Cross-lingual Transfer Learning for Irony Detection

Chen Hu
University of Minnesota
huxxx853@umn.edu

Jiaqi Liu
University of Minnesota
liu00687@umn.edu

Keyang Xuan
University of Minnesota
xuan0008@umn.edu

Abstract

Irony is a literary technique that is widely used across languages. A text's ironic intent is defined by its context incongruity. Accurate irony detection is crucial to effective sentiment analysis as well as harassment and hate speech detection in social media. However, detecting ironic statements is tough for a machine as irony enables one speaker or writer to conceal their true intention of negativity under the guise of overt positive representation. In this project, we aim to study this common feature of context incongruity in ironic sentences among different languages and formulate a universal multilingual model which is of paramount importance to increase the overall performance of irony detection. The preliminary result showed that irony detection benefited from mixed language datasets and multilingual models. In order to enhance the model's recognition of context incongruity, we proposed to use prompt tuning as our major technique. By inserting a learnable soft prompt at the beginning of each sentence, the fine-tuning is considered to be more directed to its downstream task. However, our prompt tuning results did not improve the performance significantly. The number of tokens appended and mask construction might have a big impact on our results. Future works should focus on the mask construction so that the soft prompt tuning could function as a hint for model to train.

1 Introduction

The development of the social website has been a rich source of non-literal language such as irony and sarcasm. As a result, research in automatic irony detection has thrived in recent years, for the purpose of better understanding and producing human language. The irony is defined in various ways, but one common agreement on irony identifies it as a figurative language whose actual meaning is different from its superficial meaning (Kong and Qiu, 2011). Accurate irony detection could have a border impact on different research aspects. For example, in order to detect irony accurately, advanced text-mining techniques need to be applied. Besides, failure in irony detection would influence accurate sentimental analysis, and might further affect online harassment detection which has realistic usage in social media.

One of the challenges in irony detection comes from the inconsistency between contextual meaning and literal meaning. Sentiment polarity contrast is a common feature used to determine ironic language (Joshi et al., 2015). For example, in the sentence "I love being ignored", incongruity comes from the contradiction between the positive polarity word "love" and the negative polarity word "ignored". Other features including lexical factors (Kreuz and Caucci, 2007), punctuation marks, and syntactic patterns (Davidov et al., 2010) were investigated in English-based irony detection.

While studies based on English have provided a relatively comprehensive understanding of irony detection, there is still a lack of a systematic map about how irony detection could be applied to non-English language. Even though irony is a common linguistic phenomenon that appears in almost every language, some of its features still vary in different cultures and in structural properties of a specific language (Xing and Xu, 2015; Calvo et al., 2020; Cignarella et al., 2018). For example, words with all capitalized characters typically would be recognized to have non-literal meaning in English (Karoui et al., 2019) while other languages such as Chinese or Japanese do not have such capitalized characters. Chinese irony detection is more challenging because it is either composed of short statements in social media (Li et al., 2019), or it contains widely-used emoji that have unique meanings in the Chinese linguistic environment.

Although there are some papers investigating the models and features specific to Chinese ironic detection, our interest mainly focuses on investigating the pattern similarity between irony sentences of different languages. We proposed a novel application of the multilingual model to learn English and Chinese irony. The rest of the paper is organized as follows: we discussed some related works in Section 2. In Section 3, we describe our overall objective, data resources, and proposed methods in detail. In Section 4, we showed our results in which the performance of monolingual and multilingual models was compared. We also showed our soft prompt tuning results. In Section 5, we concluded our work and discussed some potential failure reasons as well as future directions.

2 Related Work

2.1 Tokenization

One challenge in our project is that the tokenization methods for Chinese and English are different. Unlike English which naturally has space as a sign for tokenization, Chinese sentence as a collection of characters is more ambiguous to segment. The most popular type of tokenization adopted by pretrained-language models(PLM) including Chinese is sub-word tokenization, such as byte pair encoding (BPE) (Sennrich et al., 2016), Word- Piece (Schuster and Nakajima, 2012) and unigram language model segmentation (Kudo, 2018). Apart from sub-word tokenization, a simple character tokenizer is also common to use (Sun et al., 2019). Another difference is that Chinese words do not need to do lemmatization and stemming. That is to say, Chinese characters do not need to deal with tense or plural.

