Bi-directional LSTM-CNNs-CRF for Italian Sequence Labeling and Multi-Task Learning

Pierpaolo Basile*
Università degli Studi di Bari Aldo Moro

Pierluigi Cassotti**
Università degli Studi di Bari Aldo Moro

Lucia Siciliani†
Università degli Studi di Bari Aldo Moro

Giovanni Semeraro‡
Università degli Studi di Bari Aldo Moro

In this paper, we propose a Deep Learning architecture for several Italian Natural Language Processing tasks 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. This architecture provided state of the art performance in several sequence labeling tasks for the English language. We exploit the same approach for the Italian language and extend it for performing a multi-task learning involving PoS-tagging and sentiment analysis. 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.

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, such as part-of-speech (PoS) tagging and Named Entity Recognition (NER), can be solved as sequence labeling problem. 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, Kumar, and McCallum 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

* Department of Computer Science, Via E. Orabona, 4 - 70125 Bari (Italy).
E-mail: pierpaolo.basile@uniba.it
** Department of Computer Science, Via E. Orabona, 4 - 70125 Bari (Italy).
E-mail: pierluigicassotti@gmail.com
† Department of Computer Science, Via E. Orabona, 4 - 70125 Bari (Italy).
E-mail: lucia.siciliani@uniba.it
‡ Department of Computer Science, Via E. Orabona, 4 - 70125 Bari (Italy).
E-mail: giovanni.semeraro@uniba.it

and CRF. The method is able to achieve state of the art performance in sequence labeling tasks for the English with no need of using 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). We have already proposed an evaluation on these tasks (Basile, Semeraro, and Cassotti 2017) using the same architecture, but without correctly optimizing hyperparameters due to the lack of a validation set. In this paper, we describe a procedure for hyperparameters optimization based on k-fold cross-validation. Moreover, we extend this architecture for performing a multi-task learning (Zhang and Yang 2017; Ruder 2017) involving PoS-tagging and sentiment analysis. This has been possible by using training data with multiple levels of annotations. In particular, we exploit training data about tweets annotated with PoS-tag, polarity and irony.

Our research goal is twofold: 1) to prove the effectiveness of the DL architecture in a different language - in this case Italian - without using language specific features; 2) to investigate the performance of the architecture in the context of multi-task learning, when multiple levels of annotations on the same training set are exploited.

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 using no 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, Xu, and Yu 2015). The architecture is sketched in Figure 1: The input level of the Convolutional Neural Network (CNN) is represented by the character-level representation. A dropout layer (Srivastava et al. 2014) with convolution and max pooling 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 (bi-LSTM) layer as sketched in Figure 2. The dropout layer is also applied to output vectors from the LSTM layer. The output layer is based on Conditional Random Fields (CRF) 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, usually a noun follows an article in PoS-tagging, or the I-ORG tag[1] cannot follow the I-PER tag in the NER task.

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: One for PoS-tagging and one

