How "BERTology" Changed the State-of-the-Art also for Italian NLP

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

Abstract

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 evaluating the impact of these models in enhancing application performance for Italian establishing the new state-of-the-art for some fundamental NLP tasks.

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, boosting the performance of almost all applications and especially those based on statistical analysis and Deep Neural Networks (DNN).

A recent study (He and Choi, 2019) tried to determine the new baselines for several NLP tasks for English fixing the new state-of-the-art for the examined tasks. This work aims at doing a similar process also for Italian. We considered a number of relevant tasks applying state-of-the-art neural models available to the community and fed them with all the contextualised word embeddings specifically developed for Italian.

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*) trained by using 200 millions tweets from 2012 to 2015 (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 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) Web-corpus project consisting of more than 11

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

billions of tokens. 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 Evaluation Tasks

Following the work of He and Choi (2019), we selected some basic tasks both for word and sentence/text classification. We mainly concentrated our efforts on tasks for which evaluation procedures were well established in the Italian community and reliable evaluation benchmark were available. We choose (a) two very basic word-classification tasks, namely part-of-speech (PoS) tagging and Named Entity Recognition (NER), (b) the dependency parsing task and (c) two very important tasks for social-media text classification, namely Sentiment Analysis (Subjectivity/Polarity/Irony classification) and Hate Speech Detection (HSD).

We mainly relied on some benchmark proposed in one of the past EVALITA evaluation challenges[5] or the Universal Dependencies (UD) project[6].

After the influential paper from (Reimers and Gurevych, 2017) it is clear to the community that reporting a single score for each DNN training session could be heavily affected by the system initialisation point and we should instead report the mean and standard deviation of various runs with the same setting in order to get a more accurate picture of the real systems performance and make more reliable comparisons between them. Thus any new result proposed in this paper is presented as the mean and standard deviation of at least 5 runs.

With regard to the dataset splitting, if a specific dataset was already split in training/validation/test set, we adopted this subdivision, while, if the dataset was split only in development and test set, we split it and used the training/validation sets for training and tuning the stopping epoch and, once fixed that parameter, we retrained the system on the entire development set maintaining the same epoch for the early stopping.

3.1 Part-of-Speech Tagging

The first task we worked on is the part-of-speech tagging. This is a very basic task in NLP and a lot of applications rely on precise PoS-tag assignments. There are various data sets available for this task taken from one of the EVALITA 2007 tasks (Tamburini, 2007) and from the UD annotated corpora.

System	EVALITA 2007
(Tamburini, 2016)	98.18
Fine-Tuning$_{\text{giUC}}$	98.75$\pm$0.04
Fine-Tuning$_{\text{bertMC}}$	98.80$\pm$0.05
Fine-Tuning$_{\text{umC}}$	**99.10**$\pm$0.04

Table 1: PoS-tagging Accuracy for the EVALITA 2007 benchmark.

	UD-ISDT v2.5	
System	**UPOS**	**XPOS**
Fine-Tuning$_{\text{giUC}}$	98.72$\pm$0.03	98.65$\pm$0.02
Fine-Tuning$_{\text{bertMC}}$	98.73$\pm$0.05	98.69$\pm$0.05
Fine-Tuning$_{\text{umC}}$	**98.78**$\pm$0.08	**98.73**$\pm$0.02

Table 2: PoS-tagging Accuracy for UD-ISDT v2.5 corpus both considering UPOS and XPOS.

	UD-PoSTW v2.5	
System	**UPOS**	**XPOS**
(Cimino and Dell'Orletta, 2016a)	93.19	-
(Basile et al., 2017)	93.34	-
Fine-Tuning$_{\text{giUC}}$	94.77$\pm$0.07	94.57$\pm$0.05
Fine-Tuning$_{\text{bertMC}}$	96.37$\pm$0.09	96.18$\pm$0.06
Fine-Tuning$_{\text{umC}}$	**97.29**$\pm$0.33	**97.27**$\pm$0.04

Table 3: PoS-tagging Accuracy for UD-PoSTWITA v2.5. N.B.: the baselines from the literature refer to the previous PoSTWITA version used in EVALITA 2016 campaign.

