Multi-Task Learning in Deep Neural Networks at EVALITA 2018

Andrea Cimino, Lorenzo De Mattei and Felice Dell'Orletta
Italia**NLP** Lab - *www.italianlp.it*
Instituto di Linguistica Computazionale "Antonio Zampolli" - CNR - Pisa
Dipartimento di Informatica - Università di Pisa

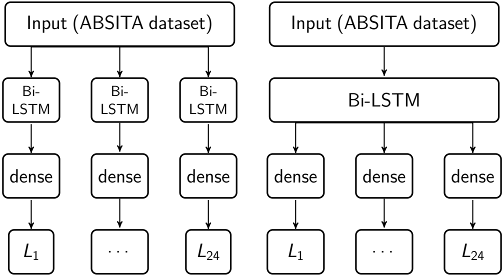
EVALITA, Turin, November 12th, 2018

Our participation at EVALITA 2018

- In EVALITA edition the majority of task were classfication (binary) tasks, among them:
  - **ABSITA** (Aspect Based Sentiment Analysis)
  - **HaSpeeDe** (Hate Speech Detection)
  - **GxG** (Gender X-Genre, author profiling in terms of gender)
  - **IronITA** (Irony and Sarcasm Detection)
- We designed and developed a general purpose system based on deep neural networks and evaluated the performance on these 4 shared tasks

In this presentation. . .

- I will describe our multi task learning our architecture
- I will report the resources and experiments we performed on the **ABSITA 2018** shared task

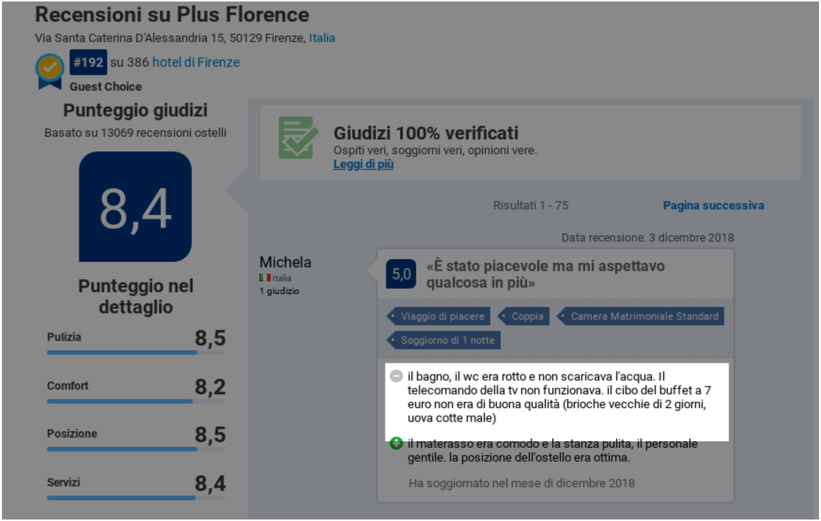


(a) STL Model (b) MTL Model

- We tackled the task as a 24 binary labels classification problems
- We resorted to a MTL architecture with the aim of:
  - Reducing the complexity of the architecture
  - Exploiting possible hidden relations among the tasks

Booking specific word embeddings

Booking specific word embeddings



ITALIAN NATURAL LANGUAGE PROCESSING LAB@ILC

www.italianlp.it

Booking specific word embeddings

www.italianlp.it

Booking specific word embeddings

- 538,835 Booking reviews scraped from the web
- More specifically:
  - 338,494 **positive** reviews
  - 200,341 **negative** reviews
- Starting from the positive and the negative reviews, we finally obtained two different word embedding lexicons.

