App2Check @ ATE_ABSITA 2020: Aspect Term Extraction and Aspect-based Sentiment Analysis

Emanuele Di Rosa
Chief Technology Officer
emanuele.dirosa
@app2check.com

Alberto Durante
Research Scientist
alberto.durante
@app2check.com

Abstract

In this paper we describe and present the results of the system we specifically developed and submitted for our participation to the ATE_ABSITA 2020 evaluation campaign on the Aspect Term Extraction (ATE), Aspect-based Sentiment Analysis (ABSA), and Sentiment Analysis (SA) tasks. The official results show that App2Check ranks first in all of the three tasks, reaching a F1 score which is 0.14236 higher than the second best system in the ATE task and 0.11943 higher in the ABSA task; it shows a Root-Mean-Square Error (RMSE) that is 0.13075 lower than the second classified in the SA task.

1 Introduction

User reviews are becoming more important for all consumer-oriented industries. Thanks to the expansion of a review culture, collecting and sharing a feedback from a buyer of a product/service can both help the seller to improve and other customers who can take advantage of the reviews for their purchase decisions. However, having automatic tools to process reviews and extract useful insights to analysts, especially where large amounts of reviews are available, becomes relevant for any consumer-oriented industry.

Aspect-Term Extraction and Aspect-Based Sentiment Analysis tasks are, respectively, focused on the extraction of the main aspects in a sentence and to assign a specific sentiment to each of them. These are essential tools to understand the reasons behind the success or the failure of a product or service, or anyway that allow to take actions, finalized to to improve the customer perception. The former helps analysts to go beyond the traditional "word cloud" that is available in most of text analytic tools and that focuses just on the most recurrent words in a collection. Aspect-Term Extraction, similarly to the Named-Entity Recognition task, detects a sequence of word tokens that conceptually identify an "aspect" of the sentence. The Sentiment Analysis task maintains its importance on a higher level, where it can substitute user rating, which can be sometimes incoherent to the opinions expressed in the review text. Anyway, it represents just the overall polarity of an opinion, which is very often the result of different polarities on multiple aspects. The assignment of a specific and, in general, different polarity to each aspect in the sentence, leads to the ABSA task, which is highly dependent on the ATE task, but can take advantage of the learning obtained by an SA model. In the last few years, deep learning-based models proved to be the best technical approach for natural language processing and understanding and are very promising also for the ATE, SA and ABSA tasks.

In this paper, we present the system that we specifically developed and submitted for our participation to the ATE_ABSITA 2020 evaluation campaign (De Mattei et al., 2018), which is part of EVALITA 2020 (Basile et al., 2020), on the Aspect Term Extraction (ATE), Aspect-based Sentiment Analysis (ABSA), and Sentiment Analysis (SA) tasks. To this aim, we decided to focus just on deep learning-based approaches to train a specific model for each task. More specifically, we take advantage of the most recent approach in which pre-trained language models, largely recognized as bringing NLP to a new era (Qiu et al., 2020), are used as the main component for the 3 tasks. In particular, about the ATE task, in order to select the best performing pre-trained models to use for our submission, we performed an extensive experimental analysis and comparison. The experimen-

tal evaluation shows some interesting and unpredictable results, discussed in section 2, which also represent an added value of this paper. In fact, we can summarize that in the dev set:

1. the NER fine-tuned model shows lower performance than general-purpose pre-trained models without a specific NER fine-tuning

2. a language specific, Italian-native model shows a lower performance than multilingual models fine-tuned on Italian in the specific downstream tasks

3. the biggest and most recent, multilingual XLM-Roberta model shows the best performance when fine-tuned on the downstream tasks

While the last result, related to the fact that bigger models –in terms of number of parameters– are more effective than smaller models, is quite common (with the exception of distilled models) and known in literature (see also the recent GPT3 vs GPT2 comparison (Brown et al., 2020)), the first two results are quite surprising. In fact, we expected that the multilingual model specifically fine-tuned on the NER task on another language could take advantage of a previous training in another language as shown in (Pires et al., 2019). Moreover, the native Italian pre-trained language model GilBERTo (based on Facebook RoBERTa architecture (Liu et al., 2019) and CamemBERT text tokenization approach (Martin et al., 2020)) later fine-tuned on the NER task with italian training set, shows a performance that is 4% lower than the XML-Roberta multilingual pre-trained model later trained on a NER training set in Italian.