2.2 Irony Detection

The task of irony detection is to classify a piece of text as ironic or non-ironic. Current approaches to irony detection can be classified into three classes, namely rule-based approaches, classical feature-based machine learning methods, and deep neural network models (Zhang et al., 2019). Rule-based approaches generally rely on linguistic features such as sentiment lexicon or hashtags to detect irony (Sulis et al., 2016) while classical feature-based machine learning approaches use hand-crafted features for irony detection, such as sentiment lexicon, subjectivity lexicon, emotional category features, emotional dimension features or structural features (Farias and Rosso, 2017). In this project, we more focus on deep learning-based approaches where (deep) features are automatically derived from texts using neural network models (Zhang et al., 2019).

Within deep learning-based models, some researchers use a pre-trained convolutional neural network for extracting sentiment, emotion and personality features for irony detection (Poria et al., 2016) while other researchers use CNN-LSTM structures for irony detection (Ghosh and Veale, 2017). However, context-based models utilize both content and contextual information required for irony detection, which leads to a growing interest in using neural language models for pre-training for various tasks in natural language processing. Given the highlighted importance of context to capture figurative language phenomena and the difficulties of data annotation, transfer learning approaches such as transformers are gaining attention in various domain adaptation problems. People (Potamias et al., 2020) propose Recurrent CNN Roberta (RCNN-RoBERTa), a hybrid neural architecture building on RoBERTA architecture, which is further enhanced with the employment and devise of a recurrent convolutional neural network. They report performance with an accuracy of 79% on the SARC dataset (Khodak et al., 2017). Similarly, an ensemble of Roberta and Albert on *GetitOffMyChest* dataset (Jaidka et al., 2020) achieve a performance of 85% accuracy with an F1 score of 0.55 (Dadu and Pant, 2020). BERT is used along with aspect-based sentiment analysis to extract the relation between context dialogue sequence and response. They obtain an F1 score of 0.73 on the Twitter dataset and 0.73 on the Reddit dataset (Javdan et al., 2020). However, among all sentimental polarity tests, irony classification shows its hardness, as the overall performance is over 10% lower (Zhang et al., 2019).

Few researchers perform the task of cross-lingual irony detection. It is shown that monolingual models trained separately in different languages using multilingual word representation can open the door to irony detection in different languages (Ghanem et al., 2020). The effectiveness of dependency-based syntactic features is also found in irony detection in a multilingual perspective (Cignarella et al., 2020). However, among all these approaches, modern cross-lingual transformer-based models have seldom been applied and overall performance barely reaches 70%.

It is our objective to study the pattern similarities between irony sentences of different languages and formulate a novel model that increases the accuracy of cross-lingual irony detection.

2.3 Prompt Tuning

Due to the rich knowledge obtained by Pretrained Language Models (PLMs), prompt tuning was proposed by a series of studies to bridge the gap between pre-training objectives and down-stream tasks (Hu et al., 2021; Schick and Schütze, 2020; Liu et al., 2021). Prompt turning shows excellent performance in few-shot learning and zero-shot learning. Among all prompt tuning mechanisms, "p-tuning" or "prompt tuning" learns "soft prompts" to condition frozen language models to perform specific downstream tasks (Lester et al., 2021). These mechanics improved model generalization and avoid over-fitting to a specific task domain. In our cases, we will use prompt tuning as it may explicitly reveal the context incongruity in an ironic sentence and thus improve the model performance. In this case, the multilingual irony patterns can be emphasized and learned with a limited amount of data inputs. The detail is explained in the following section.

3 Problem Formulation

In this section, we present our training objective, general pipeline, and the data resources we use. For the whole project, we follow our pipeline tightly to explore the transfer-ability of irony property between English and Chinese and we aim to improve the performance on the task compared to the baseline model by enhancement techniques such as prompt tuning.