1 The IOB2 schema for data annotation is usually adopted in the NER task.

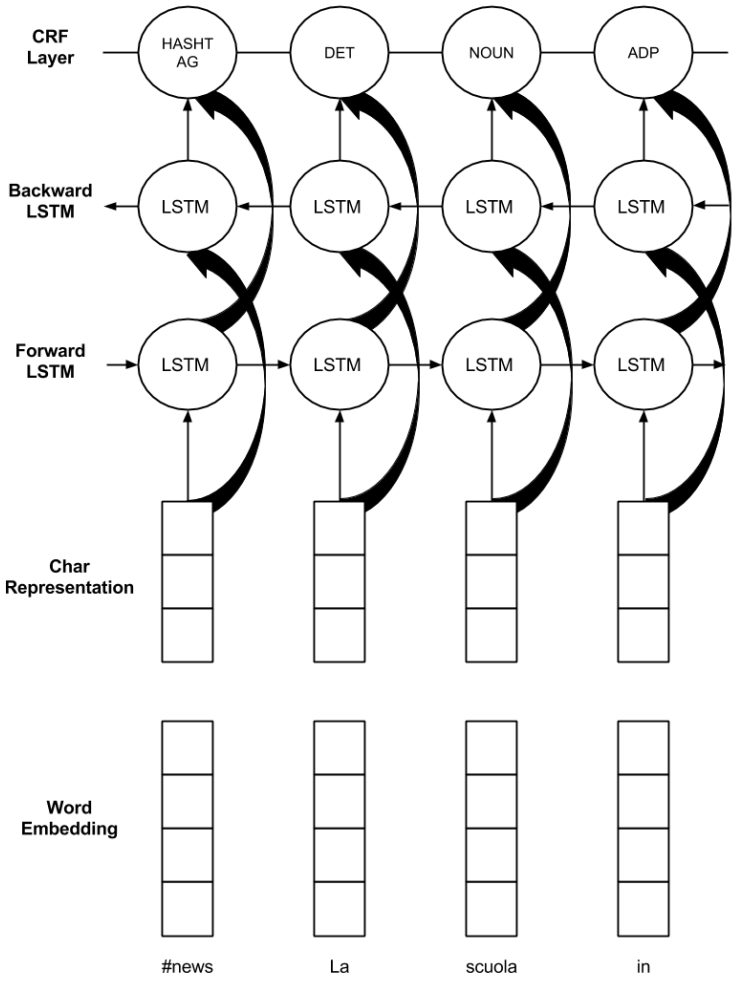

Figure 1
The DL architecture for sequence labeling.

for NER and SST. Moreover, we re-implemented[2] the architecture by using the Keras[3] framework and Tensorflow[4] as back-end.

2.1 Multi-task learning

We extend the previous architecture for performing multi-task learning. In particular, we want to jointly learn PoS-tag, polarity and irony. It is important to underline that PoS-tag is assigned to each token occurring in the sentence, while polarity and irony are

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/`

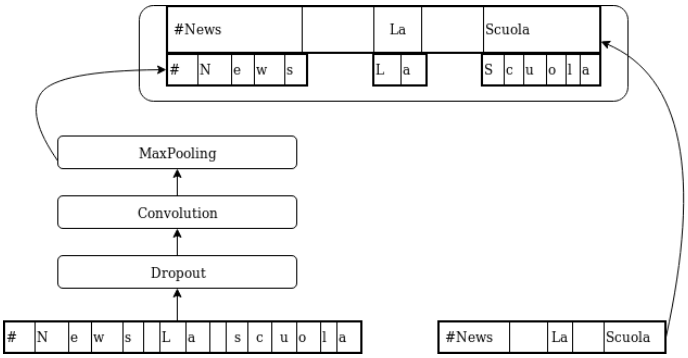

Figure 2
The input level of the DL architecture.

assigned to the whole sentence. We follow a hard parameter sharing approach (Ruder 2017) in which we have some shared layers in the bottom of the network and task-specific layers on the top.

The proposed architecture depends on the particular sentiment analysis task (Barbieri et al. 2016) that we want to perform. The task is deeply described in Section 3.4. Here, we want to point out that we need to solve four binary classification tasks: Subjectivity (true/false), positive polarity (true/false), negative polarity (true/false) and irony (true/false). We want to train a classifier for these classes jointly with the PoS-tagging task.

For this purpose, we add a parallel layer to the CRF one. In particular, a new layer based on a bi-LSTM is added using the same dimension of the first LSTM layer. Then, a dropout layer is applied and the final classes probabilities are computed by a binary cross entropy function for each class. In this case the last layer does not predict a tag for each token, but it predicts only one tag[5] for each classification task (subjectivity, positive, negative, irony). The multi-task architecture is sketched in Figure 3.

We can notice that the output of the first bi-LSTM layer is the input of the CRF layer for predicting PoS-tags, while each binary sentiment task is implemented by a new bi-LSTM layer and a cross entropy function.

3. Evaluation

We provide an evaluation in the context of four tasks for the Italian language. The first three tasks concern sequence labeling: 1) PoS-tagging of Italian tweets; 2) NER of Italian news 3) Super Sense Tagging. The fourth task concerns sentiment classification on Twitter. In such task, we try to jointly learn PoS-tagging and sentiment classification.