About the SA task, we take advantage of a previously trained predictive model we had at App2Check, an evolution of the one presented in (Di Rosa and Durante, 2017), which is now based on the Multilingual BERT model and later fine-tuned on a 1 to 5 sentiment scale on a big amount of product reviews. This model has been additionally trained on the training set of the competition in order to have a domain-specific training. For the SA task, we decided to not perform any additional experimental comparison with other pre-trained models. Finally, about the ABSA task, we created a special encoding to map the output of our available SA model in order to be additionally fine-tuned on the ABSA training set of the competition: this helped to take advantage of a transfer learning from the SA task to the ABSA task.

This paper is structured as follows: in sections 2, 3 and 4, we describe each of the three tasks of the competition, the details of our training, system implementation and present the results in both the dev set and the competition results. Finally, we show the conclusions in section 5.

2 Aspect-Term Extraction

Aspect Term Extraction (ATE) is the task of identifying an "aspect" in a text without knowing a priori the list of aspects that contains it. According to the literature definition, a term/phrase is considered as an aspect when it co-occurs with "opinion words" that indicate a sentiment polarity on it.

Our approach has been to consider the ATE task as a Named Entity Recognition task (NER) and fine-tune already existing pre-trained language models on the NER task, by using the training set of the competition. More specifically, we decided to investigate four different classes of models:

1. Native Italian pre-trained language models, with no specific NER fine-tuning

2. Multilingual pre-trained language model, with no specific NER fine-tuning

3. Native Italian pre-trained language models, with a specific NER fine-tuning

4. Multilingual pre-trained language model, with a specific NER fine-tuning

To implement all of these approaches, we based on the Hugging Face transformers library (Wolf et al., 2019) and, in order to simplify our work, we looked for pre-trained models made available publicly by the Hugging Face. With the exception of item 3, for which we could not find any publicly available model in the HuggingFace models list, we considered more than one state-of-the-art model for each type of encoding that we further trained/fine-tuned on the competition training set.

For type 1, we considered dbmdz/bert-base-italian-xxl-uncased[1] and GilBERTo[2]. For type 2, we considered two implementations of RoBERTa:

[1] https://github.com/dbmdz/berts

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

xml-roberta-large[3] (Conneau et al., 2020), xml-roberta-base[4] (Liu et al., 2019), and multilingual BERT[5] (Pires et al., 2019). We wanted to try xml-roberta-large with a 512 maximum sequence length, but an out of memory exception prevented us for using it. For type 4 we considered wietsedv/bert-base-multilingual-cased-finetuned-conll2002-ner[6].

K	Len Model	Ep	F1-T	F1-D
1	512 B-BERT ita unc.	11	0.961	0.663
1	512 GilBERTo unc.	10	0.941	0.697
1	512 GilBERTo unc.	15	0.973	0.6700
2	512 B-xlmRoBERTa	8	0.981	0.687
2	256 L-xlmRoBERTa	12	0.965	0.728
2	256 **L-xlmRoBERTa**	15	**0.980**	**0.708**
2	512 B-mBERT	20	0.991	0.679
4	512 B-mBERT NER	30	0.910	0.657
4	512 B-mBERT NER	45	0.965	0.623

Table 1: Aspect-Term Extraction performance on development set.

All models have been trained on a cloud platform using an Nvidia Tesla P100-PCIE as GPU accelerator. In Table 1 we show the results obtained by the models on the training and development set, highlighting in bold the model chosen for the competition.

The value in column K, Len and Ep are associated respectively to the kind of pre-trained model used, the maximum sequence length used in the training and the number of epochs of the training. The F1-T and F1-D columns contains the F1-scores on training set and development set. For each model, the prefixes *L* and *B* indicates whether the base or large version has been used; if an uncased version of the pre-trained model has been used, the model name is labeled with *unc.*.

The Italian Base Bert and GilBERTo approaches, both of class 1, show similar results on both training and development set. Interestingly, on the development set, the multilingual Base Bert model in class 2 shows very similar results to the best model in class 1 which is specifically trained on Italian.

The xlm RoBERTa Large multilingual model shows a F1-score on the development set that is higher than the Base version of the same model, even if they show almost the same performance on the training set. The model in class 4, multilingual Bert Base specifically trained on the NER task, shows the worst performance on the development set, even if trained with a much higher number of epochs.

Thanks to the F1 score reached on the development set, the xlm RoBERTa Large multilingual model has been chosen as our competition model, so it has been further trained on the development set and tested on the competition test set.

