Motivation

Deep Learning (DL) gained a lot of attention in last years for its capacity to generalize models without the need of feature engineering and its ability to provide good performance. On the other hand good performance can be achieved by accurately designing the architecture used to perform the learning task. In Natural Language Processing (NLP) several DL architectures have been proposed to solve many tasks, ranging from speech recognition to parsing. Some typical NLP tasks can be solved as sequence labeling problem, such as part-of-speech (PoS) tagging and Named Entity Recognition (NER). Traditional high performance NLP methods for sequence labeling are linear statistical models, including Conditional Random Fields (CRF) and Hidden Markov Models (HMM) (Ratinov and Roth, 2009; Passos et al., 2014; Luo et al., 2015), which rely on hand-crafted features and task/language specific resources. However, developing such task/language specific resources has a cost, moreover it makes difficult to adapt the model to new tasks, new domains or new languages. In (Ma and Hovy, 2016), the authors propose a state of the art sequence labeling method based on a neural network architecture that benefits from both word- and character-level representations through the combination of bidirectional LSTM, CNN and CRF. The method is able to achieve state of the art performance in sequence labeling tasks for the English without the use of hand-crafted features.

In this paper, we exploit the aforementioned architecture for solving three NLP tasks in Italian: PoS-tagging of tweets, NER and Super Sense Tagging (SST). Our research question is to prove the effectiveness of the DL architecture in a different language, in this case Italian, without using language specific features. The results of the evaluation prove that our approach is able to achieve state of the art performance and in some cases it is able to overcome the best systems developed for the Italian without the usage of specific language resources.

The paper is structured as follows: Section 2 provides details about our methodology and summarizes the DL architecture proposed in (Ma and Hovy, 2016), while Section 3 shows the results of the evaluation. Final remarks are reported in Section 4.

2 Methodology

Our approach relies on the DL architecture proposed in (Ma and Hovy, 2016), where the authors combine two aspects previously exploited separately: 1) the use of a character-level representation (Chiu and Nichols, 2015); 2) the addition of an output layer based on CRF (Huang et al., 2015). The architecture is sketched in Figure 1: the input level of the Convolution Neural Network is represented by the character-level representation. A dropout layer (Srivastava et al., 2014) is applied before feeding the CNN with character embeddings. Then the character embeddings are concatenated with the word embeddings to form the input for the Bi-directional LSTM layer. The dropout layer is also applied to output vectors from the LSTM layer. The output layer is based on Conditional Random Fields and it modifies the output vectors of the LSTM in order to find the best output sequence. The CRF layer is useful for learning correlations between labels in neighborhoods, for example generally a noun follows an article in PoS-tagging, or the I-ORG tag[1] cannot follow the I-PER tag in the NER task.

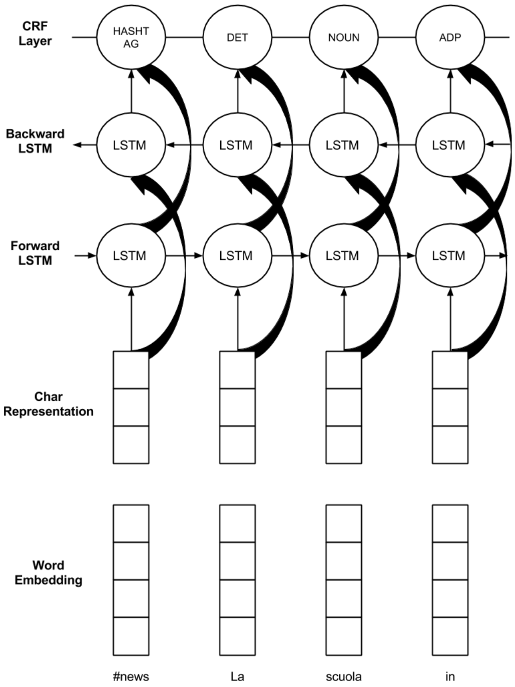

Figure 1: The DL architecture for the sequence labeling.

The aforementioned architecture can be easily adapted to other languages since it does not rely on language dependent features. The only components outside the architecture are the word embeddings that can be built by relying on a corpus of documents of the specific language. In Section 3, we provide details about the setup of the architecture parameters and the building of word embeddings for Italian, in particular we adopt two different word embeddings: ones for PoS-tagging and ones for NER and SST. Moreover, we re-implement[2] the architecture by using the Keras[3] framework and Tensorflow[4] as back-end.

