CSECU-DSG at SemEval-2022 Task 3: Investigating the Taxonomic Relationship Between Two Arguments using Fusion of Multilingual Transformer Models

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{aziz.abdul.cu, akram.hossain.cse.cu}@gmail.com,
and nowshed@cu.ac.bd

Abstract

Recognizing lexical relationships between words is one of the formidable tasks in computational linguistics. It plays a vital role in the improvement of various NLP tasks. However, the diversity of word semantics, sentence structure as well as word order information make it challenging to distill the relationship effectively. To address these challenges, SemEval-2022 Task 3 introduced a shared task PreTENS focusing on semantic competence to determine the taxonomic relations between two nominal arguments. This paper presents our participation in this task where we proposed an approach through exploiting an ensemble of multilingual transformer models. We employed two fine-tuned multilingual transformer models including XLM-RoBERTa and mBERT to train our model. To enhance the performance of individual models, we fuse the predicted probability score of these two models using weighted arithmetic mean to generate a unified probability score. The experimental results showed that our proposed method achieved competitive performance among the participants' methods.

1 Introduction

Lexical relations are regarded as the most important semantic relations to infer the meanings of words effectively (Khoo and Na, 2006). Therefore, identifying such relations is beneficial to understand the semantic competence and distill the underlying context of the textual expression. If we consider the NLP applications, it has a significant impact on several tasks including semantic search, automatic question/answering, story generation, relation extraction, and machine translation (Barkan et al., 2020). However, most of the prior works focused mainly on syntactic (Luu et al., 2014) and contextual relation (Maksimov et al., 2018) to infer the lexical relation in between words.

Sentence	Language	Label / Score
Sub-task 1		
Apprezzo il vino , ma non il Chianti.	It	1
J' aime les chats, sauf les beagles.	Fr	0
I like movies, but not comedies.	En	1
I like earrings, except socks.	En	0
Sub-task 2		
Amo i merli piÃ¹ degli uccelli.	It	1.9
J'aime les chats, et aussi les canards.	Fr	6.17
I like cats, but not sparrows.	En	4.77

Table 1: Examples of sub-task 1 and sub-task 2.

Considering this PreTENS shared task at SemEval 2022 (Zamparelli et al., 2022) introduce a new task that focuses mainly on semantic competence with specific attention on the estimating taxonomic relations between two nominal arguments. The taxonomic relation here means one argument is a supercategory of the other, or in extensional terms, one argument is a superset of the other. The task is divided into two subtasks. The first one is a binary classification subtask, where a system needs to predict the acceptability label for given text considering the taxonomic relation. The second one is a regression subtask, where a system needs to predict the percentage of acceptability label on a seven-point Likert-scale for a given text considering the same scenario. Besides, the task addresses the challenges of multilingual expression and comprises a dataset of three different languages including Italian (It), French (Fr), and English (En). To illustrate a clear view of the task definition and research goal, we articulate a few examples from different languages and corresponding labels for each subtask in Table 1.

We articulate the rest of the contents as follows: Section 2 describes our proposed approach

**The first two authors have equal contributions.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 255 - 259

whereas, in Section 3, we present our experimental setup and conduct performance analysis against the various settings and participants' methods. Finally, we conclude our work with some future directions in Section 4.

2 Proposed Framework

In this section, we describe our proposed approach for the PreTENS shared task. Our goal is to determine the semantic competence between two arguments focusing on the taxonomic relations. The task is articulated into both the binary classification subtask and the regression subtask. We depict the overview of our proposed framework in Figure 1.

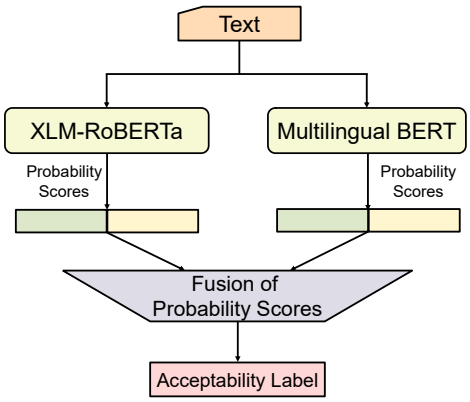

Figure 1: Overview of our proposed framework.