All tasks are performed using Italian datasets. In particular we exploit data coming from the last edition (2016) of EVALITA[6] (Basile et al. 2016) and the previous ones (2009 (Magnini and Cappelli 2009) and 2011[7]). EVALITA[8] is a periodic evaluation campaign of NLP and speech tools for the Italian language. The usage of a standard benchmark

5 Assigned to the whole text.
6 https://github.com/evalita2016/data
7 http://www.evalita.it/2011/working_notes
8 http://www.evalita.it/

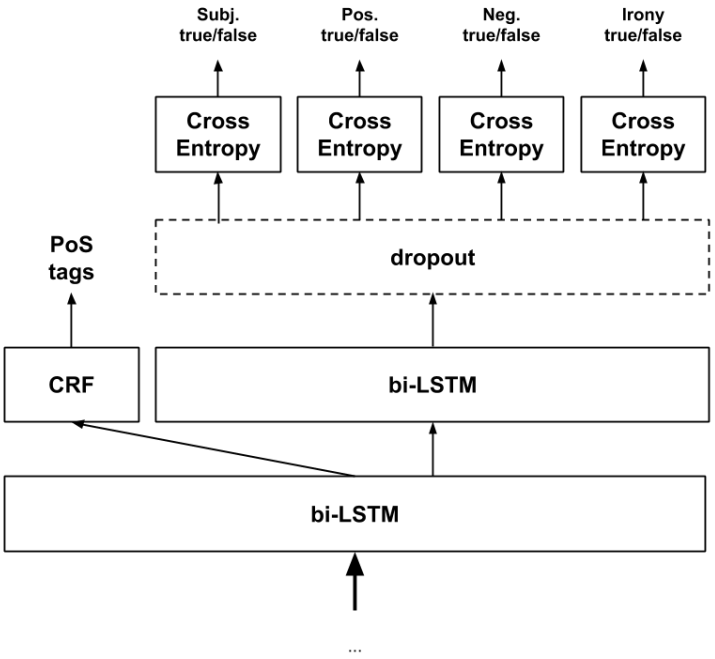

Figure 3
The DL architecture for multi-task learning.

allows us to compare our system with the state of the art approaches for the Italian language. In particular, EVALITA 2016 contains some shared-task data. We exploit these data for performing the multi-task evaluation.

Each task has its specific parameters, with parameters in Table 1 that are shared by all tasks.

Table 1
Parameters' values.

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)
Gradient clipping	5.0
Dropout	0.5

We perform parameters optimization using 5-fold cross-validation on training data since EVALITA does not provide a validation set. In particular, we perform optimization in order to choose the best optimization algorithm evaluating among Adadelta, Adagrad, Adam and SGD. Regarding SGD, we test several values of the initial learning

rate in the set {0.01, 0.0125, 0.15} and values for the decay rate in the set {0.01, 0.05, 0.1}. Moreover, we optimize the number of epochs (we set the maximum number of epochs to 60).

Results of the optimization procedure are reported in Table 2. Results about SGD parameters are removed since they give rise to lower performance.

Table 2
Results of the optimization

Task	Opt.Alg.	Epochs
PoS-tagging	Adadelta	60
NER	Adadelta	57
Supersense	Adagrad	60
Multi-task	Adam	60

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 canonical PoS-tagging task 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 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 testing. A training sample is reported in Figure 4, where each token is annotated with its PoS-tag. The metric used for the evaluation is the classical tagging accuracy defined as the number of correct PoS-tag assignments divided by the total number of tokens in the test set. Participants can predict only one tag for each token.

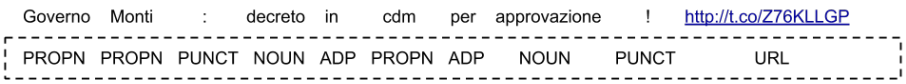

Figure 4
A PoS-tagging training sample.

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 3. Our approach (*DL-ita*) is able to provide results in line with the first three PoSTWITA participants. In (Basile, Semeraro, and Cassotti 2017), we report an accuracy of .9334 using the same system, but running 100 epochs. In this work, we set the maximum number of epochs to 60 during the optimization step in order to reduce the computation time. Nevertheless, results prove the effectiveness of the proposed architecture without exploiting task/language specific resources. The only used resource is a corpus of 70M tweets randomly extracted from Twita, a collection of about 800M tweets, for building the word embeddings.

Table 3
Results for the PoSTWITA task.