3 Evaluation

We provide an evaluation in the context of three sequence labeling tasks: 1) PoS tagging of Italian tweets; 2) NER of Italian news and 3) Super Sense Tagging. All tasks are performed using Italian datasets, in particular we exploit data coming from the last edition (2016) of EVALITA[5] (Basile et al., 2016) and previous ones (2009 (Magnini and Cappelli, 2009) and 2011[6]). EVALITA[7] is a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language. The usage of a standard benchmark allows to compare our system with the state of the art approaches for the Italian language.

Each task has its specific parameters, but there are some ones that are in common as reported in Table 1. We do not perform any parameters optimization and we use the values proposed in the English evaluation (Ma and Hovy, 2016). We choose this strategy in order to not reduce the training set since validation set is not provided in all the tasks.

3.1 PoS tagging of Tweets

The goal of the task is to perform PoS-tagging of tweets. The task is more challenging with respect to the classical PoS-tagging due to the short and noisy nature of tweets. For the evaluation we adopt the dataset used during the EVALITA 2016 PoSTWITA task (Bosco et al., 2016) in order

[1] Generally, the NER task uses the IOB2 schema for data annotation.

[2] The code is available on line: https://github.com/pippokill/bilstm-cnn-crf-seq-ita

[3] https://keras.io/

[4] https://www.tensorflow.org/

[5] https://github.com/evalita2016/data

[6] http://www.evalita.it/2011/working_notes

[7] http://www.evalita.it/

Parameter	**Value**
Framework	Keras 2.0.1
Back-end	Tensorflow 1.1.0
Char embed. dimension	30
Word embed. dimension	300
Window size	3
LTSM dimension	200 (bi-LTSM 400)
Optimization	Adadelta
Gradient clipping	5.0
Epochs	100 (PoS), 60 (NER and SST)

Table 1: Parameters' values.

System	Accuracy
UNIBA-twita	**.9334**
UNIBA-itwiki	.9199
UNIBA-random300	.8790
ILC-CNR	.9319
UniDuisburg	.9286
UniBologna UnOFF	.9279

Table 2: Results for the PoSTWITA task.

to compare our system with the other EVALITA participants. The dataset contains 6,438 tweets (114,967 tokens) for training and 300 tweets (4,759 tokens) for test. The metric used for the evaluation is the classical tagging accuracy: it is defined as the number of correct PoS tag assignment divided by the total number of tokens in the test set. Participants can predict only one tag for each token.

All the top-performing PoSTWITA systems are based on Deep Neural Networks and, in particular, on LSTM, moreover most systems use word or character embeddings as inputs for their systems. This makes other systems more similar to the one proposed in this paper.

Results of the evaluation are reported in Table 2, our best approach (*UNIBA-twita*) is able to overcome the first three PoSTWITA participants. (*UNIBA-twita*) exploits a corpus of 70M tweets randomly extracted from Twita, a collection of about 800M tweets, for building the word embeddings. It is important to underline that the best system (*ILC-CNR*) (Cimino and Dell'orletta, 2016) in PoSTWITA uses a biLSTM and a RNN by exploiting both word and char embeddings, moreover it use further features based on morpho-syntactic category and spell checker. The good performance of our system probably depends by the CRF layer and the corpus used for building the word embeddings. This hypothesis is supported by the fact that the configuration (*UNIBA-itwiki*) based on word embeddings extracted from Wikipedia obtains the worst result. The configuration *UNIBA-random300* adopts random embeddings, we report this result in order to underline the importance of pre-trained word embeddings. Moreover, the second best system (*UniDuisburg*) (Horsmann and Zesch, 2016) in PoSTWITA exploits a CRF classifier using several features without a DL architecture, while the system *UniBologna UnOFF* (Tamburini, 2016) uses a BiLSTM with a CRF layer by exploiting word embeddings and additional morphological features.

3.2 NER Task

Three tasks about named entities have been organized during the EVALITA evaluation campaigns, respectively in 2007 (Speranza, 2007), 2009 (Speranza, 2009), and 2011 (Lenzi et al., 2013). In this paper we take into account the 2009 edition since the I-CAB dataset [8]used in the evaluation is the same adopted in 2009. In 2007 a different version of I-CAB was used, while in 2011 the task was focused on data transcribed by an ASR system. The I-CAB dataset consists of a set of news manually annotated with four kinds of entities: GPE (geo-political), LOC (location), ORG (organization) and PER (person). The dataset contains 525 news for training and 180 for testing for a total number of 11,410 annotated entities for training and 4,966 ones for testing. The dataset is provided in the IOB2 format.

We build word embeddings by exploiting the Italian version of Wikipedia. Word2vec is used for creating embeddings with a dimension of 300, we remove all words that have less than 40 occurrences in Wikipedia, for the other parameters we adopt the standard values provided by word2vec.