3.1 Training Objective

Even though there is no previous work exploring the transferability of irony detection between two languages, many XLM approaches can be used for this task. Among those, we use XLM-RoBERTa (*XLM-R*) as the base model since it was pretrained on CommonCrawl data which contained more amount of data compared to the Wiki-100 (Wenzek et al., 2020) and *XLM-R* has been proved to have better performance on many cross-lingual tasks comparing to other XLM approaches (Conneau et al., 2020). Our goal is to create a better model architecture, which can beat the performance of *XLM-R* in our cross-lingual irony detection task.

Prompt tuning is recognized as an effective tool for few-shot learning and zero-shot learning tasks. It adapts the downstream tasks by inserting text pieces i.e. template, to the input and transforms a classification problem into a masked language modeling problem. In our case, since the training datasets only have a few thousand examples and we generally think irony pattern such as context incongruity has been learned by large multilingual models, we will utilize prompt tuning to explicitly direct the model to our specific task domain.

3.2 Data Resources

In this case, we pick English and Chinese as the linguistic basis for irony resources. For the English irony resource, we combine the Reddit ironic corpus which is composed of about 2000 Reddit comments (Wallace et al., 2014) and Twitter ironic dataset from SemEval-2018 Task 3 (Van Hee et al., 2018) which contained about 4000 tweets. For the Chinese irony resource, we make a combined irony dataset by both Ciron which is collected from Weibo for irony annotation in simplified-style (Xiang et al., 2020) and NTU irony corpus which consists of messages in traditional Chinese version from the microblogging platform based on emoticons (Tang and Chen, 2014). Also, to keep the same format, we first convert all posts in the NTU irony corpus from the traditional version to the simplified Chinese version. Then, since the original scale in Ciron is from 1 to 5 which corresponds to not irony to strongest irony, we manually re-scale the label from which we convert the original 1,2 labels to -1 in the new dataset representing not irony and we convert the original 4,5 labels to 1 represent irony. For the rest sentences which are labeled as 3, we manually delete those due to their neutral meaning.

3.3 General Pipeline

Fig 1. showed the general pipeline for our training process. For our task, the general pipeline is composed of these steps: Firstly, we train a multilingual model based on either the Chinese training set, English training set, or Mix training set and at the same time train a monolingual model by the same training data. Then the fine-tuning model of the multilingual model will be separately tested on the Chinese testing set and English testing set and its performance will compare to the result from the monolingual fine-tuning model. The perfor-