System	Accuracy
DL-ita	.9265
ILC-CNR	.9319
UniDuisburg	.9286
UniBologna UnOFF	.9279

It is important to underline that the best system (*ILC-CNR*) (Cimino and Dell'Orletta 2016) in PoSTWITA uses a bi-LSTM and an RNN by exploiting both word and character embeddings, moreover it uses further features based on morpho-syntactic categories and spell checker. 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 bi-LSTM 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, Speranza, and Sprugnoli 2013). In this paper we take into account the 2009 edition since the I-CAB dataset [9]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: the tag B (for "begin") denotes the first token of a Named Entity, I (for "inside") is used for all other tokens in a Named Entity, and O (for "outside") is used for all other words. The Entity type tags are: PER (for Person), ORG (for Organization), GPE (for GeoPolitical Entity), or LOC (for Location). A training sample is reported in Figure 5.

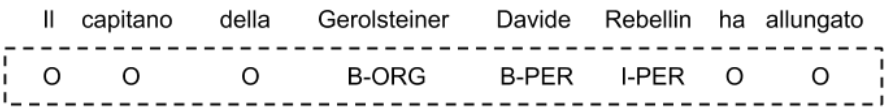

Figure 5
A NER training sample.

9 http://ontotext.fbk.eu/icab.html

We build word embeddings by exploiting the Italian version of Wikipedia. Word2vec (Mikolov et al. 2013) 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.

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

System	ALL			GPE	LOC	ORG	PER
	P	R	F1	F1	F1	F1	F1
DL-ita	.8236	.8197	.8217	.8579	.5905	.6673	.9203
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

Results of the evaluation are reported in Table 4, where our system (*DL-ita*) is compared with respect to the other EVALITA 2009 participants. The system outperforms 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, Pianta, and Giuliano 2009) combines two classifiers (HMM and CRF), the second participant (Gesmundo 2009) uses a Perceptron algorithm, while the third (Mehdad, Scurtu, and Stepanov 2009) adopts Support Vector Machine and feature selection. We can conclude that the DL architecture is more effective in 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, Severyn, and Moschitti 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, Severyn, and Moschitti 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 achieve 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, Di Pietro, and Simi 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. A training sample is reported in Figure 6, where each token occurring in WordNet is annotated with its super sense. The metric adopted for the evaluation is the F1. Results of the evaluation are reported in Table 5.

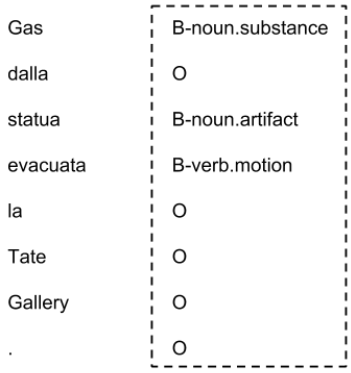

Figure 6
A SST training sample.

As word embeddings, we use the same ones adopted for the NER task and built upon Wikipedia with lowercase. Moreover, we exploit PoS-tags as additional features.

Table 5
Results for the Super-Sense Tagging task.

System	F1
DL-ita	.7864
UNIBA-SVMcat	.7866
UNIPI-run3	.7827

Our system (*DL-ita*) is very close to the best system in EVALITA 2011 SST task *UNIBA-SVMcat*. This system combines lexical and distributional features through an SVM classifier, in particular it exploits specific features such us: lemma, contextual PoS-tags, the super-sense assigned to the most frequent sense of the word and information about the grammatical conjugation of verbs. We plan to introduce this kind of features into the DL system in order to understand if this difference in performance still emerges. The second system (*UNIPI-run3*) (Attardi et al. 2011) exploits lexical features and a Maximum Entropy classifier.

3.4 Multi-task Evaluation

In this evaluation, we exploit shared data coming from EVALITA 2016. In particular, we use PoS-tagging and sentiment analysis data. We choose these two tasks because they share the largest training set. We plan to investigate more annotation layers when the number of shared examples in training/testing set will increase. PoS-tagging data have been previously described in Section 3.1, while sentiment data are taken from SENTIPOLC (Barbieri et al. 2016). SENTIPOC (*SENTIment POLarity Classification*) is a sentiment analysis task where systems are required to automatically annotate tweets with a tuple of boolean values indicating the message's subjectivity, its polarity (positive or negative), and whether it is ironic or not. For example, the following tweet "*Dopo due*

ore che stavo studiando italiano, mi sono accorta che avevo preso il libro sbagliato. #benecosi"[10] is annotated with the subjectivity, positive and irony tags.

