UoR-NCL at SemEval-2022 Task 3: Fine-Tuning the BERT-Based Models for Validating Taxonomic Relations

Thanet Markchom
University of Reading, UK
t.markchom@pgr.reading.ac.uk

Huizhi Liang
Newcastle University, UK
huizhi.liang@newcastle.ac.uk

Jiaoyan Chen
University of Oxford, UK
jiaoyan.chen@cs.ox.ac.uk

Abstract

In human languages, there are many presuppositional constructions that impose a constrain on the taxonomic relations between two nouns depending on their order. These constructions create a challenge in validating taxonomic relations in real-world contexts. In SemEval2022-Task3 Presupposed Taxonomies: Evaluating Neural Network Semantics (PreTENS), the organizers introduced a task regarding validating the taxonomic relations within a variety of presuppositional constructions. This task is divided into two subtasks: classification and regression. Each subtask contains three datasets in multiple languages, i.e., English, Italian and French. To tackle this task, this work proposes to fine-tune different BERT-based models pre-trained on different languages. According to the experimental results, the fine-tuned BERT-based models are effective compared to the baselines for classification. For regression, the fine-tuned models show promising performances with the possibility of improvement.

1 Introduction

Taxonomic relations are one of the significant lexical relationships that have been used in many applications such as question answering (Yih et al., 2013), sentiment analysis (Araque et al., 2019) and biomedical ontologies (Bodenreider, 2004). In natural languages, there are many constructions that constrain the taxonomic relation between two nouns based on the order of these two nouns. For instance, given a sentence "I have a dog, not a pet". The construction "I have a ..., not a ..." implies that the taxonomic relation does not hold between "dog" and "pet". This can be seen as a presupposition imposed by the construction. However, this is not true since dogs are pets. Thus, with various presuppositional constructions, validating taxonomic relations becomes more complicated in the real world.

To address this issue, SemEval2022-Task3 Presupposed Taxonomies: Evaluating Neural Network Semantics (PreTENS) (Zamparelli et al., 2022) introduces the task where taxonomic relations have to be validated in different presuppositional constructions. This task proposes novel datasets in multiple languages (i.e., English, Italian and French) containing sentences with different two-noun constructions. Each sentence is labeled by an acceptability label for classification and an acceptability score for regression. Two challenges have been raised in this task: (1) a taxonomic relation between two nouns in the sentence must be detected, and (2) the construction which embeds the two nouns must also be validated.

To effectively validate taxonomic relations in such constructions, understanding the contexts or semantic meanings of these constructions are the key. Many previous studies have shown that pre-trained models comprise the prior knowledge of context comprehension (Yang et al., 2019). Recently, the language mode called BERT has been widely used in several tasks. The BERT model can be pre-trained with the self-supervised method to generate word/token or sentence representations enriched with prior knowledge. Then, they can be fine-tuned specifically for many downstream tasks including validating taxonomic relations. Therefore, in this work, we adopt the pre-trained BERT-based models in different languages to utilize the prior knowledge from the resources that they were pre-trained with. Then, elaborating on the pre-training, we fine-tune these pre-trained models to predict the acceptability of each sentence.

2 Related Work

Pre-trained language models such as GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018) have been widely used to generate word/sentence

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 260 - 265
July 14-15, 2022 ©2022 Association for Computational Linguistics

representations for many NLP applications. These representations have been proven to be effective since they are enriched with knowledge from the pre-training resources. Recently, the transformer architecture (Vaswani et al., 2017), a multi-layer multi-head self-attention, has made the major success in NLP. With this architecture as a fundamental, a language model called BERT (Devlin et al., 2019) was proposed. Such model can be first pre-trained with two self-supervising tasks, masked LM and next sentence prediction. After that, they can be fine-tuned with additional output layers to create new models for various downstream tasks. Due to the huge success of the BERT model, many language models stemming from it have been proposed. These include RoBERTa (Liu et al., 2019), the more robust BERT model, and DistilBERT (Sanh et al., 2020), the modified BERT model using knowledge distillation methods.

All of these BERT-based models can be pre-trained on different corpora/resources depending on various purposes. Previously, there are variations of the BERT-based models pre-trained in different languages. MDZ Digital Library team at the Bavarian State Library introduced the variations of the BERT-based models pre-trained on Italian and German corpora[1]. Le et al. proposed FlauBERT, pre-trained on a large French corpus consists of texts in diverse topics and writing styles. Although they have been used to solve several tasks in many languages, how to use these models in validating taxonomic relations in real-world contexts still remains an open issue.