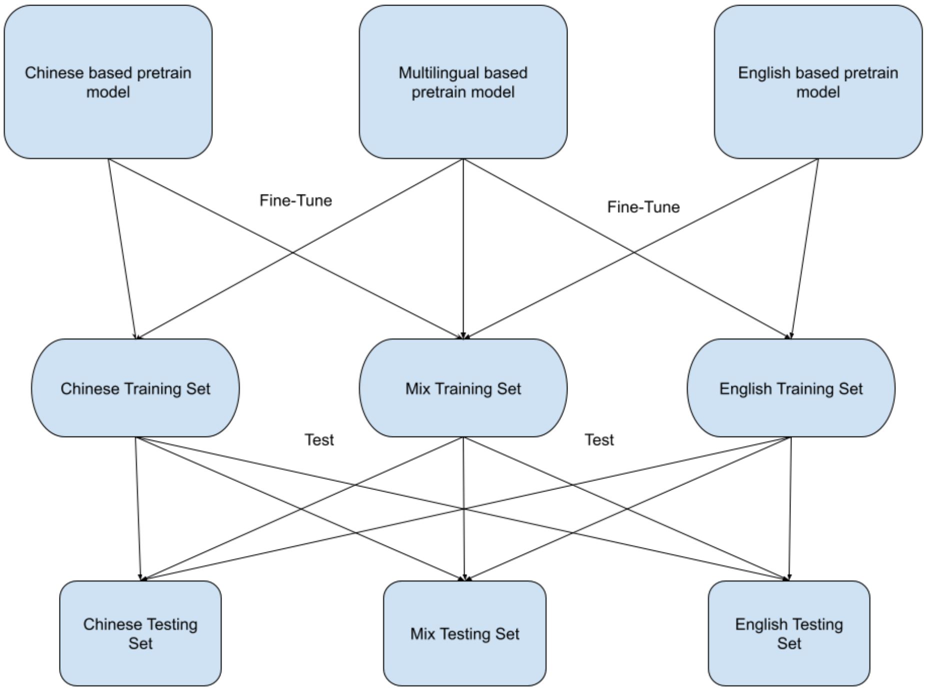

Figure 1: General Pipeline

mance of the monolingual model and *XLM-R* are the benchmark for us to evaluate our self-defined approach's performance.

Fig 2. showed our general pipeline for prompt tuning. We generally followed the algorithm stated in the paper (Lester et al., 2021) and placed a 20-token-lengths prompt in front of each text embedding. The prompt weight was randomly initialized, tuned during training steps, and universally the same for all input sentences.

4 Results

4.1 Preliminary results - Comparison between Monolingual and Multilingual Models

We first started with the experiment via two monolingual models and one multilingual model. The idea is to compare the performance of the monolingual and multilingual models, and whether the models mentioned above could benefit from a mixture use of Chinese and English compared with only using the single dataset. Table 1. summarized the datasets that we used in this experiment.

GTP-2 (Radford et al., 2019) was used to test how a model pretrained by English could detect Chinese irony, English irony, and Mix dataset. We trained GTP-2 on either the English training set or the Mixture training set since it is not reasonable to fine-tune it on the Chinese training set. Similarly,

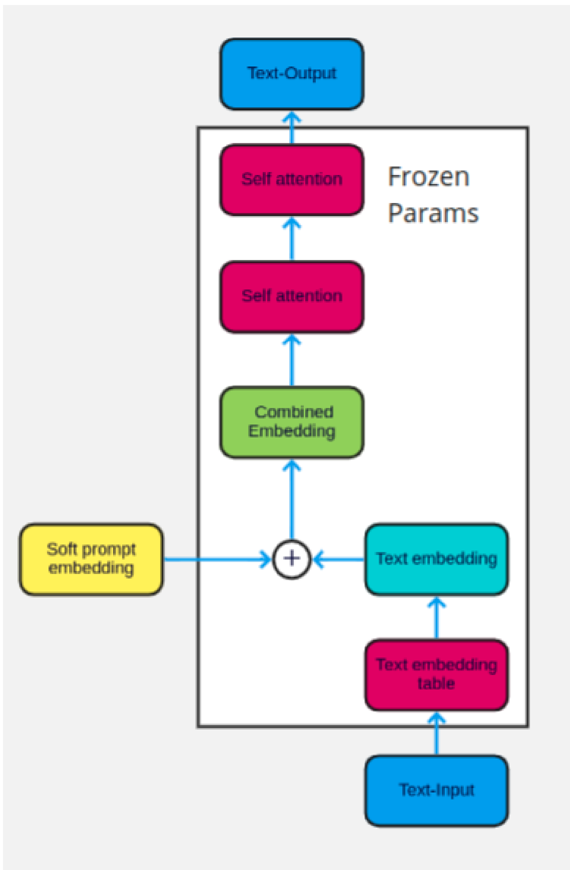

Figure 2: Soft Tuning Flowchart (Lester et al., 2021)

Training set			
	Irony	Non-irony	Total
English	2308	3068	5376
Chinese	1487	5582	7069
Mix	3795	8650	12445
Testing set			
English	441	733	1174
Chinese	389	1379	1768
Mix	830	1112	2942

Table 1: Training and testing dataset size

GPT-2			
	English	Chinese	Mix
English	0.66/0.34	0.78/0.13	0.78/0.13
Mix	0.65/0.28	0.89/0.70	0.79/0.50
Bert-Chinese			
Chinese	0.58/0.13	0.74/0.12	0.70/0.41
Mix	0.64/0.54	0.90/0.77	0.81/0.65
XLM-Roberta			
English	0.70/0.61	0.66/0.41	0.67/0.48
Mix	0.71/0.65	0.89/0.74	0.82/0.68

