LingJing at SemEval-2022 Task 3: Applying DeBERTa to Lexical-level Presupposed Relation Taxonomy with Knowledge Transfer

Fei Xia[1,2]*, Bin Li[3]*, Yixuan Weng[1]*, Shizhu He[1,2],
Bin Sun[3], Shutao Li[3], Kang Liu[1,2], Jun Zhao[1,2]
[1] National Laboratory of Pattern Recognition Institute of Automation, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] College of Electrical and Information Engineering, Hunan University
{libincn, shutao_li, sunbin611}@hnu.edu.cn, wengsyx@gmail.com,
xiafei2020@ia.ac.cn, {shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

This paper presents the results and main findings of our system on SemEval-2022 Task 3 Presupposed Taxonomies: Evaluating Neural Network Semantics (PreTENS)[1]. This task aims at semantic competence with specific attention on the evaluation of language models, which is a task with respect to the recognition of appropriate taxonomic relations between two nominal arguments. Two sub-tasks including binary classification and regression are designed for the evaluation. For the classification sub-task, we adopt the DeBERTa-v3 pre-trained model for fine-tuning datasets of different languages. Due to the small size of the training datasets of the regression sub-task, we transfer the knowledge of classification model (i.e., model parameters) to the regression task. The experimental results show that the proposed method achieves the best results on both sub-tasks. Meanwhile, we also report negative results of multiple training strategies for further discussion. All the experimental codes are open-sourced at https://github.com/WENGSYX/Semeval.

1 Introduction

In order to dive into the capability of the current language models (Bengio et al., 2000; Howard and Ruder, 2018), we take part in and apply the pre-training language model on the SemEval-2022 Tasks 3: the Presupposed Taxonomies: Evaluating Neural Network Semantics (PreTENS) (Zamparelli et al., 2022). To evaluate the model performance comprehensively, two sub-tasks are designed in this PreTENS task:

Sub-task 1) Binary classification task[2], which consists in predicting the acceptability label assigned to each sentence of the test set. For example, "I like trees, and in particular birches" is acceptable while " I like oaks, and in particular trees" is unacceptable, so they are labeled 1 and 0, respectively.

Sub-task 2) Regression sub-task[3], which consists in predicting the average score assigned by human annotators on a seven-point Likert scale (Joshi et al., 2015) with respect to the subset of data evaluated via crowdsourcing. For example,"I like governors, an interesting type of politician" is more acceptable than "I like politicians, an interesting type of farmer", so the former will also have a higher score (6.16) than the latter (1.42).

It is noted that both sub-tasks comprise datasets in 3 languages: English, Italian, French, where French and Italian are slightly adapted translations of the English dataset. For each sub-task, every sample is formed as the 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, where the argument nouns are taken from various semantic categories.

The most similar tasks are the Natural Language Inference (MacCartney, 2009; Bowman et al., 2015; Conneau et al., 2017) and Taxonomy Expansion & Enrichment (Zhang et al., 2018; Shen et al., 2018; Yu et al., 2020), where the former requires the model to differentiate the relationship between a premise sentence and a hypothetical sentence, while the latter shall identify the relationship between different concepts. There are many powerful language models for accomplishing these tasks (Devlin et al., 2018; He et al., 2020), and well-formed semantic representations can be obtained for the downstream tasks. However, it is a challenge for the model to decide the acceptance of the given sentence, as the semantic meaning is hard to be distinguished. Moreover, the performance of downstream tasks when fine-tuning is limited by the size of the training dataset (Xie et al., 2020).

To solve the above problems, we propose a

*These authors contribute equally to this work.
[1]https://sites.google.com/view/semeval2022-pretens/
[2]https://codalab.lisn.upsaclay.fr/competitions/1292
[3]https://codalab.lisn.upsaclay.fr/competitions/1290

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 239 - 246

method that applies DeBERTa to lexical-level presupposed relation taxonomy with knowledge transfer. Specifically, the powerful DeBERTa-v3 pre-trained model (He et al., 2021) is fine-tuned with datasets of different languages in the classification task. For the regression task, due to the limited size of the training datasets, we fine-tune the trained model of the classification task in the regression task. As a result, the proposed method achieves a global top score of 94.173 in sub-task 1 and a global top score of 0.802 in sub-task 2. Our method wins on two sub-tasks. In addition, we present the negative results of multiple training strategies when fine-tuning and provide further discussions.