3 System Overview

This task consists of two sub-tasks (1) a binary classification sub-task and (2) a regression sub-task. For sub-task 1, each sentence in the datasets is labeled with 1 if it is acceptable and 0 otherwise. For sub-task 2, each sentence is labeled with the average score assigned by human annotators. The score is on a seven point Likert-scale ranging from 1 that means "not at all acceptable" to 7 that means "completely acceptable".

This paper proposes two similar approaches of fine-tuning the BERT-based models for taxonomic relation classification and regression. For both sub-tasks, we select three pre-trained BERT-based models that were pre-trained on three corpora with different languages. For English, we choose *DistilBERT-Base-Uncased* [2], pre-trained on Toronto Book Corpus and full English Wikipedia. The model has 6 layers, 12 heads, and 768 embedding dimension and has 66M parameters in total. For Italian, we select *BERT-Base-Italian-XXL-Uncased* [3], the Italian BERT model pre-trained on texts from a recent Wikipedia dump and the OPUS corpora collection. This model consists of 12 layers, 12 attention heads and 768 embedding dimension. The total number of parameters is 110M parameters. Lastly, for French, we select *FlauBERT-Base-Uncased* [4], pre-trained on text corpus consists of 24 sub-corpora gathered from different sources such as Project Gutenberg [5] and Common Crawl [6]. This model has 12 layers, 12 attention heads and 768 embedding dimension. The total number of parameters is 137M parameters. Based on these pre-trained models, an additional layer is added in each model to fine-tune these models. Each sub-task has different settings for fine-tuning.

Sub-task 1: Binary Classification To fine-tune the models for sub-task 1, a fully-connected layer is added on top of the pooled output (the sequence embedding, i.e., the "[CLS]" token embedding from the pre-trained model). This layer has an output size 2 and the softmax activation function. It outputs the probability of each class (1 and 0). The loss function for fine-tuning is the binary cross-entropy loss. The final prediction of each sentence is made by selecting the class with the maximum probability.

Sub-task 2: Regression Similarly to sub-task 1, we add a fully-connected layer on top of the pooled output for model fine-tuning. This layer has an output size 1. This output is the predicted score for a regression task. The mean squared error loss function is used for fine-tuning this model.

4 Experiments

We conducted experiments on the training and test sets provided by the task organizers. For each sub-task, there are three training sets and three test sets for three different languages. The training set of each language has 5,837 samples while the test set has 14,560 samples. Both training and test sets consist of sentences with different presuppositional

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

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

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

[4] https://github.com/getalp/Flaubert

[5] https://www.gutenberg.org/

[6] https://data.statmt.org/ngrams/deduped2017/

Construction	Sentence
andtoo	I like forests, and cities too.
butnot	I like sports, but not football.
comparatives	I like movies less than videogames.
drather	I would rather have beagles than rabbits.
except	I like pets, with the only exception of hamsters.
generally	I like bracelets, and more generally jewelry.
particular	I like fruits, and more specifically lemons.
prefer	I do not like cauliflower, I prefer apples.
type	I can stand rainstorms, an interesting type of rain.
unlike	Unlike cats, ducks are often mentioned in this text.
ingeneral	I like mountains, and nature in general.

Table 1: Examples of sentences with different constructions

Test set	Sub-task 1					Sub-task 2			
	Model	Precision	Recall	F1	F1-macro	Model	MSE	RMSE	ρ
English	Baseline	0.642	**0.866**	0.734	0.734	Baseline	4.45	2.11	0.23
	En-C	**0.900**	0.824	**0.860**	**0.873**	En-R	**2.71**	**1.65**	**0.24**
Italian	Baseline	0.557	0.877	0.682	0.682	Baseline	4.18	2.05	**0.40**
	It-C	**0.820**	**0.938**	**0.875**	**0.874**	It-R	**3.86**	**1.97**	0.04
French	Baseline	0.651	0.899	0.755	0.755	Baseline	4.66	2.16	**0.30**
	Fr-C	**0.763**	**0.905**	**0.828**	**0.823**	Fr-R	**3.65**	**1.91**	0.23