Table 2: Monolingual and multilingual model results (column - training sets, row - testing sets)

we used Bert Chinese Based model (CPT) (Shao et al., 2021) to test how a model pretrained by Chinese could detect irony in all training sets. We only fine-tuned the CPT model on the Chinese and Mixture datasets. XLM-RoBERTa (Conneau et al., 2019) was the multilingual model we selected for irony detection. It was trained on three datasets and tested on three datasets respectively as well. We used accuracy and F1 score as the evaluation metric. It was notable that due to the application of irony detection, missing an irony case was more harmful than misclassifying a non-irony case. That is to say, the recall was more important in such cases. Thus, we would value the F1 score heavily instead of accuracy.

4.1.1 Discussion

Fig 3. showed the accuracy and F1 score for Chinese irony detection via GPT-2, CPT, and XRoberta models trained on either the English dataset (orange line) or the Mixture dataset (blue line). As mentioned above, there was no result for CPT trained in English. Based on the plot, one thing to notice was that models trained on a mixture dataset outperformed the models trained on the English dataset. The accuracy of Chinese irony detection by XRoberta trained in English slightly decreased compared to GTP-2 models while the F1 score increased on the other hand from 0.12 to 0.50. Since accuracy is the sum of correctly identifying true positive cases and true negative cases, one possible reason could be due to the inability of XRoberta to classify non-irony cases while the recall and precision remained relatively high. Besides, accuracy and F1 score did not change dramatically across different models when the models were trained on mixture data. One possible reason might be due to the information contained in the mixture dataset is sufficient for models to make predictions on Chinese irony detection.

Fig 4. showed the accuracy and F1 score for English irony detection via GPT-2, CPT, and XRoberta models trained on either the Chinese dataset (orange line) or the mixture dataset (blue line). On the contrary with results from models tested on Chinese irony, there is a significant increase in accuracy and F1 scores from monolingual models to multilingual models no matter what training set was used. Similarly, models trained on the mixture dataset had better accuracy and F1 scores compared to the ones trained on the single dataset.

We summarized all our monolingual and multilingual model results in Table 2. In conclusion, we found that the multilingual model would have higher performance than the monolingual model (in the condition of the same training and testing dataset). Besides, even though Chinese and English are two languages with quite different language patterns, when we incorporated more information in the training (either by using the multilingual model or training with mix dataset), the performance was always better. This can indicate that there is a common pattern between irony in those two quite different languages, such as contextual incongruity.

4.1.2 False Cases and Error Analysis

From the testing results, we generated some mistakenly classified examples to see whether there was any potential pattern that could make the model confuse. Table 3. showed a group of falsely classified sentences with their ground label and predicted results. Based on these samples, several patterns are worth mentioning: Firstly, in the sentence "I just love being ignored [smile]", the emoji mark contains a large portion of the ironic sentiment of the whole sentence. However, such kind of expres-