The best results for the EVALITA 2007 data set has been obtained by (Tamburini, 2016) using a BiLSTM-CRF system based on word2vec word embeddings enriched with morphological information. For UD corpora we considered the ISDT corpus v2.5 and PoSTWITA: there are no evaluation data in literature for the ISDT corpus while for PoSTWITA the best results were obtained by

[4]https://github.com/musixmatchresearch/umberto
[5]http://www.evalita.it
[6]https://universaldependencies.org

(Basile et al., 2017) using a BiLSTM-CRF system and by the best system at EVALITA 2016 (Cimino and Dell'Orletta, 2016a).

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[7]. We did not employ any hyperparameter tuning, the validation set has been used only for determining the stopping criterion.

Tables 1, 2 and 3 show the results obtained by fine tuning the considered BERT-derived models for this task. A very relevant increase in performance w.r.t. the literature is evident by looking at the results and UmBERTo is consistently the best system.

3.1.1 PoS-tagging on Speech Data

We participated to the EVALITA 2020 KIPOS challenge (Bosco et al., 2020) for evaluating PoS-taggers on speech data by using exactly the same tagger. In this case, we did not make any parameter tuning: we used the basic parameters and stopped the training phase after 10 epochs. After the challenge, we evaluated all the BERT-derived models in order to propose a complete overview of the available resources.

Tables 4 show 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 again the best system.

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. Table 5 shows the results compared with the other two participating systems. Again, the simple fine tuning of a BERT-derived model, namely UnBERTo, exhibits the best performance on Sub-task B. The scarcity of data could probably affect the results on Subtask A.

3.2 Named Entity Recognition

The second task we considered is Named Entity Recognition. For system evaluation we relied on the nice evaluation benchmark used in the EVALITA 2009 campaign (Bartalesi Lenzi et al., 2009). The best results gathered from literature are due to (Basile et al., 2017) that used a BiLSTM-CRF system and to the best system at the EVALITA 2009 campaign (Zanoli et al., 2009).

For this task we used exactly the same script of the previous task, being both tasks simple word-classification tasks, and did not apply any hyperparameter tuning at all, fixing a priori the number of epoch to 10.

Table 6 outlines the obtained results. Again a simple fine tuning of BERT-derived models is enough powerful to guarantee relevant increases of performance with respect to the previous literature and, again, UmBERTo resulted the model producing the best performance.

[7] https://huggingface.co/transformers/

System	Main Task Accuracy		
	Form.	**Inform.**	**Both**
(Izzi and Ferilli, 2020)	81.58	79.37	80.43
(Proisl and Lapesa, 2020)	87.56	88.24	87.91
Fine-Tuning$_{\text{bertMC}}$	91.67	88.05	89.79
Fine-Tuning$_{\text{alUC}}$	90.02	89.82	89.92
Fine-Tuning$_{\text{giUC}}$	92.96	89.92	91.38
Fine-Tuning$_{\text{umC}}$	**93.49**	**91.13**	**92.26**

Table 4: PoS-tagging Accuracy for the EVALITA KIPOS 2020 benchmark for the Main Task.

System	Sub-Task A Accuracy		
	Form.	**Inform.**	**Both**
(Izzi and Ferilli, 2020)	78.73	75.79	77.20
Fine-Tuning$_{\text{umC}}$	86.47	83.16	84.75
(Proisl and Lapesa, 2020)	**87.37**	**87.58**	**87.48**
	Sub-Task B Accuracy		
(Izzi and Ferilli, 2020)	77.11	77.50	77.31
(Proisl and Lapesa, 2020)	87.81	88.10	87.96
Fine-Tuning$_{\text{umC}}$	**89.74**	**89.52**	**89.63**

Table 5: PoS-tagging Accuracy for the EVALITA KIPOS 2020 benchmark for the two Sub-Tasks A and B.

