# Italian Event Detection Goes Deep Learning

## Abstract

This paper reports on a set of experiments with different word embeddings to initialize a state-of-the-art Bi-LSTM-CRF network for event detection and classification in Italian, following the EVENTI evaluation exercise. The net- work obtains a new state-of-the-art result by improving the F1 score for detection of 1.3 points, and of 6.5 points for classification, by using a single step approach. The results also provide further evidence that embeddings have a major impact on the performance of such architectures.

## Introduction

Current societies are exposed to a continuous flow of information that results in a large production of data (e.g. news articles, micro-blogs, social media posts, among others), at different moments in time. In addition to this, the consumption of information has dramatically changed: more and more people directly access information through social media platforms (e.g. Facebook and Twitter), and are less and less exposed to a diversity of perspectives and opinions. The combination of these factors may easily result in information overload and impenetrable “filter bubbles”. Events, i.e. things that happen or hold as true in the world, are the basic components of such data stream. Being able to correctly identify and classify them plays a major role to develop robust solutions to deal with the current stream of data (e.g. the storyline framework BIBREF0 ), as well to improve the performance of many Natural Language Processing (NLP) applications such as automatic summarization and question answering (Q.A.).

Event detection and classification has seen a growing interest in the NLP community thanks to the availability of annotated corpora BIBREF1 , BIBREF2 , BIBREF3 , BIBREF4 and evaluation campaigns BIBREF5 , BIBREF6 , BIBREF7 , BIBREF8 , BIBREF9 , BIBREF10 . In the context of the 2014 EVALITA

Workshop, the EVENTI evaluation exercise BIBREF11 was organized to promote research in Italian Temporal Processing, of which event detection and classification is a core subtask.

Since the EVENTI campaign, there has been a lack of further research, especially in the application of deep learning models to this task in Italian. The contributions of this paper are the followings: i.) the adaptation of a state-of-the-art sequence to sequence (seq2seq) neural system to event detection and classification for Italian in a single step approach; ii.) an investigation on the quality of existing Italian word embeddings for this task; iii.) a comparison against a state-of-the-art discrete classifier. The pre-trained models and scripts running the system (or re-train it) are publicly available. .

## Task Description

We follow the formulation of the task as specified in the EVENTI exercise: determine the extent and the class of event mentions in a text, according to the It-TimeML $<$ EVENT $>$ tag definition (Subtask B in EVENTI).

In EVENTI, the tag $<$ EVENT $>$ is applied to every linguistic expression denoting a situation that happens or occurs, or a state in which something obtains or holds true, regardless of the specific parts-of-speech that may realize it. EVENTI distinguishes between single token and multi-tokens events, where the latter are restricted to specific cases of eventive multi-word expressions in lexicographic dictionaries (e.g. “fare le valigie” [to pack]), verbal periphrases (e.g. “(essere) in grado di” [(to be) able to]; “c'è” [there is]), and named events (e.g. “la strage di Beslan” [Beslan school siege]).

Each event is further assigned to one of 7 possible classes, namely: OCCURRENCE, ASPECTUAL, PERCEPTION, REPORTING, I(NTESIONAL)_STATE, I(NTENSIONAL)_ACTION, and STATE. These classes are derived from the English TimeML Annotation Guidelines BIBREF12 . The TimeML event classes distinguishes with respect to other classifications, such as ACE BIBREF1 or FrameNet BIBREF13 , because they expresses relationships the target event participates in (such as factual, evidential, reported, intensional) rather than semantic categories denoting the meaning of the event. This means that the EVENT classes are assigned by taking into account both the semantic and the syntactic context of occurrence of the target event. Readers are referred to the EVENTI Annotation Guidelines for more details.

## Dataset

The EVENTI corpus consists of three datasets: the Main Task training data, the Main task test data, and the Pilot task test data. The Main Task data are on contemporary news articles, while the Pilot Task on historical news articles. For our experiments, we focused only on the Main Task. In addition to the training and test data, we have created also a Main Task development set by excluding from the training data all the articles that composed the test data of the Italian dataset at the SemEval 2010 TempEval-2 campaign BIBREF6 . The new partition of the corpus results in the following distribution of the $<$ EVENT $>$ tag: i) 17,528 events in the training data, of which 1,207 are multi-token mentions; ii.) 301 events in the development set, of which 13 are multi-token mentions; and finally, iii.) 3,798 events in the Main task test, of which 271 are multi-token mentions.