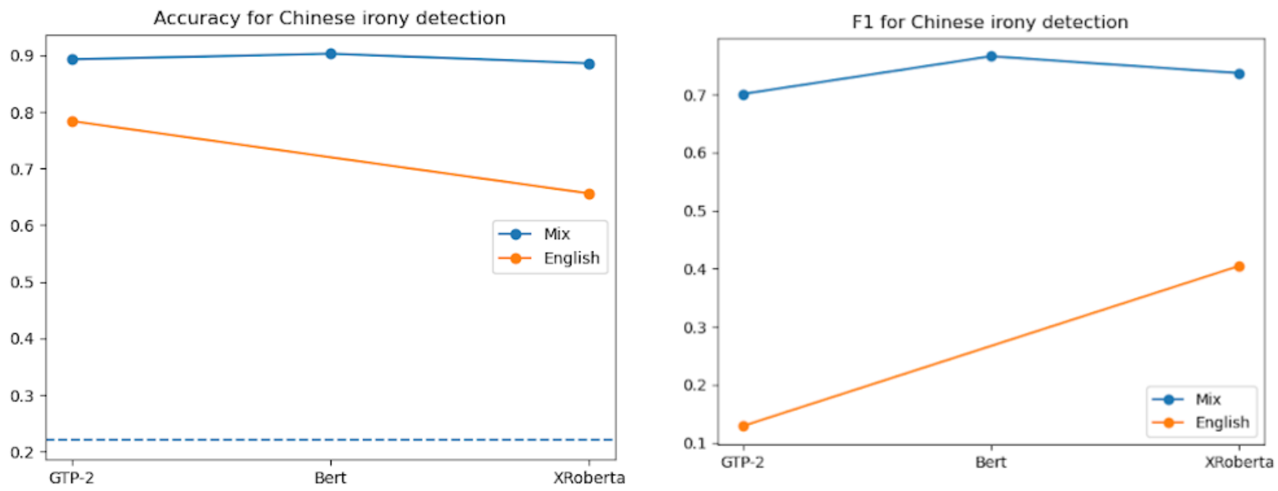

Figure 3: Accuracy and F1 for chinese irony detection

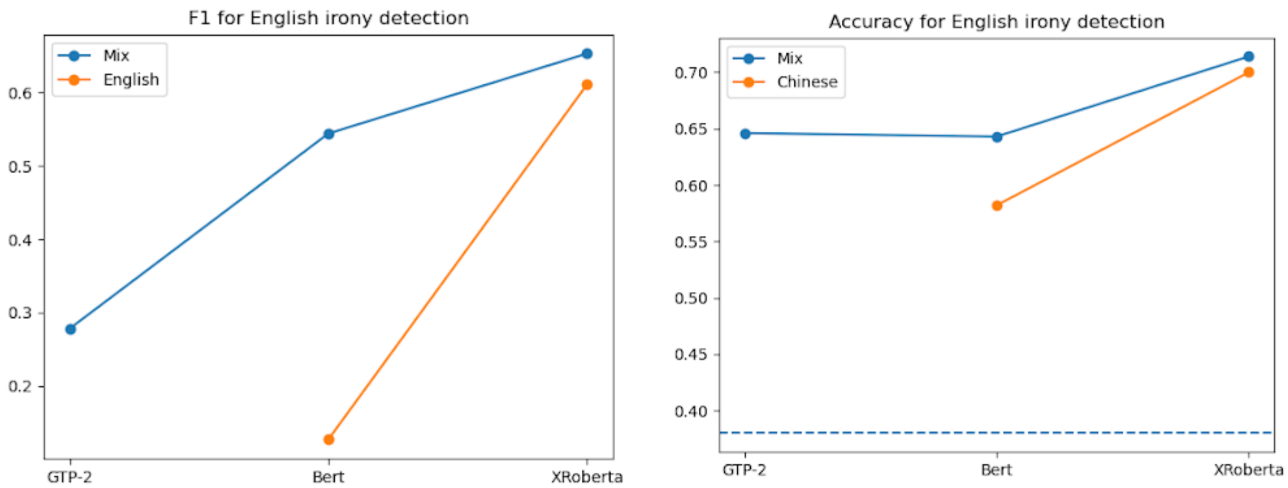

Figure 4: Accuracy and F1 for English irony detection

Error Analysis		
Content	True Label	Predicted Label
I just love being ignored [smile]	1	0
I just drank a healthy, homemade, all-fruit smoothie..in a @Budweiser beer glass#irony	1	0
I am so ready for Monday. #sarcasm	1	0
@GalloSays this game is pathetic. How are they losing this game?	0	1
@robinhosking where did THAT come from?!	0	1
天气可以再热一点没关系 :-&	1	0
这个酒柜的设计如何？真的很棒是吧？	0	1
晚上的电视节目 可以再无聊一点点！ :-&	1	0

Table 3: Failure cases

sion is difficult for the model to learn and capture the meaning which probably is the reason for false prediction. Secondly, post such as "I just drank a healthy, homemade, all-fruit smoothie..in a @Budweiser beer glass" really needs more supplementary evidence or details to help make a correct classification. In detail, the original context of such kind of post delivers trivial patterns related to irony semantics, and most time these posts are affiliated with other elements including pictures or material so we can only detect its meaning correctly only by analyzing these elements together. Thirdly, since the rhetorical question is one of the commonest formats of irony sentiment, our model might be overly sensitive to the question mark. As a result, some general questions without any irony semantic may also be classified into irony.