Pos.	Name	F1 score
1	App2Check	0.68222
2	ghostwriter19	0.53986
3	SentNa	0.34027
4	*Baseline*	*0.2556*

Table 2: Aspect-Term Extraction on the test set of the competition.

In Table 2 we show the official results of the Aspect-Term Extraction task in (De Mattei et al., 2018). App2Check model ranked first with a F1 score that is 0.14236 higher than the second best system.

3 Sentiment Analysis

The SA task is about the detection of the opinion expressed in a text review. According to the typical user rating, which is here used as the reference value for the polarity, the score is defined on a five-value scale from 1 (very negative) to 5 (very positive).

About our implementation for this task, we took advantage of a previously trained predictive model we had at App2Check. It is an evolution of the one presented in (Di Rosa and Durante, 2017), which is now based on the Multilingual BERT model based on 104 languages and 110M parameters, and later fine-tuned on a 1 to 5 sentiment scale on a big amount of product reviews. This model has been additionally trained on the training set of the competition in order to have a domain-specific training. We decided to not perform any additional experimental comparison with other pre-trained models, since it has been already compared with other approaches in the past and also because of the little time at our disposal.

In Table 3 we show the results of the competition for the Sentiment Analysis task. The root-

[3]https://huggingface.co/xlm-roberta-large
[4]https://huggingface.co/xlm-roberta-base
[5]bert-base-multilingual-cased
[6]https://github.com/chambliss/Multilingual_NER

Pos	Name	RMSE
1	App2Check	0.66458
2	SentNa	0.79533
3	ghostwriter19	0.81394
4	*Baseline-AVG score*	*0.10040*
5	*Baseline-AlBERTo*	*0.10806*
6	*Baseline-Freq score*	*0.12800*

Table 3: Sentiment Analysis on the test set of the competition.

mean-square error of App2Check is 0.13073 lower than the error of the second best system, ranking in first position.

4 Aspect-Based Sentiment Analysis

The Aspect-Based Sentiment Analysis task is an extension of both the ATE and the SA tasks. In fact, the aim of the Aspect-Based Sentiment Analysis task is to detect the sentiment polarity associated to each aspect extracted, thanks to the ATE task discussed in Section 2. The possible polarity values are:

Polarity	Value
neutral	[0,0]
positive	[1,0]
negative	[0,1]
mixed	[1,1]

Similarly to what we have done with the Aspect Category Polarity task at ABSITA 2018 (Di Rosa and Durante, 2018), we assumed that the sentiment score of every aspect detected in Section 2 is the one associated to the portion of text in which it is contained. In order to do so, we split portions of the review using strong punctuation marks and some conjunctions (especially the ones leading to sentiment inversion). For example, in the case of:

> *Ottimo prodotto di marca, la qualità é veramente notevole. Non è molto capiente ma si può prendere un'altra versione. È provvisto di una tasca piccola davanti e quella grande*[7]

The aspect *capiente*[8] has the same polarity score as *Non è molto capiente*, while the aspect *qualità*[9] has the same polarity score as *Ottimo prodotto di marca, la qualità é veramente notevole*.

The same assumption has been applied to the training set: the polarity of each portion of a review has been associated to the contained aspect. If a portion of a review does not contain any aspect, it has been ignored.

The submitted ABSA system has been based on a single sentiment classification model, rather than two binary models for positive and negative polarities. The final model is a four-class retraining of the sentiment model presented in section 3 which has been originally trained on user reviews with five levels (strong positive, positive, mixed/neutral, negative, strong negative) using multilingual BERT (Pires et al., 2019). In this way, we take advantage of some transfer learning about positive, negative and neutral sentiment learned on reviews.

Pos.	Name	F1 score
1	App2Check	0.61878
2	ghostwriter19	0.49935
3	SentNa	0.28632
4	*Baseline*	*0.20000*

Table 4: Aspect-Based Sentiment Analysis on the test set of the competition.

In Table 4 we show the results of the Aspect-Based Sentiment Analysis of the competition. App2Check system is in first position, with a F1 score that is 0.11943 higher than the second best system.