Given an input sentence containing two nominal arguments, we employ two transformer models including XLM-RoBERTa and multilingual BERT (mBERT) to extract the diverse contextual features. Later, a feed-forward linear layer is employed in each model to estimate the probability score of each class. In Figure 2, we illustrate an overview of the setup of mBERT transformer model. Finally, we fuse these models' predictions by taking the weighted arithmetic mean of these scores to determine the final label.

2.1 XLM-RoBERTa

The Facebook AI launched the XLM-RoBERTa as an upgrade to their initial XLM-100 model (Conneau et al., 2020). It is a scaled cross-lingual sentence encoder. Using self-supervised training approaches, it offers state-of-the-art performances in cross-lingual understanding where a model is taught in one language and then applied to multiple languages with no additional training data. This model showed increased performance on numerous NLP applications. XLM-RoBERTa creates the

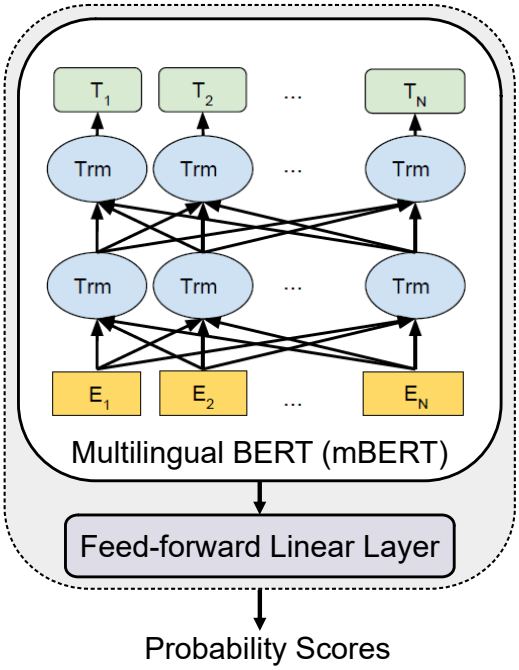

Figure 2: Multilingual BERT (mBERT) model.

possibility for a one-model-for-many-languages approach rather than a single model per language.

However, XLM-R maps the same sentence in different languages to similar representations which is crucial for this PreTENS task to learn semantic competence in cross-lingual form. Here, we use the HuggingFace's implementation of the XLM-RoBERTa-base model (Conneau et al., 2020). It is composed of 12-layers (i.e. transformer block), the dimension of hidden size is 768, the number of the self-attention head is 12, the size of vocabulary is 250K, and containing 270M parameters.

2.2 Multilingual BERT

Multilingual-BERT (mBERT) (Devlin et al., 2019) is a version of BERT that is gaining popularity for effective contextual representation of textual contents in various multilingual tasks including natural language inference in cross-lingual characteristics, named entity recognition in multilingual corpora, and dependency parsing (Chi et al., 2020). It is pretrained on 104 different languages in Wikipedia. It provides an effective path to zero-shot cross-lingual model transfer.

Multilingual-BERT representations are influenced by high-level grammatical features that are not manifested in any one input sentence which is critical to learn taxonomic relations in PreTENS task. In our approach, we employ the huggingface

https://huggingface.co/xlm-roberta-base

implementation of the bert-base-multilingual-cased model and perform finetuning the model with the task-specific data.

2.3 Fusion of Transformer Models

To enhance the performance of individual models, we fuse the predicted probability score of two models to generate a unified probability score for each class. We use the weighted arithmetic mean to average both model's probability scores to determine the each class confidence of the fused model. After calculate the both class probability score of fused model, our system predict the best confidence class, which is our final label. The estimation is computed as follows:

$$f(x_i, y_i) = \begin{cases} 0, & \text{if } W_0 > W_1 \\ 1, & \text{otherwise} \end{cases}$$

$$W_i = \frac{x_i * M + y_i * R}{M + R}$$

x_i and y_i correspond to the mBERT and XLM-RoBERTa probability score, where M and R represents their weight respectively. W_i (i.e. i = {0, 1}) is the unified probability score for each class.