2 Main method

In this section, we will elaborate on the main methods for the two sub-tasks of the PreTENS task. The training strategies are included at the end of this section.

2.1 Sub-task 1 - Binary classification using DeBERTa

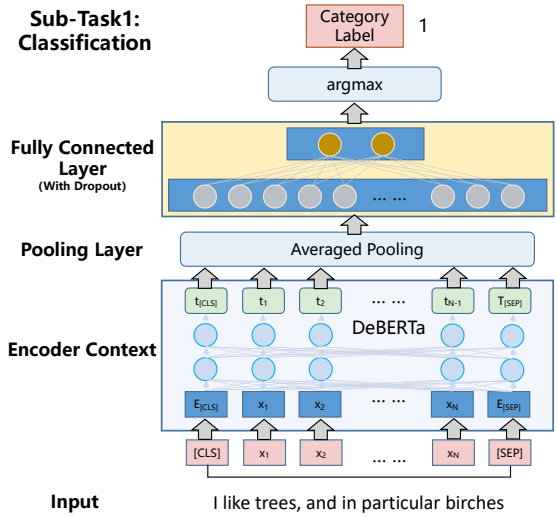

Figure 1: Main structure of the method in sub-task 1.

Sub-task 1 is a classic classification task, where two acceptability labels are required to be classified. We adopt the DeBERTa-v3 (He et al., 2021) model for processing this binary classification, where the main method structure is shown in Figure 1. The given sentence is separated into tokens and then sent to the pre-trained model as the input. To obtain the complete meaning of the whole sentence, we take the output embedding of each token to be averaged by the averaged pooling layer. The binary classification task is designed by sending the averaged encoding into the fully connected layer with dropout.

2.2 Sub-task 2 - Regression with knowledge transferring

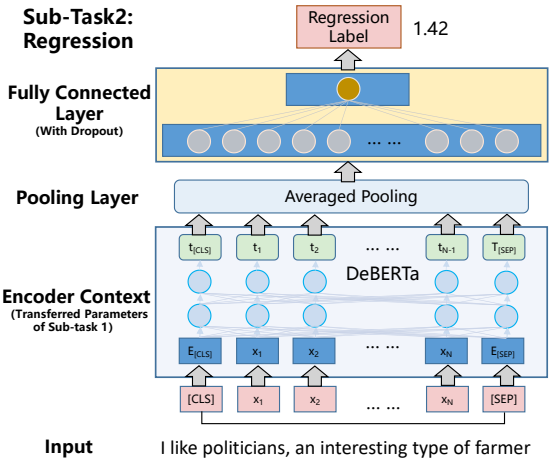

Figure 2: Main structure of the method in sub-task 2.

Sub-task 2 is a regression task, where a seven-point Likert-scale which ranges from 1 (not at all acceptable) to 7 (completely acceptable) is used to perform the regression. There is a subset of 1,533 sentences of the entire dataset for this regression task, where the total number is relatively small. It is a wise choice to transfer the knowledge from the pre-trained model of sub-task 1 into the regression task. The reason is that the fine-tuned model of sub-task 1 learns quite a few patterns of sentence acceptance that come from the extra knowledge. As shown in Figure 2, the DeBERTa-v3 model trained in sub-task 1 is designed to perform the regression task. The fully connected layer with dropout for obtaining the output logits is used for the regression task.

2.3 Multiple training strategies

In this section, we will introduce some training strategies used in the competition, which includes data augmentation with translation, adversarial training and child-tuning training.

2.3.1 Data augmentation with translation

When fine-tuning the English datasets, we translate the training sample of the Italian and the French to English one by one based on the M2M-1.2B model (Ott et al., 2019; Fan et al., 2020). It is a process that provides more useful datasets from the same

Example	Sample	Label
Classification	I like trees, and in particular birches	1
	I like oaks, and in particular trees	0
Regression	I like politicians, an interesting type of farmer	1.42
	I like governors, an interesting type of politician	6.16

Table 1: The dataset sample example.

resources. As a result, the DeBERTa model fine-tuned the datasets in English may achieve better results.