The SENTIPOLC training set consists of 7,410 tweets (6,412 are shared with the PoSTWITA task), while the test set contains 2,000 tweets (300 are shared with PoSTWITA). In conclusion, we are able to train the multi-task architecture using 6,412 tweets, while the accuracy of PoS-tag is evaluated on 300 tweets and the performance on SENTIPOLC is computed on 2,000 tweets.

Table 6
Results for the PoSTWITA task using the multi-task architecture.

System	Accuracy
DL-ita-MT	.9246
ILC-CNR	.9319
UniDuisburg	.9286
UniBologna UnOFF	.9279
DL-ita	.9265

Table 6 reports the accuracy on PoS-Twita. The multi-task architecture (DL-ita-ML) obtains results similar to the single-task learning architecture (*DL-ita*). Results show that PoS-tag is not able to exploit information about polarity and irony for improving performance.

Table 7
Results for the SENTIPOLC task

System	Subj.	Positive	Negative	Total	Irony
DL-ita-ML	.7176	.6361	.6521	.6441	.5120
DL-ita	.7282	.6391	.6802	.6602	.4970
Unitor.1.u	**.7444**	.6354	.6885	.6620	.4728
Unitor.2.u	.7351	.6312	.6838	.6575	.4810
samskara.1.c	.7184	.5198	.6168	.5683	-
UniPI.2.c	.6937	.6850	.6426	**.6638**	-
tweet2check16.c	.6236	.6153	.5878	.6016	**.5412**
CoMoDI.c -	-	-	-	-	.5251
tweet2check14.c	.5843	.5660	.6034	.5847	.5162

Regarding the SENTIPOLC task, results in Table 7 show that the multi-task architecture is able to improve its performance in the irony task. However, performance decreases in the polarity task. In conclusion, information about the PoS-tag is useful for irony, but not in the subjective and polarity tasks. It is not easy to interpret the DL architecture and this is made even more difficult by the multi-task learning. For example, the following tweet was correctly classified by the DL-ita-ML and incorrectly

10 In English: "After two hours I was studying Italian, I realized that I had taken the wrong book. #welldone"

classified by the DL-ita: *Io mi lamento della gente che scrive ancora "freddy mercury" ma anche quella che scrive "jhonny cash" non scherza*[11].

Moreover, Table 7 reports results for the systems at the intersection between the first three systems of each SENTIPOLC subtask. Our system (DL-ita-ML) is able to achieve good results in each subtask and ranks 4 out of 18 in the subjective task, 6 out of 25 in the polarity task and 5 out of 12 in the irony task.

The best system in the subjective task, Unitor1.u (Castellucci, Croce, and Basili 2016), reports also good performance in the polarity task but poor performance in the irony task. In particular, Unitor1.u implements a workflow of several Convolutional Neural Networks classifiers in which sentiment specific information is injected using Polarity Lexicons (Basili, Croce, and Castellucci 2017) automatically acquired through the analysis of unlabeled collection of tweets. Conversely, our system does not exploit any additional resources.

The best system in the polarity task, UniPI.2 (Attardi et al. 2016), adopts Convolutional Neural Networks as *Unitor1.u* and exploits both word embeddings and Sentiment Specific word embeddings. This system ranks eighth in the subjective task and does not participate in the irony one.

Finally, the best system in the irony task, tweet2check16.c (Di Rosa and Durante 2016), is an industrial system which combines many different classifiers, each of which is built by using different machine learning algorithms and implementing different features. This system is able to achieve moderate performance in the polarity task, while poor performance are reported in the subjective task.

The CoMoDI.c (Frenda 2016) is a rule based system specifically developed for irony detection and it is able to achieve good performance[12] in the irony task. The samskara system (Russo and Monachini 2016) is the third in the final rank of the subjective task. This system uses a Naive Bayes classifier trained on a set of structural features specifically designed for the Twitter domain. However, this system achieves the worst performance in the final rank of the polarity task and it does not participate in the irony task.

In conclusion, the multi-task architecture obtains good performance in all SENTIPOLC subtasks. However, it does not achieve the best performance in any specific task. Moreover, we observe that the information about PoS-tags is useful in the irony subtask, while the use of polarity information in the PoS-tag results in a slight decrease in accuracy.