3 Experiment and Evaluation

3.1 Dataset Description

The organizers of the SemEval-2022 PreTENS shared task 3 (Zamparelli et al., 2022) provided a benchmark dataset to evaluate the performance of the participants' systems. The dataset comprises in 3 languages including English, Italian, and French. The French and Italian are slightly adapted translations of the English dataset. The dataset statistics is summarized in Table 2.

Language	Sub-task 1		Sub-task 2	
	Train	Test	Train	Test
It	5837	14560	524	1009
Fr	5837	14560	524	1009
En	5837	14560	524	1009
Total	17511	43680	1572	3027

Table 2: The statistics of the datasets used in PreTens shared task.

Dataset of each languages contained about 20K artificially generated text samples that enforces presuppositions on the taxonomic status of their arguments A and B, e.g. comparatives (I like A more than B), exemplifications (I like A, and in particular B), generalizations (I like A, and B in general), and others. In Subtask 1, all samples are provided with an acceptability label either 1 or 0 where 1 stands for acceptable (i.e. the taxonomical relations is compatible with the construction at issue) and 0 stands for not acceptable (not compatible). Besides, a subset of 1533 samples of the whole dataset i.e. 5% of the total and representative of the patterns considered, was used for the subtask 2. It was annotated via a crowdsourcing campaign on a seven point Likert-scale, ranging from 1 (not at all acceptable) to 7 (completely acceptable). The average judgment is considered as the final label for each sample.

3.2 Experimental Settings

We now describe the details of our experiments and set of parameters that we have used to design our proposed CSECU-DSG system for each subtask.

Parameter	Optimal Value
Subtask 1: Parameters used in both models	
Learning rate	3e-5
Max-len	100
Epoch	3
Batch size	16
Manual seed	4
Subtask 2: Parameters of XLM-RoBERTa	
Learning rate	3e-5
Max-len	100
Epoch	5
Batch size	8
Manual seed	4

Table 3: Model settings for each subtask.

In Subtask 1, we utilize the Huggingface (Wolf et al., 2019) implementation of the two state-of-the-art multilingual transformer models with fine-tuning, including XLM-RoBERTa and mBERT. We finetune these models with the provided training data. To generate the unified prediction, we fuse the probability score of each model as described in Section 2.3. Since XLM-R typically perform better than the mBERT, so we don't count both model confidence weight as equal. However, from some sets of experimental result we choose the

https://huggingface.co/bert-base-multilingual-cased

weight R = 0.6 for XLM-RoBERTa and M = 0.4 for mBERT in equation 2.3. However, in Subtask 2, we only employed the XLM-RoBERTa model with finetuning strategy. The optimal parameter settings used in both subtasks are articulated in Table 3 and we used the default settings for the other parameters.

3.3 Evaluation Measures

To assess the performance of the participants' systems, PreTENS task organizers (Zamparelli et al., 2022) used different strategies and metrics for subtask 1 and subtask 2. Since the evaluation file contains instances from all three languages, the average of the F1-macro score from all the three languages is used as the global ranking score (GRS) to rank the participants' systems. We can write this as follows:

$$\text{Global Rank Score, } GRS = \frac{AS}{3}$$

where AS=(F1-macro (English) + F1-macro (French) + F1-macro (Italian)).

In subtask 2, the average of the Spearman correlation (Rho) scores from all the three languages is used as the global ranking score (GRS) to rank the participants' systems. We can write this as follows:

$$\text{Global Rank Score, } GRS = \frac{AS}{3}$$

where AS=(Rho (English) + Rho (French) + Rho (Italian)).

3.4 Results and Analysis

In this section, we analyze the performance of our proposed approaches in the PreTENS shared subtasks. The dataset comprises of 3 different languages including English, Italian, and French and the overall performance of the system is estimated considering the average score obtains in each languages dataset. Considering this, we analyze the performance of our CSECU-DSG system, based on each language. The corresponding results for subtask 1 and subtask 2 are reported in Table 4 and Table 5, respectively.

Results showed that in Subtask 1 our CSECU-DSG system achieved a pretty good score in all sets of datasets considering three languages. It demonstrates the generalizability of our approach in diverse types of languages. However, our method limited to obtain a good score in Subtask 2 for all the datasets. One plausible reason behind this is to use only single transformer models instead of the fusion approach and failed to address and analyze the challenges of subtask 2 properly.