2.3.2 Adversarial training

The common method in adversarial training is the Fast Gradient Method (FGM). The idea of the FGM (Miyato et al., 2016) is straightforward[4]. The loss is to increase the gradient so that we can take

$$\Delta x = \epsilon \nabla_x L(x, y; \theta) \tag{1}$$

where x represents the input, y represents the label, θ is the model parameter, $L(x, y; \theta)$ is the loss of a single sample, Δx is the anti-disturbance. To prevent Δx from being too large, it is usually necessary to standardize $\nabla_x L(x, y; \theta)$. The more common way is

$$\Delta x = \epsilon \frac{\nabla_x L(x, y; \theta)}{\|\nabla_x L(x, y; \theta)\|}. \tag{2}$$

2.3.3 Child-tuning training

We use the Child-tuning (Xu et al., 2021) for fine-tuning the pre-trained model and only update the parameters of the Child network through gradients mask. For the two sub-tasks, the task-independent algorithm is used. In the process of fine-tuning, the gradients mask is obtained by sampling from the Bernoulli distribution (Chen and Liu, 1997) in each step of iterative update, which is equivalent to randomly dividing a part of the network parameters when updating. The equation of the above steps is shown as follows

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t - \eta \frac{\partial \mathcal{L}(\mathbf{w}_t)}{\partial \mathbf{w}_t} \odot M_t \\ \boldsymbol{M}_t &\sim \text{ Bernoulli } (p_F). \end{aligned} \tag{3}$$

3 Experimental setup

3.1 Data description

The PreTENS dataset not only needs to judge the classification relationship between two nouns (Wang et al., 2017), but also needs to identify whether the two nouns are in line with the actual situation in the artificially constructed natural sentences. The argument nouns are taken from 30 semantic categories (e.g., dogs, birds, mammals, cars, motorcycles...).

Specifically, PreTENS is articulated into the two following sub-tasks. The classification task requires judging the acceptability of the samples. The training set and test set contain 5838 and 14556 samples. Regression Sub-task obtains scores from 1 (not at all acceptable) to 7 (completely acceptable) through human crowdsourcing, which could be affected by usability considerations, argument order, and other factors. The data set of the regression sub-task is a small amount, 524 sentences will be provided for the training set and 1,009 for the test set. The example of two sub-task datasets is shown in Table 1, and the number statistics of each sub-task is shown in Table 2.

Datasets	Classification Task		Regression Task	
Language	Train	Test	Train	Test
English	5837	14560	524	1009
French	5837	14560	524	1009
Italian	5837	14560	524	1009

Table 2: Number statistics of task dataset samples.

3.2 Evaluation metrics

For classification tasks, the official evaluation indicators include precision, recall, macro F1, and global score. The global score is the average value of macro F1.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2PrecisionRecall}{Precision + Recall}$$

[4]https://spaces.ac.cn/archives/72

$$MacroF1 = \frac{\sum_{i=1}^{n} F1_i}{n}$$

$$Global = \frac{F1(En) + F1(Fr) + F1(It))}{3}$$

where n means that the higher the total number of categories, accuracy, recall rate, and macro F1. The higher F1 the method, the better the performance.[5]

For regression, it is mainly evaluated by MSE, RMSE and Spearman correlation (rho) (Wissler, 1905).

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{R_i - Q_i}{\sigma_i}\right)^2$$

$$RMSE = \sqrt{MSE}$$

$$RHO = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n(n^2-1)}$$

where the paired values of two variables are ranked from small to large (or from large to small). R_i represents the rank of x_i, Q_i represents the rank of y_i, and R_i - Q_i is the difference between the ranks of x_i and y_i.

3.3 Baselines introduction

3.3.1 Binary classification sub-task

N-gram method The original baseline provided by the organizers is based on the n-grams (Broder et al., 1997), where n=3. A Linear Support Vector (Tang, 2013) classifier using n-grams (set to 3) as input features is used for the binary classification.

mDeBERTa model The mDeBERTa (He et al., 2021) is a multilingual version of DeBERTa (He et al., 2020) which uses the same structure as DeBERTa and was trained with CC100 multilingual data (Wenzek et al., 2020; Conneau et al., 2020). The mDeBERTa model comes with 12 layers and a hidden size of 768. It has 86M backbone parameters with a vocabulary containing 250K tokens which introduce 190M parameters in the Embedding layer. This model was trained using the 2.5T CC100 data as XLM-R (Conneau et al., 2019).