System	Macro F1
(Zanoli et al., 2009)	82.00
(Basile et al., 2017)	82.34
Fine-Tuning$_{\text{giUC}}$	82.37$\pm$0.31
Fine-Tuning$_{\text{bertMC}}$	85.07$\pm$0.29
Fine-Tuning$_{\text{umC}}$	**87.66**$\pm$0.44

Table 6: Macro-averaged F1-score for the various systems when evaluated with the EVALITA 2009 NER benchmark.

3.3 Parsing Universal Dependencies

Parsing is one of the most important tasks in NLP and the recent advances due to DNN and contextualised distributed representations allowed for large performance improvements.

Universal Dependencies project is the reference repository for standardised treebanks in various languages, thus it seemed natural to gather evaluation benchmarks from that project. As for PoS-tagging, we used two treebanks from UD v2.5, namely ISDT and PoSTWITA.

The recent work from Antonelli and Tamburini (2019) examined all the DNN parsers available at the time re-training them on some Italian dataset. In particular they showed that the neural parser from Dozat and Manning (2017) (version 1.0) was the parser exhibiting the best performance on UD-ISDT v2.1. Giving that experience, we included in our new experiments the last version (v3.0) of this parser[8] considering it as a strong baseline for this task. The word embeddings we used for these experiments were the same used in (Antonelli and Tamburini, 2019) and are computed using the ItWaC corpus (Baroni et al., 2009) and word2vec (Mikolov et al., 2013a,b).

Very recently, a new work from Vacareanu et al. (2020) showed that we can efficiently compute dependency parsing structures by treating this task as a double fine tuning task over a BERT-derived model, the first for determining the attachments and the second the edge labels, getting state-of-the-art performance. Actually, the fine-tuning DNN is more complex than in the previous tasks, consisting of a bidirectional LSTM followed by some dense layers.

We applied their method and code (PaT) for our parsing experiments using the greedy cycle removal option. We changed text case depending on the BERT-derived model case used in a specific experiment. Tables 7 and 8 show the results for all the parsing experiments.

Considering the best results obtained by the Dozat and Manning (2017) parser and those presented in (Antonelli and Tamburini, 2019), we observe a relevant increase in performance due mainly to GilBERTo and UmBERTo.

3.4 Sentiment Analysis

Three main text-classification tasks are comprised in the 'Sentiment Analysis' umbrella: Subjectivity, Polarity and Irony detection. Thanks to the EVALITA SENTIPOLC 2016 evaluation we could rely on a complete dataset annotated with respect to all the three tasks.

Given the specific nature of dataset texts, namely tweet texts, we adopted the particular preprocessing procedure introduced by AlBERTo and all the other parameters were kept as in (Polignano et al., 2019) for comparability; the only difference regards the training batch size that was 512 on TPU in the original paper and we had to use gradient accumulation on GPU (batch size = 32 and accumulation steps = 16) to avoid memory problems. Given the small size of the dataset and the high variability of the various results, for these tasks we decided to make 10 runs instead of 5.

[8] https://github.com/tdozat/Parser-v3

System	UD-ISDT v2.5	
	UAS	**LAS**
(Antonelli and Tamburini, 2019)	94.00	92.48
$\text{PaT}_{\text{bertMC}}$	94.12±0.26	91.74±0.23
(Dozat and Manning, 2018)	94.53±0.14	93.35±0.18
PaT_{umC}	95.32±0.14	93.39±0.26
PaT_{giUC}	**95.52**±0.18	**93.59**±0.28

Table 7: Parsing Un/Labeled Attachment Score (UAS/LAS) for UD-ISDT v2.5.

System	UD-PoSTW v2.5	
	UAS	**LAS**
$\text{PaT}_{\text{bertMC}}$	87.97±0.20	82.03±0.24
(Dozat and Manning, 2018)	88.04±0.13	**84.08**±0.10
PaT_{alUC}	88.19±0.32	82.66±0.38
PaT_{umC}	89.16±0.17	83.25±0.23
PaT_{giUC}	**89.29**±0.27	83.66±0.22

Table 8: Parsing Un/Labeled Attachment Score (UAS/LAS) for UD-PoSTWITA v2.5.