Results of the evaluation are reported in Table 3 and Table 4. Table 3 reports precision (P), recall (R) and F1-measure (F1) for different configurations of the system. In particular: *no-case-sensitive* does not perform lowercase of words for both word embeddings and the lookup table, while *case-sensitive* does it. The *random* configuration randomly initializes embeddings without using pre-trained embeddings, while *no char* does not adopt char embeddings. The results show that

[8] http://ontotext.fbk.eu/icab.html

Configuration	ALL			GPE	LOC	ORG	PER
	P	R	F1	F1	F1	F1	F1
no-case-sensitive	.8286	.8182	**.8234**	.8561	.6220	.6587	.9239
case-sensitive	.8220	.8084	.8151	.8444	.6305	.6421	.9178
random	.7153	.6885	.7017	.7564	.4809	.5209	.8037
no char	.8305	.7426	.7841	.8492	.6200	.5945	.8714

Table 3: Results for the Italian NER task using different configurations.

System	ALL			GPE	LOC	ORG	PER
	P	R	F1	F1	F1	F1	F1
UNIBA	.8286	.8182	**.8234**	.8561	.6220	.6587	.9239
FBK_ZanoliPianta	.8407	.8002	.8200	.8513	.5124	.7056	.8831
UniGen_Gesmundo_r2	.8606	.7733	.8146	.8336	.5081	.7108	.8741
UniTN-FBK-RGB_r2	.8320	.7908	.8109	.8525	.5224	.6961	.8689

Table 4: Results for the Italian NER task compared with other EVALITA 2009 participants.

the best performance is obtained by applying lowercase, moreover the contribution of char embeddings is significant.

Table 4 reports the result of our best configuration (*no-case-sensitive*) with respect to the other EVALITA 2009 participants. The system is able to outperform the first three EVALITA participants thanks to the best performance in recall. All the first three participants adopt classical classification methods: the first system (Zanoli et al., 2009) combines two classifiers (HMM and CRF), the second participant (Gesmundo, 2009) uses a Perceptron algorithm, while the third participant (Mehdad et al., 2009) adopts Support Vector Machine and feature selection. We can conclude that the DL architecture is more effective in the model generalization and in tackling the data sparsity problem. This behavior is supported by the good performance in recognizing LOC entities, in fact the LOC class represents about the 3% of annotated entities in both training and test. Other two systems (Nguyen and Moschitti, 2012; Bonadiman et al., 2015) able to overcome the EVALITA 2009 participants have been proposed in the literature. The former (Nguyen and Moschitti, 2012) achieves the 84.33% of F1 by using re-ranking techniques and the combination of two state-of-the-art NER learning algorithms: conditional random fields and support vector machines. The latter (Bonadiman et al., 2015) exploits a Deep Neural Network with a log-likelihood cost function and a recurrent feedback mechanism to ensure the dependencies between the output tags. This system is able to achieves the 82.81% of F1, a performance comparable with our DL architecture.

3.3 Super Sense Tagging

The Super-Sense Tagging (SST) task (Dei Rossi et al., 2011) consists in annotating each significant entity in a text, like nouns, verbs, adjectives and adverbs, within a general semantic taxonomy defined by the WordNet lexicographer classes (called super-senses, for a total of 45 senses). SST can be considered as a task half-way between NER and Word Sense Disambiguation (WSD): it is an extension of NER, since it uses a larger set of semantic categories, and it is an easier and more practical task with respect to WSD. The dataset has been tagged using the IOB2 format as for the NER task and contains about 276,000 tokens for training and about 50,000 for testing. The metric adopted for the evaluation is the F1, results of the evaluation are reported in Table 5. As word embeddings we use the same ones adopted for the NER task and built upon Wikipedia with lowercase.

System	F1
UNIBA-pos-Adagrad	**.7871**
UNIBA-pos	.7787
UNIBA	.7453
UNIBA-SVMcat	.7866
UNIPI-run3	.7827

Table 5: Results for the Super-Sense Tagging task.

The best performance (*UNIBA-pos-Adagrad*) is obtained using Adagrad instead of Adadelta (*UNIBA-pos*) as optimization method. Moreover, we exploits PoS-tags as additional features, while *UNIBA* uses only tokens and word/char embeddings. The difference in performance between *UNIBA-pos* and *UNIBA* proves the effectiveness of the PoS-tag in this task. The best system in EVALITA 2011 SST task, *UNIBA-SVMcat* (Basile, 2013, 2011), is very close to our best configuration. This system combines lexical and distributional features through an SVM classifier, while the second system (*UNIPI-run3*) (Attardi et al., 2011) exploits lexical features and a Maximum Entropy classifier.