3.3.2 Regression sub-task

N-gram method This method is provided by the organizers, where it uses a Linear Support Vector regressor with the 3-grams features is provided for the regression sub-task.

[5]Below is the specific meaning of the formula. TP: The prediction is correct and the sample is correct. FP: The prediction is wrong and the sample is correct. FN: The prediction is correct and the sample is wrong.

mDeBERTa model We adopt the mDeBERTa (He et al., 2021) for processing the sub-task 2. The architecture of this method is the same as Figure 2, where the linear layer over in the mDeBERTa is initialized before fine-tuning. A clamped step is performed for obtaining the final results of the regression.

3.4 Implementation details

We train the model based on the PyTorch (Paszke et al., 2019) and use the hugging-face (Wolf et al., 2020) framework. During training, we employ the AdamW optimizer (Loshchilov and Hutter, 2017). The default learning rate is set to 1e-5 with the warm-up (He et al., 2016). Four RTX3090 GPUs are implemented for all experiments. There are some variants of the DeBERTa-v3 model, i.e., base and large model. We adopt DeBERTa-v3-large models as our backbone, where the batch size is set to 24, and the max length of input is set to 64. We train our backbone for 6 epochs, and save the model parameters at the end of each round. We test the saved checkpoints in the evaluation phase, and select the highest score as the experimental result.

For the sub-task 1, we will implement the overall training (by mixing the datasets in different languages for training), and separate training (to fine-tune in one language and expand to the others), and conduct experiments on the training strategies such as FGM, data augmentation with translation and Child-tuning.

For the sub-task 2, we map the result of [1,7] in the label space to the minimum value of [0,1]. The results of the regression model are clamped so that the minimum value is 0 and the maximum value is 1. Moreover, we will compare the performance of models which use knowledge transfer or not.

4 Results and discussions

In this section, we studied the experimental results of the two sub-tasks and the impact of different strategies on the results, and further discussed the comparison with other participants to prove the effectiveness of the method. Finally, Studies analyze the deviation of language model and the future research direction.

4.1 Experimental results

4.1.1 Classification sub-task

Model Selection We first carry out experiments on three different models when fine-tuning, namely

Experimental Items		English			French			Italian		
Method	Global	Precision	Recall	macro F1	Precision	Recall	macro F1	Precision	Recall	macro F1
N-gram	72.34	64.19	85.64	73.38	65.07	89.92	75.50	55.71	87.73	68.15
mDeBERTa	88.58	83.33	95.30	88.92	82.72	93.33	87.71	83.38	95.67	89.11
DeBERTa-v3-base	72.88	75.66	69.91	75.03	83.60	75.18	81.16	58.60	69.02	62.45
DeBERTa-v3-large	92.90	88.10	95.83	91.94	92.31	93.89	93.42	93.11	93.48	93.35
DeBERTa-v3-large in English	92.24	96.18	**98.57**	**97.48**	84.38	**97.14**	90.19	82.36	**97.11**	89.06
DeBERTa-v3-large in French	89.33	85.12	96.35	90.36	85.23	86.79	86.67	88.03	93.58	90.98
DeBERTa-v3-large in Italian	91.78	**96.88**	89.20	93.50	92.80	82.75	88.72	**94.22**	91.06	93.12
DeBERTa-v3-large in En. and Fr.	93.21	92.18	96.73	94.59	**95.34**	92.26	94.21	92.31	87.95	90.82
DeBERTa-v3-large + FGM	88.32	83.72	97.61	89.93	81.11	96.09	87.62	80.85	96.05	87.42
DeBERTa-v3-large + Translation	92.75	92.93	96.66	94.96	88.25	94.06	91.30	89.04	94.67	92.00
DeBERTa-v3-large + Child-tuning	94.41	91.57	97.74	94.70	93.26	94.93	**94.37**	92.84	94.99	**94.18**
Ours	**95.34**	96.18	**98.57**	**97.48**	93.26	94.93	**94.37**	92.84	94.99	**94.18**

Table 3: Main experimental results of the sub-task1. From top to bottom, the first four lines are the comparison between different pre-training models, and then the best model DeBERTa-large is selected as the subsequent fine-tuning model. The four lines in the middle represent the way to use separate or overall datasets for training, not all data sets. The last three lines are some training strategies used for fine-tuning. We chose the highest score of all experiments under different data sets as "Ours" "Global Score".