System	Macro F1
$\text{TensorFlow+TPU}_{\text{alUC}}$	72.23*
$\text{Fine-Tuning}_{\text{bertMC}}$	72.92±0.86
(Castellucci et al., 2016)	74.44
$\text{Fine-Tuning}_{\text{alUC}}$	75.83±0.63
$\text{Fine-Tuning}_{\text{umC}}$	77.14±0.78
$\text{Fine-Tuning}_{\text{giUC}}$	**77.58**±1.20
(Polignano et al., 2019) (alUC)	**79.06***

Table 9: Subjectivity detection macro F1-score for EVALITA SENTIPOLC 2016. * results that we were not able to reproduce using the same code.

System	Macro F1
Fine-Tuning$_{\text{bertMC}}$	65.38$\pm$1.65
(Cimino and Dell'Orletta, 2016b)	66.38
TensorFlow+TPU$_{\text{alUC}}$	71.59*
(Polignano et al., 2019) (alUC)	72.23*
Fine-Tuning$_{\text{alUC}}$	72.60$\pm$1.38
Fine-Tuning$_{\text{umC}}$	72.74$\pm$0.88
Fine-Tuning$_{\text{giUC}}$	**74.75**$\pm$0.94

Table 10: Polarity detection macro F1-score over 4 classes for EVALITA SENTIPOLC 2016. * results that we were not able to reproduce using the same code.

System	Macro F1
Fine-Tuning$_{\text{bertMC}}$	52.17$\pm$1.55
(Di Rosa and Durante, 2016)	54.12
Fine-Tuning$_{\text{umC}}$	55.65$\pm$3.09
Fine-Tuning$_{\text{alUC}}$	56.80$\pm$1.92
TensorFlow+TPU$_{\text{alUC}}$	57.21*
Fine-Tuning$_{\text{giUC}}$	**60.60**$\pm$1.45
(Polignano et al., 2019) (alUC)	**60.90***

Table 11: Irony detection Macro F1-score for EVALITA SENTIPOLC 2016 dataset. * results that we were not able to reproduce using the same code.

We slighly modified the script '`run_glue.py`' from the version 2.7.0 of the Huggingface/Transformers package considering the three tasks as a BERT-derived model fine-tuning for text classification tasks respectively with 2, 4 and 2 classes.

Tables 9, 10 and 11 present the obtained results. We have to say that we had a lot of problems in reproducing the results in Polignano et al. (2019), both by using our script and also by using the original TPU-based script on Google Colab. In the cited tables, you can find the original results and the ones produced by us using the same script and setting marked by an asterisk (TensorFlow+TPU$_{\text{alUC}}$).

3.5 Hate Speech Detection

Hate Speech on social media has become a relevant problem in recent years and the automatic detection of such messages got a great importance in NLP.

Thanks to the dataset produced by Bosco et al. (2018) we had the possibility to test the same text classification procedures we used for Sentiment Analysis also for this task both on Facebook and Twitter data. Table 12 shows the results we obtained comparing them with the best system at the EVALITA 2018 HaSpeDe campaign (Cimino et al., 2018). GilBERTo exhibit the best performance on both subtasks.

	Macro F1	
System	FB	TW
Fine-Tuning$_{\text{bertMC}}$	77.62$\pm$0.46	76.07$\pm$0.78
(Cimino et al., 2018)	82.88	79.93
Fine-Tuning$_{\text{umC}}$	83.55$\pm$0.40	80.28$\pm$0.55
Fine-Tuning$_{\text{alUC}}$	84.23$\pm$0.37	79.00$\pm$0.84
Fine-Tuning$_{\text{giUC}}$	**84.36**$\pm$0.69	**80.86**$\pm$0.46

Table 12: Macro F1-score for the HaSpeDe EVALITA 2018 Facebook (FB) and Twitter (TW) datasets.

