UniBO @ KIPoS: Fine-tuning the Italian "BERTology" for PoS-tagging Spoken Data

Fabio Tamburini
FICLIT - University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract

English. The use of contextualised word embeddings allowed for a relevant performance increase for almost all Natural Language Processing (NLP) applications. Recently some new models especially developed for Italian became available to scholars. This work aims at applying simple fine-tuning methods for producing high-performance solutions at the EVALITA KIPOS PoS-tagging task (Bosco et al., 2020).

Italian. *L'utilizzazione di* word embedding *contestuali ha consentito notevoli incrementi nelle performance dei sistemi automatici sviluppati per affrontare vari task nell'ambito dell'elaborazione del linguaggio naturale. Recentemente sono stati introdotti alcuni nuovi modelli sviluppati specificatamente per la lingua italiana. Lo scopo di questo lavoro è valutare se un semplice* fine-tuning *di questi modelli sia sufficiente per ottenere performance di alto livello nel task KIPOS di EVALITA 2020.*

1 Introduction

The introduction of contextualised word embeddings, starting with ELMo (Peters et al., 2018) and in particular with BERT (Devlin et al., 2019) and the subsequent BERT-inspired transformer models (Liu et al., 2019; Martin et al., 2020; Sanh et al., 2019), marked a strong revolution in Natural Language Processing (NLP), boosting the performance of almost all applications and especially those based on statistical analysis and Deep Neural Networks (DNN).

This work heavily refers to an upcoming work of the same author (Tamburini, 2020) experimenting various contextualised word embeddings for Italian to a number of different tasks and it is aimed at applying simple fine-tuning methods for producing high-performance solutions at the EVALITA KIPOS PoS-tagging task (Bosco et al., 2020; Basile et al., 2020).

2 Italian "BERTology"

The availability of various powerful computational solutions for the community allowed for the development of some BERT-derived models trained specifically on big Italian corpora of various textual types. All these models have been taken into account for our evaluation. In particular we considered those models that, at the time of writing, are the only one available for Italian:

- Multilingual BERT[1]: with the first BERT release Google developed also a multilingual model ('bert-base-multilingual-cased' – *bertMC*) that can be applied also for processing Italian texts.

- AlBERTo[2]: last year a research group from the University of Bari developed a brand new model for Italian especially devoted to Twitter texts and social media ('m-polignano-uniba/bert_uncased_L-12_H-768_A-12_italian_alb3rt0' – *alUC*) (Polignano et al., 2019). Only the uncased model is available to the community. Due to the specific training of alUC, it requires a particular pre-processing step for replacing hashtags, urls, etc. that alter the official tokenisation, rendering it not really applicable to word-based classification tasks in general texts; thus, it will be used only for

[1] https://github.com/google-research/bert
[2] https://github.com/marcopoli/AlBERTo-it

working on twitter or social media data. In any case we tested it in all considered tasks and, whenever results were reasonable, we reported them.

- GilBERTo[3]: it is a rather new CamemBERT Italian model ('idb-ita/gilberto-uncased-from-camembert' – *giUC*) trained by using the huge Italian Web corpus section of the OSCAR (Ortis Suárez et al., 2019) project. Also for GilBERTo it is available only the uncased model.

- UmBERTo[4]: the more recent model developed explicitly for Italian, as far as we know, is UmBERTo ('Musixmatch/umberto-commoncrawl-cased-v1' – *umC*). As well as GilBERTo, it has been trained by using OSCAR, but the produced model, differently from GilBERTo, is cased.

3 KIPOS 2020 PoS-tagging Task

Part-of-speech tagging is a very basic task in NLP and a lot of applications rely on precise PoS-tag assignments. Spoken data present further challenges for PoS-taggers: small datasets for system training, short training sentences, less constrained language, the massive presence of interjections, etc. are all examples of phenomena that increase the difficulties for building reliable automatic systems.

The PoS-tagging system used for our experiments is very simple and consist of a slight modification to the fine tuning script '`run_ner.py`' available with the version 2.7.0 of the Huggingface/Transformers package[5]. We did not employ any hyperparameter tuning, and, as the stopping criterion, we fixed the number of epoch to 10 and chose the UmBERTo model on the basis of the previous experience (Tamburini, 2020). After the challenge, we evaluated all the BERT-derived models in order to propose a complete overview of the available resources.

Table 1 shows the results obtained by fine tuning all the considered BERT-derived models for the Main Task. A very relevant increase in performance w.r.t. the other participants is evident looking at the results and UmBERTo is consistently the best system.