Experimental Items	Original				Fine-Tuned			
Method	Global Score	RHO(EN)	RHO(FR)	RHO(IT)	Global Score	RHO(EN)	RHO(FR)	RHO(IT)
N-gram	0.309	0.265	0.317	0.344	/	/	/	/
mDeBERTa	0.430	0.412	0.350	0.529	0.720(+0.290)	0.670(+0.258)	0.783(+0.433)	0.708(+0.179)
DeBERTa-v3-base	0.108	0.216	-0.018	0.124	0.275(+0.167)	0.266(+0.050)	0.232(+0.250)	0.326(+0.202)
DeBERTa-v3-large	0.429	0.426	0.344	0.516	0.815(**+0.386**)	0.759(**+0.333**)	0.849(**+0.505**)	0.837(**+0.321**)

Table 4: Main experimental results of the sub-task2.

mDeBERTa, DeBERTa-v3-base, DeBERTa-v3-large. It can be found in the Figure 3 that the DeBERTa-v3-large model beats the N-gram method and surpass mDeBERTa in French and Italian. It shows that the larger model can bring improvements in the classification task.

Datasets Choosing We are more curious about the generalization ability of the DeBERTa model to fine-tune one language and then transfer to other languages. As a result, we fine-tune the datasets of separate and overall training in different languages. Although the performance by fine-tuning a single language is not satisfactory, it can be better than overall training. It is because there will be large gaps between different languages, and forced dataset mixing will reduce the final performance.

Training strategies After that, we compare some commonly used training strategies and the experimental results in this task, including FGM, translation, and Child-tuning. The performance of the FGM is not good, which indicates that the task pays more attention to the semantic features at the lexical level than the semantic features of sentences. All languages are translated into English, even better than using the original language. We believe that this is because the error accumulation caused by inaccurate translation will interfere with the semantic representation information. The result of Child-tuning is positive, which shows that the catastrophic forgetting problem of the model can be alleviated by eliminating some unimportant weights in the large model.

4.1.2 Regression sub-task

In our experiments of sub-task 2 shown in the Figure 4, we select three different methods for the experiments and compare them with the same experimental settings. The "Original" is the original pre-training model, the "Fine-Tuned" represents the model that is fine-tuned by sub-task1 and re-initialized the linear layer. The experimental results of three different methods show that the performance of the model can be greatly improved through knowledge transfer.

4.2 Official results

As shown in Table 5 and Table 6, our method achieved first place in subtask 1and subtask 2 and

Experimental Items		English			French			Italian		
System	Global	Precision	Recall	macro F1	Precision	Recall	macro F1	Precision	Recall	macro F1
Ours(LingJing)	**94.173**	97.65	96.34	**96.988**	93.15	92.48	**92.817**	91.68	93.77	**92.714**
YingluLi	92.310	93.29	91.89	92.582	93.34	91.77	92.546	92.94	90.69	91.802
injySarhanUU	91.247	90.54	95.26	92.839	85.83	93.14	89.335	92.69	90.47	91.567
piano	90.842	97.72	96.15	96.926	78.85	96.51	86.792	84.86	93.14	88.807
csecudsg	90.714	89.21	93.26	91.189	88.87	91.84	90.334	90.32	90.92	90.620
holdon	89.686	92.66	96.05	94.325	81.25	94.89	87.541	80.81	94.67	87.193

Table 5: System comparison on the three datasets of binary classification sub-task.

System	Global Score	RHO(EN)	RHO(FR)	RHO(IT)
Ours(LingJing)	**0.802 (1)**	**0.758 (1)**	**0.841 (1)**	**0.807 (1)**
qiaoxiaosong	0.757 (2)	0.706 (2)	0.805 (2)	0.759 (2)
huawei_zhangmin	0.669 (3)	0.636 (3)	0.740 (3)	0.631 (3)
injySarhanUU	0.221 (5)	0.478 (4)	-0.062 (16)	0.246 (5)
daydayemo	0.206 (6)	0.212 (8)	0.284 (5)	0.121 (9)
aidenqiu	0.205 (7)	0.211 (9)	0.284 (6)	0.121 (9)
Baseline	0.309 (4)	0.265 (6)	0.317 (4)	0.344 (4)

Table 6: System comparison on the three datasets of regression sub-task.