4 Discussion and Conclusions

The starting idea of this work was to derive the new state-of-the-art for some NLP tasks for Italian after the 'BERT-revolution' thanks to the recent availability of Italian BERT-derived models. Looking at the results presented in previous sections for some very important tasks, we can certainly conclude that BERT-derived models, specifically trained on Italian texts, allow for a large increase in performance also for some important Italian NLP tasks. On the contrary, 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 for the various tasks 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. And we did not apply any parameter tuning at all! The only optimisation regards the early stopping definition on validation set. By tuning all the hyperparameters, it is reasonable we can further increase the overall performance.

For the future, it would be interesting to eval-

uate end-to-end systems, for example for solving PoS-tagging + Parsing and PoS-tagging + NER by using the BERT-derived model fine tuning code and PaT for both end-to-end tasks.

A lot of scholars are working in studying new transformer-based models or training the most promising ones on different languages; there are brand new Italian models that were made available very recently not yet included into our evaluations like the one produced by Stefan Schweter at CIS, LMU Munich[9]; it would be interesting to insert them into our tests.

Acknowledgments

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

References

O. Antonelli and F. Tamburini. 2019. State-of-the-art Italian dependency parsers based on neural and ensemble systems. *Italian Journal of Computational Linguistics*, 5(1):33–55.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

V. Bartalesi Lenzi, M. Speranza, and R. Sprugnoli. 2009. EVALITA 2009 The Entity Recognition Task. In *Proceedings of the EVALITA 2009 Workshop*, Reggio Emilia, Italy.

P. Basile, G. Semeraro, and P. Cassotti. 2017. Bidirectional LSTM-CNNs-CRF for Italian Sequence Labeling. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, pages 18–23, Roma, Italy.

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.

C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In *In Proc. of the EVALITA 2018 Workshop*, Torino, Italy.

[9]https://github.com/stefan-it/fine-tuned-berts-seq

G. Castellucci, D. Croce, and R. Basili. 2016. Context–aware Convolutional Neural Networks for Twitter Sentiment Analysis in Italian. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italy.

A. Cimino, L. De Mattei, and F. Dell'Orletta. 2018. Multi-task Learning in Deep Neural Networks at EVALITA 2018. In *In Proc. of the EVALITA 2018 Workshop*, Torino, Italy.

A. Cimino and F. Dell'Orletta. 2016a. Building the state-of-the-art in POS tagging of Italian Tweets. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italy.

A. Cimino and F. Dell'Orletta. 2016b. Tandem LSTM-SVM Approach for Sentiment Analysis. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italy.

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.

E. Di Rosa and A. Durante. 2016. Tweet2Check evaluation at Evalita Sentipolc 2016. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italy.

T. Dozat and C.D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *Proceedings of the 2017 International Conference on Learning Representations*.

T. Dozat and C.D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 484–490, Melbourne, Australia.

H. He and J.D. Choi. 2019. Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with BERT. In *The Thirty-Third International Flairs*

Conference, AAAI Publications, pages 228–233.

G.L. Izzi and S. Ferilli. 2020. A hybrid approach for part-of-speech tagging. In *Proceedings of the Seventh International Workshop EVALITA 2020*.

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.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proc. of Workshop at ICLR*.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

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.

T. Proisl and G. Lapesa. 2020. Klumsy: Experiments on part-of-speech tagging of spoken italian. In *Proceedings of the Seventh International Workshop EVALITA 2020*.

N. Reimers and I. Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. ACL.

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. 2007. EVALITA 2007: the Part-of-Speech Tagging Task. *Intelligenza Artificiale*, IV(2):4–7.

F. Tamburini. 2016. (Better than) State-of-the-Art PoS-tagging for Italian Texts. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 280–284, Napoli, Italy.

R. Vacareanu, G.C. Gouveia Barbosa, M.A. Valenzuela-Escárcega, and M. Surdeanu. 2020. Parsing as tagging. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5225–5231, Marseille, France. ELRA.

R. Zanoli, E. Pianta, and C. Giuliano. 2009. Named entity recognition through redundancy driven classifiers. In *Proceedings of the Workshop EVALITA 2009*, Reggio Emilia, Italy.