4.1.3 Limitations

Although the multilingual model trained on a mixture dataset had relatively high accuracy and F1 score, the performance was not what we expected. One limitation is that in the English testing dataset, some hashtags including irony were not deleted. Based on the failure cases shown, it seemed that having irony as a hashtag did not improve the performance. We would still try to avoid such situations in any datasets. Another limitation is that the models still relied heavily on superficial meaning instead of contextual meaning. An example could be the usage of emojis in the testing data mentioned above. One way to improve the understanding of contextual meaning could be grouping together some combinations of words. This could be either done by tokenization or by adding a soft prompt so that the model had the freedom to learn patterns with a small subset of training samples.

4.2 Enhancement by Prompt Tuning

4.2.1 Evaluation in GPT-3 Interface

Due to the limitation and the attempt to use soft tuning, We first explored several prompt tuning techniques including the aforementioned one in GPT-3 interface to verify whether prompt tuning is appropriate for our specific task. To be notifiable, we decided to use failure cases in the preliminary result and checked whether a proportion of them could be recognized by the model when a prompt is added.

According to Fig.5 and Fig.6, we could find that when the input was a single ironic sentence, GTP-3 was unable to correctly identify the property. However, after giving several examples of ironic sentences, there was a certain chance for GTP-3 to identify whether the input sentences are ironic or not. Thus, prompt tuning might be appropriate for our ironic detection tasks.

Check whether the sentence is ironic or not
Input: A member of PETA wears leather shoes.

The sentence is not ironic.

Figure 5: Example of irony detection via single prompt

Check whether the sentence is ironic or not

Example: A marriage counselor files for divorce.
Output: Yes

Example:The police station gets robbed.
Output: Yes

Example: A fire station burns down.
Output: Yes

Input: A member of PETA wears leather shoes.

Yes

Figure 6: Example of irony detection via several prompts

4.2.2 Soft Prompt Tuning

In this step, We utilized the "soft prompt tuning" technique, in which a trainable length of embedding will be added to the front of each input sentence embedding, as depicted in Fig. 2. The prompt weight was randomly initialized, tuned during training steps, and universally the same for all input sentences. We compared the results with a pre-trained fine-tuning model to analyze the efficiency of soft prompt tuning techniques.

In table 4, we showed results on soft prompt tuning with GPT-2 and CPT models. Due to the time limitation and the incompatibility of the model, we did not accomplish testing prompt tuning on the multilingual XLM-Roberta model. Overall, compared to the preliminary results shown in Table 2,

GPT-2			
	English	Chinese	Mix
English	0.66/0.04	0.78/0.21	0.75/0.14
Mix	0.45/0.31	0.69/0.14	0.67/0.11
Bert-Chinese			
Chinese	0.56/0.14	0.79/0.09	0.69/0.08
Mix	0.47/0.2	0.7/0.11	0.64/0.12
XLM-Roberta			
English	/	/	/
Mix	/	/	/

Table 4: Prompt tuning results (column - training sets, row - testing sets)

we found that the general performance with prompt tuning was not better than without it. Among the 12 trials, only CPT model trained on the Chinese set and tested on the Chinese set had a 0.05 improvement in accuracy. We attributed this to insufficient prompt length tuning and inadequate prompt initialization and the detailed analysis is explained in the next section.