Tables 1 and 1 report, respectively, the distribution of the events per token part-of speech (POS) and per event class. Not surprisingly, verbs are the largest annotated category, followed by nouns, adjectives, and prepositional phrases. Such a distribution reflects both a kind of “natural” distribution of the realization of events in an Indo-european language, and, at the same time, specific annotation choices. For instance, adjectives have been annotated only when in a predicative position and when introduced by a copula or a copular construction. As for the classes, OCCURRENCE and STATE represent the large majority of all events, followed by the intensional ones (I_STATE and I_ACTION), expressing some factual relationship between the target events and their arguments, and finally the others (REPORTING, ASPECTUAL, and PERCEPTION).

## System and Experiments

We adapted a publicly available Bi-LSTM network with a CRF classifier as last layer BIBREF14 . BIBREF14 demonstrated that word embeddings, among other hyper-parameters, have a major impact on the performance of the network, regardless of the specific task. On the basis of these experimental observations, we decided to investigate the impact of different Italian word embeddings for the Subtask B Main Task of the EVENTI exercise. We thus selected 5 word embeddings for Italian to initialize the network, differentiating one with respect to each other either for the representation model used (word2vec vs. GloVe; CBOW vs. skip-gram), dimensionality (300 vs. 100), or corpora used for their generation (Italian Wikipedia vs. crawled web document vs. large textual corpora or archives):

As for the other parameters, the network maintains the optimized configurations used for the event detection task for English BIBREF14 : two LSTM layers of 100 units each, Nadam optimizer, variational dropout (0.5, 0.5), with gradient normalization ( $\tau $ = 1), and batch size of 8. Character-level embeddings, learned using a Convolutional Neural Network (CNN) BIBREF22 , are concatenated with the word embedding vector to feed into the LSTM network. Final layer of the network is a CRF classifier.

Evaluation is conducted using the EVENTI evaluation framework. Standard Precision, Recall, and F1 apply for the event detection. Given that the extent of an event tag may be composed by more than one tokens, systems are evaluated both for strict match, i.e. one point only if all tokens which compose an $<$ EVENT $>$ tag are correctly identified, and relaxed match, i.e. one point for any correct overlap between the system output and the reference gold data. The classification aspect is evaluated using the F1-attribute score BIBREF7 , that captures how well a system identify both the entity (extent) and attribute (i.e. class) together.

We approached the task in a single-step by detecting and classifying event mentions at once rather than in the standard two step approach, i.e. detection first and classification on top of the detected elements. The task is formulated as a seq2seq problem, by converting the original annotation format into an BIO scheme (Beginning, Inside, Outside), with the resulting alphabet being B-class_label, I-class_label and O. Example "System and Experiments" below illustrates a simplified version of the problem for a short sentence:

input problem solution

Marco (B-STATE $|$ I-STATE $|$ ... $|$ O) O

pensa (B-STATE $|$ I-STATE $|$ ... $|$ O) B-ISTATE

di (B-STATE $|$ I-STATE $|$ ... $|$ O) O

andare (B-STATE $|$ I-STATE $|$ ... $|$ O) B-OCCUR

a (B-STATE $|$ I-STATE $|$ ... $|$ O) O

casa (B-STATE $|$ I-STATE $|$ ... $|$ O) O

. (B-STATE $|$ I-STATE $|$ ... $|$ O) O

## Results and Discussion

Results for the experiments are illustrated in Table 2 . We also report the results of the best system that participated at EVENTI Subtask B, FBK-HLT BIBREF23 . FBK-HLT is a cascade of two SVM classifiers (one for detection and one for classification) based on rich linguistic features. Figure 1 plots charts comparing F1 scores of the network initialized with each of the five embeddings against the FBK-HLT system for the event detection and classification tasks, respectively.

The results of the Bi-LSTM-CRF network are varied in both evaluation configurations. The differences are mainly due to the embeddings used to initialize the network. The best embedding configuration is Fastext-It that differentiate from all the others for the approach used for generating the embeddings. Embedding's dimensionality impacts on the performances supporting the findings in BIBREF14 , but it seems that the quantity (and variety) of data used to generate the embeddings can have a mitigating effect, as shown by the results of the DH-FBK-100 configuration (especially in the classification subtask, and in the Recall scores for the event extent subtask). Coverage of the embeddings (and consequenlty, tokenization of the dataset and the embeddings) is a further aspect to keep into account, but it seems to have a minor impact with respect to dimensionality. It turns out that BIBREF15 's embeddings are those suffering the most from out of vocabulary (OVV) tokens (2.14% and 1.06% in training, 2.77% and 1.84% in test for the word2vec model and GloVe, respectively) with respect to the others. However, they still outperform DH-FBK_100 and ILC-ItWack, whose OVV are much lower (0.73% in training and 1.12% in test for DH-FBK_100; 0.74% in training and 0.83% in test for ILC-ItWack).

