Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation

Valentin Barriere, Alexandra Balahur

European Commission's Joint Research Center - Ispra

COLING2020

Main Principle

- We propose the use a multilingual pre-trained transformer instead of a monolingual one, so that it is possible to:
  - Adapt the model to the task by pre-training it over a huge annotated dataset of tweets in English
  - Adapt the model to other languages with a data-augmentation technique using automatic translation

Machine Translation for Data-Augmentation

- We proceed to data-augmentation by translating all the tweets from their native language to the 4 other languages used for testing.
- The translations from the source language to the 4 other languages were made by the automatic translation tool of the European Commission.

Lang.	Tweet
English	I'd rather dump gasoline all over myself and run into a burning building than use Excel.
French	Je préférerais jeter de l'essence partout et tomber dans un immeuble en feu plutôt que d'utiliser Excel.
German	Ich würde lieber Benzin auf mich werfen und in ein brennendes Gebäude laufen, als Excel zu benutzen.
Spanish	Prefiero tirar gasolina sobre mí mismo y correr hacia un edificio en llamas que usar Excel.
Italian	Preferirei buttarmi la benzina addosso e correre in un edificio in fiamme piuttosto che usare Excel.

Table: Examples of automatically translated tweets (original language in bold)

Tweets Sentiment Analysis Datasets in 5 Languages

- We trained our models over 10 datasets and tested them over five different test sets in five languages: French, English, German, Spanish and Italian.
- This makes a total of 339,215 training examples when using data-augmentation with automatic translation.

Dataset	Language	Train	Dev	Test	All
SB-10k	German	4925	330	1315	6570
TASS-2019 TASS-2018	Spanish	2133	506	581	3220
DEFT-2015	French	6489	407	2938	9427
Sentipolc-16	Italian	6534	436	1964	8934
SemEval-2017 SemEval-2013 SemEval-2014 SemEval-2015 SemEval-2016	English	47762	2000	12284	62046

Table: Datasets used in our experiments

Base models

We use as classifiers:

- XLM-R [2] as a multilingual model
- Its monolingual counterparts CamemBERT [4] for French and RoBERTa [3] for English.
- AlBERTo [6] (BERT initialization) for Italian.

Results -- Table

Language	Model	Using English	D-A	Rec_{avg}	$F1_{mac}$	$F1_{PN}$
English	[1] (winner SemEval-2017)	✓	✗	68.1	∅	68.5
	[5] (SOTA)	✓	✗	**73.2**	∅	**72.8**
	Monolingual	✓	✗	**72.8**	**71.7**	**72.3**
	Multilingual	✓	✗	71.9	70.0	70.3
		✓	✓	71.6	69.3	70.2
German	Multilingual	✗	✗	72.6	73.9	67.1
		✓	✗	74.1	**74.8**	**68.7**
		✓	✓	**74.2**	74.7	68.5
Spanish	Multilingual	✗	✗	63.5	63.2	72.7
		✓	✗	68.3	68.1	76.0
		✓	✓	**69.8**	**69.6**	**78.2**
French	Monolingual	✗	✗	72.9	72.8	71.6
	Multilingual	✗	✗	72.5	72.4	71.0
		✓	✗	73.8	73.7	72.2
		✓	✓	**74.4**	**74.5**	**72.8**
Italian	Monolingual	✗	✗	66.3	66.4	61.7
	Multilingual	✗	✗	63.0	60.7	55.3
		✓	✗	67.1	64.4	60.2
		✓	✓	**68.1**	**66.1**	**62.0**
All (non English)	Multilingual	✗	✗	68.0	67.6	66.6
		✓	✗	70.8	70.3	69.3
		✓	✓	**71.6**	**71.2**	**70.4**

Table: Results of the different configurations. All the models were originally pre-trained over general text data.

Results -- Comments

- Using English tweets to pre-train improves the results of the multilingual model.
- Data-Augmentation using Machine Translation allows once again to reach higher performances.
- The English monolingual model stays the most competitive.

Results -- Analysis

- Pre-training a multilingual model over English is a good option with a small target language training set (less than 6500).
- If there is enough of available data, it is better to use a monolingual model.
- Data-augmentation improves slightly the results for almost every language in different proportions. Our intuition is that the improvements follow the performances of the MT system.
- The utilization of English external data and data-augmentation allows to obtain better performances than the monolingual models for French and Italian.

Future Work

- Compare a zero-shot setting using English with/out data-augmentation.
- Extend to other European languages.

That's it

Question?

Bibliography I

M. Cliche.
BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs.
SemEval-2017, (2014):573–580, 2017.

G. Lample and A. Conneau.
Cross-lingual Language Model Pretraining.
2019.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov.
RoBERTa: A Robustly Optimized BERT Pretraining Approach.
(1), 2019.

L. Martin, B. Muller, O. S. P. Javier, Y. Dupont, L. Romary, E. Villemonte de la Clergerie, D. Seddah, and B. Sagot.
CamemBERT: a Tasty French Language Model.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

D. Q. Nguyen, T. Vu, and A. T. Nguyen.
BERTweet: A pre-trained language model for English Tweets.
2020.

Bibliography II

M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile.
AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets.
CEUR Workshop Proceedings, 2481, 2019.