5 Future Directions and Conclusions

Unexpectedly, our soft prompting tuning results did not have a significant improvement compared to the preliminary results. One possible explanation could be the number of tokens appended to the original prompts. Based on the previous paper, the number of tokens was a variable that need to be tuned and had a huge impact on model performances. Unfortunately, we spent lots of time on model tuning and training and only had limited time to tune the model hyperparameters. Besides, the way that model initialized the appended tokens would also affect the model performance. The current model randomly assigned token values to the prompt which did not introduce additional useful information to the models in the first place. From Fig.6, we could find that the example prompts need to have the same language structure and patterns as the input prompt in order to have GTP-3 work.

Thus, one future direction that needs to pay more attention to is the initial prompt construction. For example, one prompt tuning technique called composition converts the task into several sub-tasks. For the irony classification task with this technique, we can aggregate the sentence with the prompt such as "This sentence is [Mask] from context. This sentence is [Mask] from meaning. So this sentence is [Mask]". The first and second masks can choose from "positive" and "negative" while the third mask can choose from "irony" and "non-irony". Even though this initialized prompt will change during the training process, based on research in (Lester et al., 2021), the final prompt would still have a high chance to be localized around the initial prompt. In this way, we explicitly formulate the mechanism of detecting irony and expose that to the model. We expect this type of prompt tuning will help generalize the model and improve performance as the soft prompting functions as a hint for our model to train in the process.

To conclude, we found that the multilingual model would generally have higher performance than the monolingual model. Even though Chinese and English are two languages with quite different language patterns, the better performance with mixed datasets indicates that there is a common pattern between irony and different languages, such as contextual incongruity. Despite our prompt tuning techniques did not work as expected, we still think it is a strong model enhancement technique and we want to investigate more in the future.

References

Hiram Calvo, Omar J Gambino, and Consuelo Varinia García Mendoza. 2020. Irony detection using emotion cues. *Computación y Sistemas*, 24(3):1281–1287.

Alessandra Teresa Cignarella, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, and Farah Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. *arXiv preprint arXiv:2011.05706*.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.

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 Asso-*

ciation for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Tanvi Dadu and Kartikey Pant. 2020. Sarcasm detection using context separators in online discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 51–55.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116.

DI Hernández Farias and Paolo Rosso. 2017. Irony, sarcasm, and sentiment analysis. In *Sentiment Analysis in Social Networks*, pages 113–128. Elsevier.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Paolo Rosso, and Véronique Moriceau. 2020. Irony detection in a multilingual context. In *European Conference on Information Retrieval*, pages 141–149. Springer.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:2108.02035*.

Kokil Jaidka, Iknoor Singh, Ni Chhaya, Lyle Ungar, and Jiahui Lu. 2020. A report of the cl-aff offmychest shared task: Modeling supportiveness and disclosure.

Soroush Javdan, Behrouz Minaei-Bidgoli, et al. 2020. Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection. In *Proceedings of the second workshop on figurative language processing*, pages 67–71.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762.

Jihen Karoui, Farah Benamara, and Veronique Moriceau. 2019. *Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media*. John Wiley & Sons.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*.

Lingpeng Kong and Likun Qiu. 2011. Formalization and rules for recognition of satirical irony. In *2011 International Conference on Asian Language Processing*, pages 135–138. IEEE.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, pages 1–4.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

An-Ran Li, Emmanuele Chersoni, Rong Xiang, Chu-Ren Huang, Qin Lu, et al. 2019. On the "easy" task of evaluating chinese irony detection.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, 108:132–143.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Yi-jie Tang and Hsin-Hsi Chen. 2014. Chinese irony corpus construction and ironic structure analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1269–1278, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. 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.

Rong Xiang, Xuefeng Gao, Yunfei Long, Anran Li, Emmanuele Chersoni, Qin Lu, and Chu-Ren Huang. 2020. Ciron: a new benchmark dataset for Chinese irony detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5714–5720, Marseille, France. European Language Resources Association.

Frank Z Xing and Yang Xu. 2015. A logistic regression model of irony detection in chinese internet texts. *Res. Comput. Sci.*, 90:239–249.

Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. 2019. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633–1644.