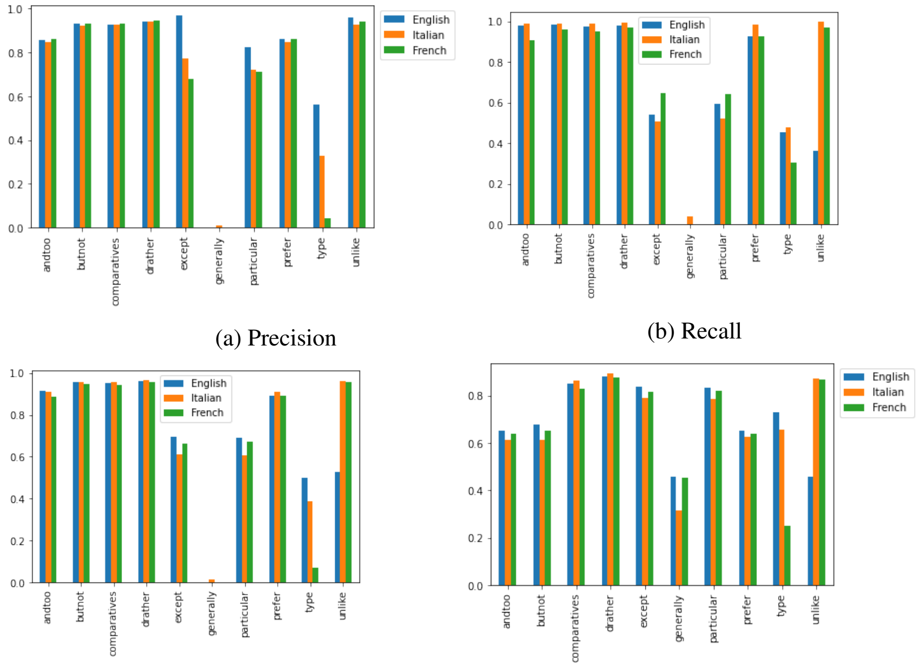
Table 2: Results of sub-task 1 and 2

constructions. For sub-task 1, there are 10 constructions, i.e., “andtoo”, “butnot”, “comparatives”, “drather”, “except”, “generally”, “particular”, “prefer”, “type” and “unlike”. For sub-task 2, there are 7 constructions, i.e., “andtoo”, “butnot”, “comparatives”, “ingeneral”, “particular”, “type” and “unlike”. Table 1 shows examples of sentences with different constructions. All the models were fine-tuned by using the Adam optimizer for 3 epochs with the batch size 16 and the learning rate 5e-5. For sub-task 1, the models were evaluated by Precision, Recall, F1 and F1-macro. For sub-task 2, they were evaluated by MSE, RMSE and Spearman Correlation (ρ). It is worth noting that ρ is used to measure the rank correlation between actual labels and predictions. It ranges between -1 to 1 where the higher value means the labels and predictions have a similar rank (or identical when it is 1) and the lower value means they have a dissimilar rank. (or fully opposed when it is -1) For each sub-task, a simple classification model using n-grams as features was used as a baseline.

4.1 Sub-Task 1 Results and Discussion

For this sub-task, we named the fine-tuned DistilBERT-Base-Uncased for classification as **En-C**, the fine-tuned BERT-Base-Italian-XXL-Uncased for classification as **It-C** and the fine-tuned FlauBERT-Base-Uncased for classification as **Fr-C**. Table 2 shows the results of sub-task 1. From this table, It-C and Fr-C performed better than the baseline in every evaluation metric. Meanwhile, En-C outperformed the baselines in terms of Precision, F1 and F1-macro.

We further investigated the performance of our approaches by comparing the results of En-C, It-C and Fr-C on each construction. The results are shown in Figure 1. From this figure, we can see that En-C, It-C and Fr-C performed particularly poorly on “generally” construction. To identify the mistake, we examined the confusion matrices of their performance on “generally” construction as shown in Figure 2. This figure shows that all of them failed in predicting the true positive cases of this construction. To answer why they failed to predict the true positive cases, we further examined the attention weights at the last layer of these