Resources and DNN Features

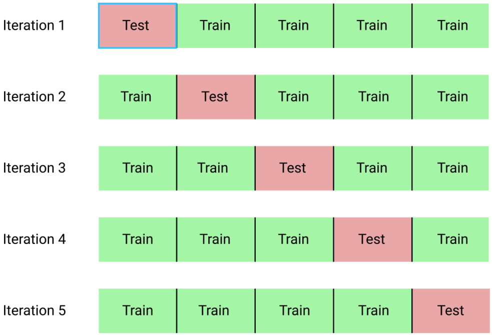
- We downloaded 50.000.000 Tweets: built 3 lexicons for neutral positive and negative tweets using seed words or emojis
  - Each TW_{POS}, TW_{NEG}, TW_{NEU} entry contains the relative frequency of the word in each lexicon
  - What we expect: $freq_{POS}(bello) > freq_{NEU}(bello) > freq_{NEG}(bello)$
  - We used this as feature in the DNN classifier
- Booking positive reviews corpus word embeddings
- Booking negative reviews corpus word embeddings
- ItWac corpus word embeddings
- Automatically postagged the dataset $\rightarrow$ POStags

Train STL and MTL

- STL: we performed three different training steps, one for each task.
- MTL: we run a shared training by iteratively optimizing at each step a loss function for each task. For the MTL the global loss function is given by the sum of the three individual loss functions.

In STL and MTL architectures, we stopped the training after 100 epochs without improvements of the loss function on the validation set, choosing the parameters with the best performances.

5-fold vote approach



- Train 5 different models, using 5 different validation sets
- Use a vote approach in classification phase
- Each validation fold follows the label distribution of the training set

Results (Development Set)

Configuration	ACD	ACP
baseline	0.313	0.197
linear SVM	0.797	0.739
STL	0.821	0.795
MTL	**0.824**	**0.804**
MTL NO K-FOLD	0.819	0.782
MTL NO BOOKING-WE	0.817	0.757

Classification results (micro f-score) of the different learning models on our development set

Results (Test Set)

Configuration	ACD	ACP
baseline	0.338	0.199
2nd best participant	0.806	0.745
linear SVM	0.772*	0.686*
STL	**0.814**	0.765
MTL	0.811*	**0.767***
MTL NO K-FOLD	0.801	0.755
MTL NO BOOKING-WE	0.808	0.753

Classification results (micro f-score) of the different learning models on the official test set.

ITALIAN NATURAL LANGUAGE PROCESSING LAB@ILC
www.italianlp.it

Results (Test Set)

Configuration	ACD	ACP
baseline	0.338	0.199
2nd best participant	0.806	0.745
linear SVM	0.772*	0.686*
STL	**0.814**	0.765
MTL	0.811*	**0.767***
MTL NO K-FOLD	0.801	0.755
MTL NO BOOKING-WE	0.808	0.753

Classification results (micro f-score) of the different learning models on the official test set.

Results (Test Set)

Configuration	ACD	ACP
baseline	0.338	0.199
2nd best participant	0.806	0.745
linear SVM	0.772*	0.686*
STL	**0.814**	0.765
MTL	0.811*	**0.767***
MTL NO K-FOLD	0.801	0.755
MTL NO BOOKING-WE	0.808	0.753

Classification results (micro f-score) of the different learning models on the official test set.

ITALIAN NATURAL LANGUAGE PROCESSING LAB@ILC
www.italianlp.it

Results (Test Set)

Configuration	ACD	ACP
baseline	0.338	0.199
2nd best participant	0.806	0.745
linear SVM	0.772*	0.686*
STL	**0.814**	0.765
MTL	0.811*	**0.767***
MTL NO K-FOLD	0.801	0.755
MTL NO BOOKING-WE	0.808	0.753

Classification results (micro f-score) of the different learning models on the official test set.

Conclusions

- We presented the first deep multi task learning system at EVALITA
- Our system ranked 1st both in the **ACD** and **ACP** subtasks
- Surprinsingly, MTL architecture did not show improvements w.r.t STL ...
  - But just 1 LSTM instead of 24, training and evaluation time extremely reduced!
- Domain specific word embeddings and the *k*-fold techinque contributed to improve classification results
- **TODO**: Incorporate relations between tasks in the loss function to further improve performances

Thanks for your attention!
Questions?