[3]https://github.com/idb-ita/GilBERTo

[4]https://github.com/musixmatchresearch/umberto

[5]https://github.com/huggingface/transformers

System	Main Task Accuracy		
	Form.	**Inform.**	**Both**
Fine-Tuning$_{\text{umC}}$	**93.49**	**91.13**	**92.26**
Fine-Tuning$_{\text{giUC}}$	92.96	89.92	91.38
Fine-Tuning$_{\text{alUC}}$	90.02	89.82	89.92
Fine-Tuning$_{\text{bertMC}}$	91.67	88.05	89.79
2nd ranked system	87.56	88.24	87.91
3rd ranked system	81.58	79.37	80.43

Table 1: PoS-tagging Accuracy for the EVALITA KIPOS 2020 benchmark for the Main Task. The Fine-Tuning$_{\text{umC}}$ has been submitted for the challenge as the system "UniBO".

We did not participate at the official challenge for the two subtasks, but we included the results of our best system also for these tasks into this report. Tables 2 and 3 show the results compared with the other two participating systems.

System	Sub-Task A Accuracy		
	Form.	**Inform.**	**Both**
Other Participant 1	**87.37**	**87.58**	**87.48**
Fine-Tuning$_{\text{umC}}$	86.47	83.16	84.75
Other participant 2	78.73	75.79	77.20

Table 2: PoS-tagging Accuracy for the EVALITA KIPOS 2020 benchmark for the Sub-Task A.

System	Sub-Task B Accuracy		
	Form.	**Inform.**	**Both**
Fine-Tuning$_{\text{umC}}$	**89.74**	**89.52**	**89.63**
Other participant 1	87.81	88.10	87.96
Other Participant 2	77.11	77.50	77.31

Table 3: PoS-tagging Accuracy for the EVALITA KIPOS 2020 benchmark for the Sub-Task B.

Again, the simple fine tuning of a BERT-derived model, namely UnBERTo, exhibits the best performance on Sub-task B. The small amount of data could probably affect the results on Sub-task A.

We collected the most frequent errors produced by the proposed system: Table 4 shows that, unexpectedly, the most frequent misclassifications involve grammatical words. The typical behaviour of the classical PoS-taggers tend to wrongly classify lexical words, namely nouns, verbs and adjectives, intermixing their classes. Apparently, on this dataset, grammatical words appear to be more complex to classify than lexical words. This be-

haviour should be investigated more appropriately by using bigger datasets and better consistency checks on the annotated data.

Formal		
#mistakes	**Gold tag**	**System tag**
19	ADP_A	ADP
16	CCONJ	ADV
12	PROPN	X
10	NOUN.LIN	X
10	ADJ	VERB
Informal		
#mistakes	**Gold tag**	**System tag**
59	PRON	SCONJ
38	ADP_A	ADP
22	ADV	CCONJ
15	NUM	DET
15	INTJ	PARA
15	CCONJ	ADV
12	NOUN	PROPN
10	VERB_PRON	VERB

Table 4: Error Analysis

4 Discussion and Conclusions

The starting idea of this work was to design the simplest DNN model for Italian PoS-tagging after the 'BERT-revolution' thanks to the recent availability of Italian BERT-derived models. Looking at the results presented in previous sections, we can certainly conclude that BERT-derived models, specifically trained on Italian texts, allow for a relevant increase in performance also when applied to spoken language by simple fine-tuning procedures. The multilingual BERT model developed by Google was not able to produce good results and should not be used when are available specific models for the studied language.

A side, and sad, consideration that emerges from this study regards the complexity of the models. All the DNN models used in this work involved very simple fine-tuning processes of some BERT-derived model. Machine learning and Deep learning changed completely the approaches to NLP solutions, but never before we were in a situation in which a single methodological approach can solve different NLP problems always establishing the state-of-the-art for that problem. Moreover, we did not apply any parameter tuning at all and fixed the early stopping criterion on 10 epochs without any optimisation. By tuning all the hyperparameters, it is reasonable we can further increase the overall performance.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

C. Bosco, S. Ballarè, M. Cerruti, E. Goria, and C. Mauri. 2020. KIPoS@EVALITA2020: Overview of the Task on KIParla Part of Speech tagging. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

L. Martin, B. Muller, P.J. Ortiz Suárez, Y. Dupont, L. Romary, E. de la Clergerie, D. Seddah, and B. Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

P.J. Ortis Suárez, B. Sagot, and L. Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in*

the Management of Large Corpora (CMLC-7), Cardiff, United Kingdom.

M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT 2018*, pages 2227–2237, New Orleans, Louisiana.

M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile. 2019. ALBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.

F. Tamburini. 2020. How "BERTology" Changed the State-of-the-Art also for Italian NLP. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Bologna, Italy.