Language	F1-macro	F1
It	91.113	90.620
Fr	90.732	90.334
En	91.506	91.189
All (CSECU-DSG)	91.117	-

Table 4: Results of our proposed CSECU-DSG system on individual monolingual tracks (subtask 1).

Language	Spearman Cor. (Rho)
It	0.207
Fr	0.081
En	0.191
All (CSECU-DSG)	0.160

Table 5: Results of our proposed CSECU-DSG system on individual monolingual tracks (subtask 2).

Next, the obtained results of our proposed CSECU-DSG system in the PreTENS shared task along with other top performing and competitive participants systems in subtask 1 and subtask 2 are articulated in Table 6 and Table 7, respectively. Following the benchmark of the PreTENS shared task, participants' systems are ranked based on the primary evaluation measures of each subtask.

Team (Rank)	F1-macro			
	Global	It	Fr	En
CSECU-DSG (4th)	91.117	91.113	90.732	91.506
Performance of team based on F1-macro score				
LingJing (1st)	94.485	93.047	93.236	97.172
injySarhanUU (3rd)	91.574	92.118	89.529	93.076
holdon (6th)	89.579	86.903	87.281	94.551
cnxupupup (10th)	86.676	86.755	86.390	86.883
breaklikeafish (15th)	77.985	74.398	82.345	77.213
Baseline (18th)	67.394	59.588	72.126	70.468

Table 6: Comparative results with other selected participants (Sub-task 1).

Results showed that our system ranked 4th among the participants' systems in subtask 1. This deduces the efficacy of our approach in addressing

Team (Rank)	Spearman Correlation (Rho)			
	Global	It	Fr	En
CSECU-DSG (8th)	0.160	0.207	0.081	0.191
Comparative performance of team based on Rho score				
LingJing (1st)	0.802	0.807	0.841	0.758
huawei_zhangmin (3rd)	0.669	0.631	0.740	0.636
daydayemo (6th)	0.206	0.121	0.284	0.212
breaklikeafish (11th)	0.078	0.186	-0.010	0.059
thanet.markchom (13th)	0.056	0.089	-0.043	0.122
Baseline (4th)	0.309	0.344	0.317	0.265

Table 7: Comparative results with other selected participants (Sub-task 2).

the challenges of the PreTENS shared task. However, in subtask 2, our system obtained a poor score though we ranked 8th in this task.

To further analyze the effectiveness of the components used in our approach, we estimate the performance of each model used in the fusion approach in subtask 1. The results are reported in Table 8. It shows that our fusion strategy improves the overall performance of the model compared to the performance of the mBERT and XLM-RoBERTa. However, considering the individual model's performances XLM-RoBERTa performed better compared to the mBERT.

Method	F1-macro			
	Global	It	Fr	En
CSECU-DSG	**91.117**	**91.113**	**90.732**	**91.506**
Performance of individual model				
XLM-RoBERTa	90.217	90.322	89.853	90.475
mBERT	88.076	86.477	88.288	89.463

Table 8: Performance analysis of individual model using subtask 1 test dataset.

4 Conclusion and Future Directions

In this paper, we presented our proposed system to address the challenges of the PreTENS shared task. We employed the weighted fusion of two state-of-the-art multilingual transformer models predictions. Experimental results demonstrate the competency of our approach in Subtask 1.

In the future, we have a plan to incorporate the task specific features and technologies to address the challenges properly. We also have a plan to explore the causal inference techniques to distill the taxonomic relation.

References

Oren Barkan, Avi Caciularu, and Ido Dagan. 2020. Within-between lexical relation classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3521–3527.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.

Christopher SG Khoo and Jin-Cheon Na. 2006. Semantic relations in information science. *Annual review of information science and technology*, 40(1):157–228.

Anh Tuan Luu, Jung-jae Kim, and See Kiong Ng. 2014. Taxonomy construction using syntactic contextual evidence. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 810–819.

NV Maksimov, AS Gavrilkina, VV Andronova, and IA Tazieva. 2018. Systematization and identification of semantic relations in ontologies for scientific and technical subject areas. *Automatic Documentation and Mathematical Linguistics*, 52(6):306–317.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Robertoa Zamparelli, Shammur Absar 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 *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.