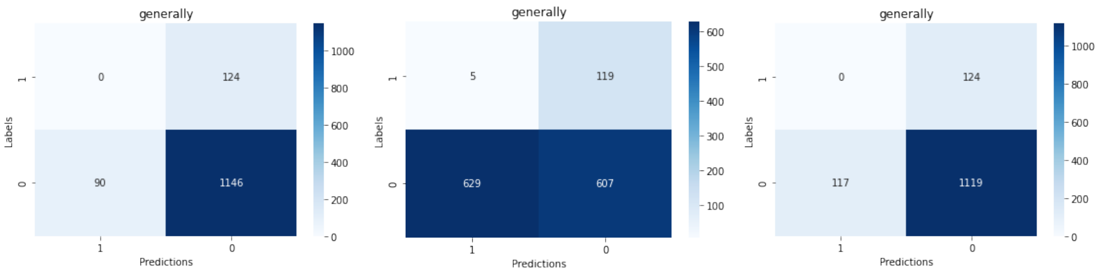
(a) Precision

(b) Recall

(c) F1

(d) F1-macro

Figure 1: Comparison of the proposed approach performance on each construction in sub-task 1



(a) En-C

(b) It-C

(c) Fr-C

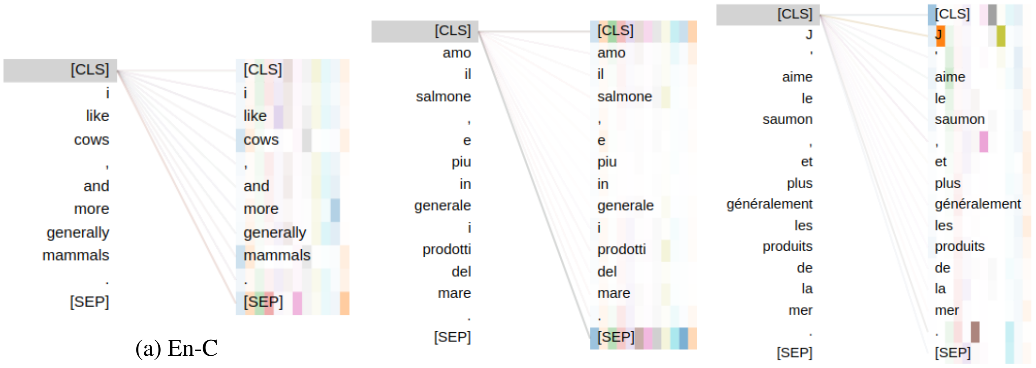
Figure 2: Confusion matrix of the results from (a) En-C, (b) It-C and (c) Fr-C on the construction “generally”

models using *bertviz*[7] library (Vig, 2019). Figure 3 illustrated the attention on the last layer of En-C, It-C and Fr-C given the acceptable sentence (labeled with 1) with “generally” construction as an input. In this figure, the attention is represented with lines connecting between the word being updated (on the left) and the word being attended to (on the right). The thickness of the lines indicates the weight. The thicker it is, the higher the weight will be. Since we use the embedding of “[CLS]” as the pooled output for fine-tuning, we only consider this token’s attention. From Figure 3a, the attention weights of “and”, “more” and “generally” tokens are relatively low compared to the other tokens. Similarly, the attention weights of “e”, “più” “in” and “generale” in It-C and the attention weights of “et”, “plus” and “généralement” in Fr-C are also low as shown in Figure 3b and 3c respectively. This suggests that these models ignored these tokens when they were fine-tuned. However, these tokens are important, since they act like keywords indicating the presuppositional “generally”. Therefore, ignoring them may result in mistakenly predicting the acceptability labels of this construction.

4.2 Sub-Task 2 Results and Discussion

For sub-task 2, we named the fine-tuned DistilBERT-Base-Uncased for regression as **En-R**, the fine-tuned BERT-Base-Italian-XXL-Uncased for regression as **It-R** and the fine-tuned FlauBERT-Base-Uncased for regression as **Fr-R**. The overall results are shown in Table 2. From this table, our approaches outperformed the baselines in terms

[7] https://github.com/jessevig/bertviz

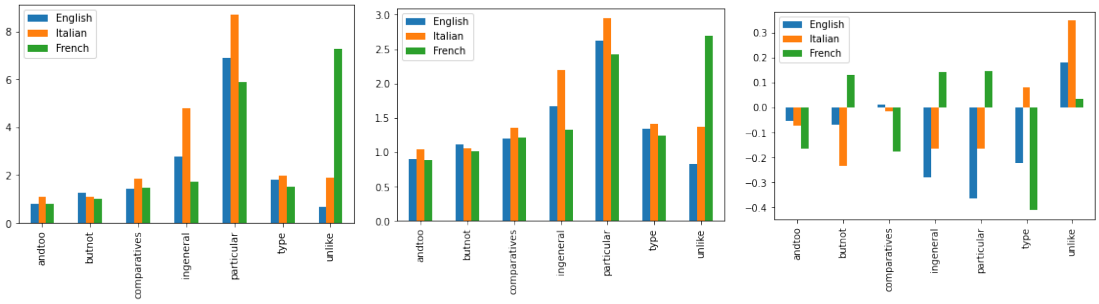


(a) En-C

(b) It-C

(c) Fr-C

Figure 3: Attention weights connecting with "[CLS]" token from the last layer of (a) En-C, (b) It-C and (c) Fr-C when the sentence with "generally" construction was given as an input