The network obtains the best F1 score, both for detection (F1 of 0.880 for strict evaluation and 0.903 for relaxed evaluation with Fastext-It embeddings) and for classification (F1-class of 0.756 for strict evaluation, and 0.751 for relaxed evaluation with Fastext-It embeddings). Although FBK-HLT suffers in the classification subtask, it qualifies as a highly competitive system for the detection subtask. By observing the strict F1 scores, FBK-HLT beats three configurations (DH-FBK-100, ILC-ItWack, Berardi2015_Glove) , almost equals one (Berardi2015_w2v) , and it is outperformed only by one

(Fastext-It) . In the relaxed evaluation setting, DH-FBK-100 is the only configuration that does not beat FBK-HLT (although the difference is only 0.001 point). Nevertheless, it is remarkable to observe that FBK-HLT has a very high Precision (0.902, relaxed evaluation mode), that is overcome by only one embedding configuration, ILC-ItWack. The results also indicates that word embeddings have a major contribution on Recall, supporting observations that distributed representations have better generalization capabilities than discrete feature vectors. This is further supported by the fact that these results are obtained using a single step approach, where the network has to deal with a total of 15 possible different labels.

We further compared the outputs of the best model, i.e. Fastext-It, against FBK-HLT. As for the event detection subtask, we have adopted an event-based analysis rather than a token based one, as this will provide better insights on errors concerning multi-token events and event parts-of-speech (see Table 1 for reference). By analyzing the True Positives, we observe that the Fastext-It model has better performances than FBK-HLT with nouns (77.78% vs. 65.64%, respectively) and prepositional phrases (28.00% vs. 16.00%, respectively). Performances are very close for verbs (88.04% vs. 88.49%, respectively) and adjectives (80.50% vs. 79.66%, respectively). These results, especially those for prepositional phrases, indicates that the Bi-LSTM-CRF network structure and embeddings are also much more robust at detecting multi-tokens instances of events, and difficult realizations of events, such as nouns.

Concerning the classification, we focused on the mismatches between correctly identified events (extent layer) and class assignment. The Fastext-It model wrongly assigns the class to only 557 event tokens compared to the 729 cases for FBK-HLT. The distribution of the class errors, in terms of absolute numbers, is the same between the two systems, with the top three wrong classes being, in both cases, OCCURRENCE, I_ACTION and STATE. OCCURRENCE, not surprisingly, is the class that tends to be assigned more often by both systems, being also the most frequent. However, if FBK-HLT largely overgeneralizes OCCURRENCE (59.53% of all class errors), this corresponds to only one third of the errors (37.70%) in the Bi-LSTM-CRF network. Other notable differences concern I_ACTION (27.82% of errors for the Bi-LSTM-CRF vs. 17.28% for FBK-HLT), STATE (8.79% for the Bi-LSTM-CRF vs. 15.22% for FBK-HLT) and REPORTING (7.89% for the Bi-LSTM-CRF vs. 2.33% for FBK-HLT) classes.

## Conclusion and Future Work

This paper has investigated the application of different word embeddings for the initialization of a state-of-the-art Bi-LSTM-CRF network to solve the event detection and classification task in Italian, according to the EVENTI exercise. We obtained new state-of-the-art results using the Fastext-It embeddings, and improved the F1-class score of 6.5 points in strict evaluation mode. As for the event detection subtask, we observe a limited improvement (+1.3 points in strict F1), mainly due to gains in Recall. Such results are extremely positive as the task has been modeled in a single step approach, i.e. detection and classification at once, for the first time in Italian. Further support that embeddings have a major impact in the performance of neural architectures is provided, as the variations in performance of the Bi-LSMT-CRF models show. This is due to a combination of factors such as dimensionality, (raw) data, and the method used for generating the embeddings.

Future work should focus on the development of embeddings that move away from the basic word level, integrating extra layers of linguistic analysis (e.g. syntactic dependencies) BIBREF24 , that have proven to be very powerful for the same task in English.

## Acknowledgments

The author wants to thank all researchers and research groups who made available their word embeddings and their code. Sharing is caring.