4 Conclusions and Future Work

We propose an evaluation of a state of the art DL architecture for sequence labeling in the context of the Italian language. In particular, we consider three tasks: PoS-tagging of tweets, Named Entity Recognition and Super-Sense tagging. All tasks exploit data coming from EVALITA a standard benchmark for the evaluation of Italian NLP systems. Our system is able to achieve good performance in all the tasks without using hand-crafted features. Analyzing the results, we observe the importance of building word embeddings on appropriate corpora and we note that the system in the SST task is not able to generalize a good model without the pos-tag feature, this underline the importance of this kind of feature in the SST task. As future work, we plan to perform a parameters optimization by reducing the training set and using a portion as validation set. Using less data for training could affect the final performance and it could be interesting to have insights on the trade-off between training on more examples versus the parameters optimization.

Acknowledgments

This work is partially supported by the project "Multilingual Entity Liking" funded by the Apulia Region under the program FutureInResearch.

References

Giuseppe Attardi, Luca Baronti, Stefano Dei Rossi, and Maria Simi. 2011. SuperSense Tagging with a Maximum Entropy Classifier and Dynamic Programming. In *Working Notes of EVALITA 2011*.

P. Basile. 2013. Super-sense tagging using support vector machines and distributional features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7689 LNAI:176–185.

P. Basile, F. Cutugno, M. Nissim, V. Patti, and R. Sprugnoli. 2016. EVALITA 2016: Overview of the 5th evaluation campaign of natural language processing and speech tools for Italian. 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)*. CEUR Workshop Proceedings, volume 1749.

Pierpaolo Basile. 2011. UNIBA: Super-sense Tagging at EVALITA 2011. In *Working Notes of EVALITA 2011*.

Daniele Bonadiman, Aliaksei Severyn, and Alessandro Moschitti. 2015. Deep neural networks for named entity recognition in italian. In *CLiC-it 2015 Proceedings of the second Italian Conference on Computational Linguistics*. page 51.

C. Bosco, F. Tamburini, A. Bolioli, and A. Mazzei. 2016. Overview of the EVALITA 2016 Part of speech on twitter for Italian 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)*. CEUR Workshop Proceedings, volume 1749.

Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional LSTM-CNNs. *arXiv preprint arXiv:1511.08308* .

A. Cimino and F. Dell'orletta. 2016. Building the state-of-the-art in POS tagging of Italian Tweets. 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)*. CEUR Workshop Proceedings, volume 1749.

Stefano Dei Rossi, Giulia Di Pietro, and Maria Simi. 2011. EVALITA 2011: Description and Results of the SuperSense Tagging Task. In *Working Notes of EVALITA 2011*.

Andrea Gesmundo. 2009. Bidirectional sequence classification for named entities recognition. In *Proceedings of the Workshop Evalita 2009*.

T. Horsmann and T. Zesch. 2016. Building a social media adapted PoS tagger using flexTag - A case study on Italian tweets. 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)*. CEUR Workshop Proceedings, volume 1749.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* .

Valentina Bartalesi Lenzi, Manuela Speranza, and Rachele Sprugnoli. 2013. Named entity recognition on transcribed broadcast news at EVALITA 2011. In *Revised Papers from EVALITA11: International Workshop on the Evaluation of Natural Language and Speech Tools for Italian*. Springer, volume 7689, pages 86–97.

G. Luo, X. Huang, C.-Y. Lin, and Z. Nie. 2015. Joint named entity recognition and disambiguation. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. pages 879–888.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* .

Bernardo Magnini and Amedeo Cappelli. 2009. Introduction to Evalita 2009. In *Proceedings of the Workshop Evalita 2009*.

Yashar Mehdad, Vitalie Scurtu, and Evgeny Stepanov. 2009. Italian named entity recognizer participation in NER task @ Evalita 09. In *Proceedings of the Workshop Evalita 2009*.

Truc-Vien T Nguyen and Alessandro Moschitti. 2012. Structural reranking models for named entity recognition. *Intelligenza Artificiale* 6(2):177–190.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv:1404.5367* .

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 147–155.

Manuela Speranza. 2007. Evalita 2007: the named entity recognition task. In *Proceedings of the Workshop Evalita 2007*.

Manuela Speranza. 2009. The named entity recognition task at evalita 2009. In *Proceedings of the Workshop Evalita 2009*.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

F. Tamburini. 2016. A BiLSTM-CRF PoS-tagger for Italian tweets using morphological information. 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)*. CEUR Workshop Proceedings, volume 1749.

Roberto Zanoli, Emanuele Pianta, and Claudio Giuliano. 2009. Named entity recognition through redundancy driven classifiers. In *Proceedings of the Workshop Evalita 2009*.