5 Conclusions

In this paper we described the approach we followed and the models we built for our participation to the ATE_ABSITA 2020 competition. We also presented the experimental evaluation we made in the context of our model selection process in the development set and show interesting results: *(i)* the NER fine-tuned model shows lower performance than general-purpose pre-trained models without a specific NER fine-tuning; *(ii)* a language specific, Italian-native model shows a lower performance than multilingual models fine-tuned on Italian in the specific downstream tasks; *(iii)* the biggest and most recent, multilingual XLM-Roberta model shows the best performance when fine-tuned on the downstream tasks. We also showed that our App2Check

[7]Translation: *Great branded product, the quality is truly remarkable. It is not very capacious but you can get another version. It has a small front pocket and a large one*

[8]Translation: *capacious*

[9]Translation: *quality*

system scored first in all of the three tasks of the competition, reaching a F1 score which is 0.14236 higher than the second best system in the ATE task and 0.11943 higher in the ABSA task; in the SA task, our system shows a Root-Mean-Square Error (RMSE) that is 0.13075 lower than the second classified.

References

Basile, Valerio and Croce, Danilo and Di Maro, Maria and Passaro, Lucia C. 2020 *EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020) CEUR.org

Lorenzo de Mattei, Graziella de Martino, Andrea Iovine, Alessio Miaschi, Marco Polignano, and Giulia Rambelli. 2020 *ATE_ABSITA@EVALITA2020: Overview of the Aspect Term Extraction and Aspect-based Sentiment Analysis Task.* Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020). CEUR.org

Emanuele Di Rosa and Alberto Durante 2018 *Aspect-based Sentiment Analysis: X2Check at ABSITA 2018* Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018

Emanuele Di Rosa and Alberto Durante. 2017. *Evaluating Industrial and Research Sentiment Analysis Engines on Multiple Sources* in Proc. of AI*IA 2017 Advances in Artificial Intelligence - International Conference of the Italian Association for Artificial Intelligence, Bari, Italy, November 14-17, 2017, pp. 141-155.

Yinhan Liu and Myle Ott and Naman Goyal and Jingfei Du and Mandar Joshi and Danqi Chen and Omer Levy and Mike Lewis and Luke Zettlemoyer and Veselin Stoyanov 2019 *RoBERTa: A Robustly Optimized BERT Pretraining Approach* CoRR, abs/1907.11692

Alexis Conneau and Kartikay Khandelwal and Naman Goyal and Vishrav Chaudhary and Guillaume Wenzek and Francisco Guzmán and Edouard Grave and Myle Ott and Luke Zettlemoyer and Veselin Stoyanov 2020 *Unsupervised Cross-lingual Representation Learning at Scale* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020

Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, Benoît Sagot *CamemBERT: a Tasty French Language Model.* ACL 2020: 7203-7219

Xipeng Qiu and Tianxiang Sun and Yige Xu and Yunfan Shao and Ning Dai and Xuanjing Huang 2020 *Pre-trained Models for Natural Language Processing: A Survey* 2003.08271, arXiv, https://arxiv.org/abs/2003.08271

Tom B. Brown and Benjamin Mann and Nick Ryder and Melanie Subbiah and Jared Kaplan and Prafulla Dhariwal and Arvind Neelakantan and Pranav Shyam and Girish Sastry and Amanda Askell and Sandhini Agarwal and Ariel Herbert-Voss and Gretchen Krueger and Tom Henighan and Rewon Child and Aditya Ramesh and Daniel M. Ziegler and Jeffrey Wu and Clemens Winter and Christopher Hesse and Mark Chen and Eric Sigler and Mateusz Litwin and Scott Gray and Benjamin Chess and Jack Clark and Christopher Berner and Sam McCandlish and Alec Radford and Ilya Sutskever and Dario Amodei 2020 *Language Models are Few-Shot Learners* CoRR 2020 https://arxiv.org/abs/2005.14165

Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and Rémi Louf and Morgan Funtowicz and Joe Davison and Sam Shleifer and Patrick von Platen and Clara Ma and Yacine Jernite and Julien Plu and Canwen Xu and Teven Le Scao and Sylvain Gugger and Mariama Drame and Quentin Lhoest and Alexander M. Rush 2019. *Transformers: State-of-the-art natural language processing*. arXiv preprint arXiv:1910.03771.

Telmo Pires and Eva Schlinger and Dan Garrette 2019 *How multilingual is Multilingual BERT?* CoRR 2019 http://arxiv.org/abs/1906.01502

Basile, Valerio and Croce, Danilo and Di Maro, Maria and Passaro, Lucia C. 2020. *EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. In Online Proceedings of Evalita 2020 Publisher: CEUR.org Editor: Basile, Valerio and Croce, Danilo and Di Maro, Maria and Passaro, Lucia C.