4. Conclusions and Future Work

We propose an evaluation of a state of the art DL architecture in the context of the Italian language. In particular, we consider three sequence labeling tasks: PoS-tagging of tweets, Named Entity Recognition and Super-Sense Tagging. We also propose a multi-task learning architecture involving PoS-tagging and sentiment analysis. 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 handcrafted features. Analyzing the results, we observe that our system is able to achieve

11 In English: *I complain about the people who still write "freddy mercury" but also the one who writes "jhonny cash" is not joking*.

12 second in the final rank

state of the art performance for the Italian language in all the sequence labeling tasks. This proves the effectiveness of the DL architecture in a different language - in this case Italian - without using language specific features

Regarding the multi-task learning, our architecture is able to achieve good performance in each subtask (subjectivity, polarity and irony) using the same architecture. In addition, the multi-task learning results show that the irony task benefits from the information provided by PoS-tags. In this work, we are able to investigate only PoS-tagging and sentiment analysis because they share the largest training set in EVALITA. We plan to investigate more annotation layers when the number of shared examples in training/testing set will increase.

As future work, we plan to investigate further multi-task learning architectures exploiting different strategies, such as the one proposed in (Hashimoto et al. 2016), where higher layers include short-cut connections to lower-level task predictions to reflect linguistic hierarchies.

Acknowledgments

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

References

Attardi, Giuseppe, 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*.

Attardi, Giuseppe, Daniele Sartiano, Chiara Alzetta, and Federica Semplici. 2016. Convolutional Neural Networks for Sentiment Analysis 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)*, Naples, Italy. aAccademia University Press.

Barbieri, Francesco, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification 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)*, Naples, Italy. aAccademia University Press.

Basile, Pierpaolo, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele 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)*, Naples, Italy. aAccademia University Press.

Basile, Pierpaolo, Giovanni Semeraro, and Pierluigi Cassotti. 2017. Bi-directional LSTM-CNNs-CRF for Italian Sequence Labeling. In Roberto Basili, Malvina Nissim, and Giorgio Satta Satta, editors, *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. Associazione Italiana di Linguistica Computazionale (AILC).

Basili, Roberto, Danilo Croce, and Giuseppe Castellucci. 2017. Dynamic polarity lexicon acquisition for advanced social media analytics. *International Journal of Engineering Business Management*, 9:1–18.

Bonadiman, Daniele, 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.

Bosco, Cristina, Fabio Tamburini, Andrea Bolioli, and Alessandro 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)*, Naples, Italy. aAccademia University Press.

Castellucci, Giuseppe, Danilo Croce, and Roberto Basili. 2016. Context-aware Convolutional Neural Networks for Twitter Sentiment Analysis in 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)*, Naples, Italy. aAccademia University Press.

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

Cimino, Andrea and Felice 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)*, Naples, Italy. aAccademia University Press.

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

Di Rosa, Emanuele and Alberto Durante. 2016. Tweet2Check evaluation at Evalita Sentipolc 2016. 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)*, Naples, Italy. aAccademia University Press.

Frenda, Simona. 2016. Computational rule-based model for Irony Detection in 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)*, Naples, Italy. aAccademia University Press.

Gesmundo, Andrea. 2009. Bidirectional sequence classification for named entities recognition. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*.

Hashimoto, Kazuma, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.

Horsmann, Tobias and Torsten 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)*, Naples, Italy. aAccademia University Press.

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

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

Luo, Gang, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint named entity recognition and disambiguation. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 879–888.

Ma, Xuezhe and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. *arXiv preprint arXiv:1603.01354*.

Magnini, Bernardo and Amedeo Cappelli. 2009. Introduction to Evalita 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*.

Mehdad, Yashar, Vitalie Scurtu, and Evgeny Stepanov. 2009. Italian named entity recognizer participation in NER task @ Evalita 09. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

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

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

Ruder, Sebastian. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Russo, Irene and Monica Monachini. 2016. Samskara Minimal structural features for detecting subjectivity and polarity in 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)*, Naples, Italy. aAccademia University Press.

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

Speranza, Manuela. 2009. The named entity recognition task at Evalita 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*.

Srivastava, Nitish, 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.

Tamburini, F. 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)*, Naples, Italy. aAccademia University Press.

Zanoli, Roberto, Emanuele Pianta, and Claudio Giuliano. 2009. Named entity recognition through redundancy driven classifiers. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*.

Zhang, Yu and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.