Transfer Learning from Pre-trained BERT for Pronoun Resolution

Xingce Bao and Qianqian Qiao
École Polytechnique Fédérale de Lausanne
xingce.bao@epfl.ch qianqian.qiao@epfl.ch

Task

The shared task Gendered Pronoun Resolution aims to classify the pronoun resolution in the sentences, hereby to find the true name referred by a given pronoun, such as she in: In May, Fujisawa joined Mari Motohashi's rink as the team's skip, moving back from Karuizawa to Kitami where she had spent her junior days. This task for pronoun resolution closely relates to the traditional coreference resolution task in natural language processing.

Motivation

It has been shown that most of the recent representative coreference systems struggled on GAP dataset with a overall mediocre performance and a large performance gap between genders. This may be due to both unbalanced training dataset used by these coreference systems or the design of the systems. Up to now, detecting and eliminating gender bias in such systems still remains a challenge.

Approach

Fine-tuned BERT

- With two different kinds of top layers (MLP-top and Positional-top) to fine-tune BERT model on GAP task.
- MLP-top extracts and aggregates vectors for all mentions by concatenation, which are then fed into a multiple layer neural network.
- Positional-top layer first maps the output of BERT into a scalar by a linear layer whose output size is 1. Then extract the value corresponding to the mention index and feed it into a softmax layer for a 3-class-probability-output.

BERT as Feature Extractor

- The contextual embeddings and the mention vectors prepared are passed to the subsequent classifier
- SVM (support vector machine) and BIDAF (bi-directional attention flow layer) as classifiers

Main Result

	M	F	T	PT
SVM 256	0.516	0.495	0.506	0.395
SVM 1024	0.619	0.574	0.596	0.475
BIDAF	0.490	0.498	0.494	0.364
BIDAF-aug	0.550	0.579	0.565	0.422
BERT-pos	0.376	0.377	0.377	**0.280**
BERT-mlp	**0.360**	**0.365**	**0.362**	0.351
Ensemble	**0.325**	**0.337**	**0.331**	**0.208**

Table 1: Evaluation results (multi-class logarithmic loss) for models.

	M	F	B	O
(Wiseman et al., 2016)	68.4	59.9	0.88	64.2
(Lee et al., 2017)	67.2	62.2	0.92	64.7
BERT-pos	**86.8**	**86.1**	0.99	**86.5**
BERT-mlp	86.3	85.9	**1.00**	86.1
Our ensemble	**88.1**	**87.9**	**1.00**	**88.0**

Table 2: Comparison to off-the-shelf resolvers, split by Masculine and Feminine Bias, and Overall.

Model

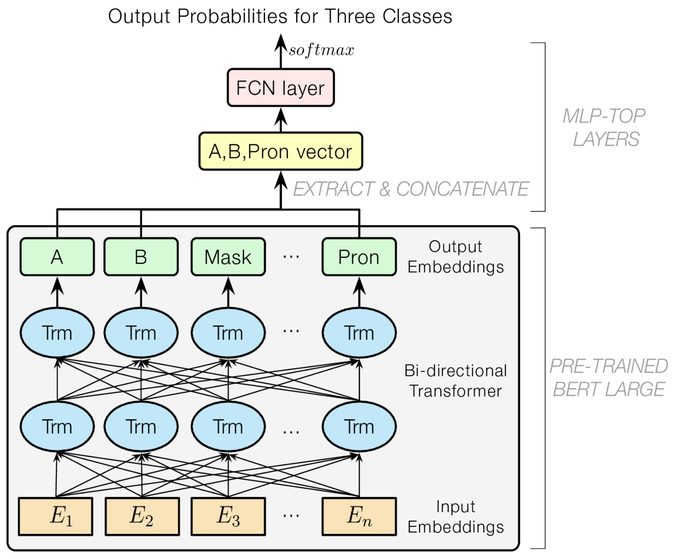

Figure 1: Fine-tuned BERT with MLP-top layer.

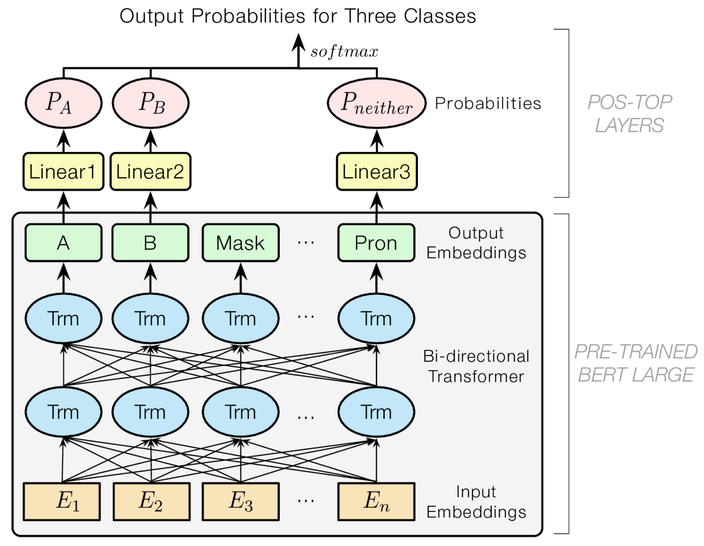

Figure 2: Fine-tuned BERT with Positional-top layer.

Analysis

Performance of Dimension Reduction

- The dimension reduction greatly enhances the result of SVM which reduces about 0.1 multi-class logarithmic loss. The SVM 1024 has a loss of 0.184 and 0.597 with respect to training and testing data, while the SVM 256 has a loss of 0.250 and 0.505. Both SVM model overfit a lot, while the dimension reduction of BERT contextual embeddings efficiently mitigate overfitting, which bridges the performance gap between training data and testing data.

Performance of Data Augmentation

- The BIDAF model performs worse when trained with the augmented training set than the original training set, due to the distribution mismatching caused by data augmentation that, the portion of the neither data is larger in the training set than in the testing set.

Efficacy of Fine-tuned Models

- Both two fine-tuned BERT models achieve much more competitive results compared to Bert as Feature Extractor models.

Efficacy of Ensemble Learning

- The ensemble learning with logistic regression greatly enhances the overall classification result.

Gender Bias

Method	GloVe	ELMo	BERT	F-BERT
WEAT	1.81^*	-0.45	0.21	0.38
SEAT	1.74^*	-0.38	0.08	0.07

Table 3: Effect sizes for male/female names with career/family task with word and sentence level embeddings. $*$: significant at 0.01. F-BERT indicates Fine-tuned BERT.

We apply WEAT and SEAT on Caliskan Test of male/female names with career and family, which corresponds to past social psychology studies. With p-values lower than 0.01, embeddings by GloVe on both word level and sentence level show significant gender bias, indicating that women are associated with family while men are associated with career.

However, p-values of all contextual embeddings including ELMo, BERT and Fined-tuned BERT are larger than 0.05, which suggests that there is no evidence suggesting existence of gender bias in these embeddings. One possible explanation is that, by training contextual word embeddings, a single word is usually represented differently in different sentences, resulting in more flexible word representations focusing on single context within a sentence rather than the overall word frequency distribution.

[1] Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California. Association for Computational Linguistics.
[2] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.