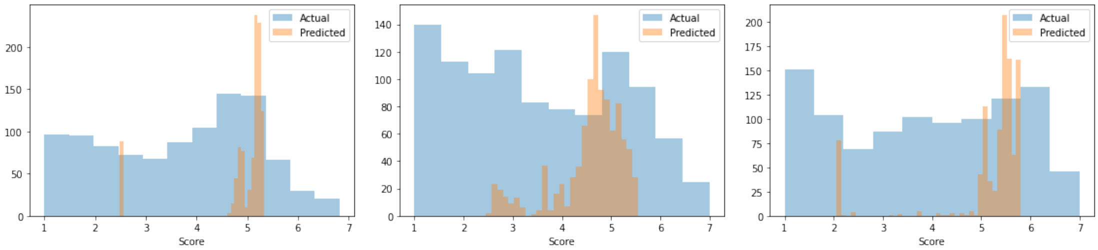


(a) MSE

(b) RMSE

(c) ρ

Figure 4: Comparison of the proposed approach performance on each construction in sub-task 2



(a) English

(b) Italian

(c) French

Figure 5: Distributions of the actual acceptability scores and the predicted acceptability scores of our approaches En-R, It-R and Fr-R on each test set (a) English, (b) Italian and (c) French respectively

of MSE and RMSE. Nonetheless, only En-R produced the results with the higher ρ than the baseline while the others failed to compete with their baselines. This suggests that the proposed approaches predicted the acceptability scores close to their actual scores but their ranks are dissimilar. Figure 5 shows the distributions of the actual acceptability scores and the predicted acceptability scores of our approaches, En-R, It-R and Fr-R on each test set, English, Italian and French. From this figure, we can see that all En-R, It-R and Fr-R tended to predict scores with low variances. This is possibly caused by using the mean squared error loss for fine-tuning these models.

As in sub-task 1, we also compared the results of them on each construction as shown in Figure 4. From Figure 4a and 4b, in terms of MSE and RMSE, En-R and It-R performed well on most constructions except "ingerneral" and "particular". Fr-R performed also well on almost every construction except "particular" and "unlike". On the other hand, in terms of ρ, the proposed models failed on most of the constructions as shown in 4c. En-R produced positive ρ on only "comparative" and "unlike". It-R only produced positive ρ on "type" and "unlike". Fr-R produced positive ρ on "butnot", "ingeneral", "particular" and "unlike". Overall, our models produced negative ρ in most of the constructions. This indicates that they failed to predict the acceptability scores with the same tendency as the actual scores. One possible reason is that the added regression layers are not suitable for fine-tuning these models.

5 Conclusion

This work proposes to fine-tune the pre-trained BERT-based models to validate taxonomic relations in different presuppositional constructions. Three different pre-trained BERT-based models are selected and fine-tuned to perform classification and regression on three different languages, English, Italian and French. According to the results, the fine-tuned models using the binary cross-entropy loss for classification are effective compared to the baseline. As for the regression subtask, the fine-tuned models using the mean squared error loss for regression performed less effectively than the baseline when evaluated with Spearman Correlation. This might be the result of using the mean squared error loss for fine-tuning. This leaves room for improvement in fine-tuning the BERT-based models for taxonomic relation regression.

References

Óscar Araque, Ganggao Zhu, and Carlos Angel Iglesias. 2019. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowl. Based Syst.*, 165:346–359.

Olivier Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*, 32:D267–70.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.

Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and Andrzej Pastusiak. 2013. Question answering using enhanced lexical semantic models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1744–1753, Sofia, Bulgaria. Association for Computational Linguistics.

Roberto Zamparelli, Shammur A. Chowdhury, Dominique Brunato, Cristiano Chesi, Felice Dell'Orletta, Arid Hasan, and Giulia Venturi. 2022. Semeval-2022 task3 (pretens): Evaluating neural networks on presuppositional semantic knowledge. In *Proceeding of SEMEVAL 2022*.