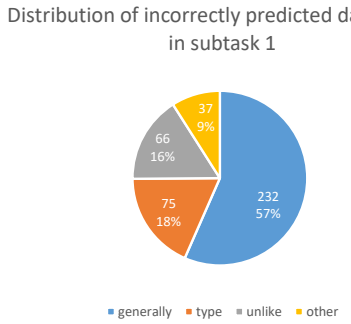

Figure 3: Case study in the sub-task1.

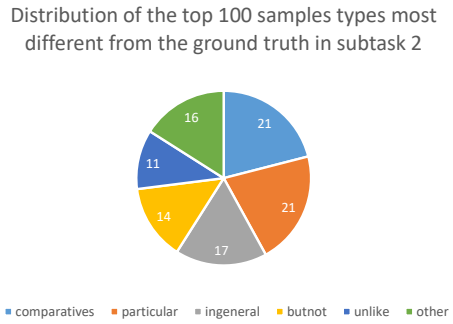

Figure 4: Case study in the sub-task2.

was substantially ahead of second place. Specifically, we earned first place in English, French and Italian in the subtask one classification task, with macroF1 values of 96.988%, 92.817% and 92.714%, respectively. Our global score of 94.173% was 1.863% above second place. For the Subtask 2 regression task, we also came first in English, French and Italian, with RHO scores of 0.758, 0.841 and 0.807, respectively. Our global rank score of Subtask 2 was 0.802, 0.045 above the second-place score.

4.3 Case studies

We counted and analyzed the mispredicted samples, and the distribution of error types is shown in Figures 3 and 4. For subtask 1, we select all the predicted error data for statistics. For subtask 2, we chose the top 100 samples with the most significant difference from the ground truth as the analysis object.

As we can see from Figure 3, the most mispredicted type in the classification task was “generally", with 57%, followed by“type" with 18% and “unlike" with 16%. Our analysis suggests that the reason for the incorrect predictions may be that “generally" sentences are less frequent in common usage and that our model did not have a large enough corpus of similar samples in the previous pre-training phase, thus leading to incorrect predictions.

From Figure 4, we can see that the top 100 data types with the greatest difference from the ground truth on the regression task are more evenly distributed, which means that the model migration is effective for subtask 2.

5 Conclusion

In this paper, we introduce the submitted system to the Semeval-22 task3 PreTENS. Based on the pre-

training DeBERTa-v3, we carry out a simple and effective classification method in sub-task 1 and apply the method of knowledge transferring to sub-task 2. The proposed systems have won first place on both sub-tasks. The experimental results show that our proposed method has better performance than other methods. In addition, we also conducted a number of comparative experiments to further explore the difficulties of the PreTENS task. In the future, we will try to explore more effective methods to perform better semantic taxonomies.

6 Acknowledgement

This work is supported by the Natural Key R&D Program of China (No.2020AAA0106400), the National Natural Science Foundation of China (No. 61922085, No.61976211) and the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006). This research work was supported by the independent research project of National Laboratory of Pattern Recognition, the Youth Innovation Promotion Association CAS and Yunnan Provincial Major Science and Technology Special Plan Projects (No. 202103AA080015).

References

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8):1157–1166. Papers from the Sixth International World Wide Web Conference.

Sean X Chen and Jun S Liu. 1997. Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica*, pages 875–892.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Bill MacCartney. 2009. *Natural language inference*. Stanford University.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for

sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle Vanni, Brian M. Sadler, and Jiawei Han. 2018. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Knowledge Discovery and Data Mining*.

Yichuan Tang. 2013. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1203, Copenhagen, Denmark. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Clark Wissler. 1905. The spearman correlation formula. *Science*, 22(558):309–311.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*.

Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. 2020. Steam: Self-supervised taxonomy expansion with mini-paths. *arXiv: Computation and Language*.

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*.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Knowledge Discovery and Data Mining*.