Comparative analysis of effectiveness of multilingual meta learning and transfer learning in automatic Chinese sarcasm detection

Zhe Han
Uppsala University
Department of Linguistics and Philology
zhe.han.3265@student.uu.se

Abstract

Though there have been various models developed for the purposes of sarcasm detection, the majority of them have been exclusively trained on English datasets collected from Twitter. Compiling a monolingual dataset in English is difficult enough, let alone one in another language. It is essential to develop efficient approaches that can accomplish sarcasm detection in low-resource languages as well, which can be done with the aid of high-resource languages using meta learning and transfer learning. The first experiment explored meta learning using three BERT-based models. Following the use of datasets in five languages in the train task, the Chinese dataset was used in the test task to verify the performance of these models. The second approach looked into transfer learning on the same three BERT-based models as the first, which were subsequently fine-tuned with the English, Korean, Arabic, and multilingual datasets. The performance of the fine-tuned language models was once again evaluated against the Chinese dataset. Both of the two experiments are binary classification tasks. In the end, the results obtained through meta learning slightly outperformed the results achieved through transfer learning.

1 Introduction

Sarcasm is the act of mocking or conveying scorn through the use of language that ordinarily means the opposite. In other words, in order to communicate a negative message, people employ positive language. Obviously, it differs from language to language and is influenced by a range of factors, including culture, gender, and many more. Due to this particularity, automatic sarcasm detection is tough in NLP tasks. Sarcasm Detection is the task of determining whether a sentence is sarcastic or non-sarcastic. Although there has been increased interest in automating sarcasm detection as an NLP task in recent years, which allows for more nuanced analysis of social media texts and other informal textual information that can be harvested from the internet, it is still impossible to acquire sarcasm datasets across all languages. Therefore, it is vital to look into techniques for leveraging the vast amount of training data available, such as English, to assist other low-resource languages. Given this situation, meta learning and transfer learning are potential avenues to explore.

Meta learning is a method for addressing the problem by locating the best initial hyper parameters from several different types of training datasets, each of which consists of only a few data samples. In this way, the model trained by meta learning method can be well generalized to the target task after learning a few target data examples. Namely, it lays emphasis on learning how to learn rather than attempting to attain the ultimate parameters for the target task. (Ravi and Larochelle, 2016; Yin, 2020) Meta learning has recently been shown to be useful for a variety of machine learning tasks as well. Nooralahzadeh et al. (2020) introduced a cross-lingual meta learning framework that demonstrates effective performance in NLI and QA tasks. Besides, van der Heijden et al. (2021) pioneered a meta learning method to implement document classification in low-resource languages, showing great effectiveness after few-shot learning.

The employment of a previously trained model on an unseen but related task is defined as transfer learning. It has become mainstream right now that sophisticated neural networks are trained by this method. Cross-lingual transfer is beneficial for boosting performance in low-resource languages (Ruder, 2017). Multilingual transfer learning experiments can thus be carried out by fine-tuning several cutting-edge models to investigate if the experiments will achieve effective transferability.

According to the final results, the Multilingual BERT performed the best in meta learning, while the SentiBERT performed the worst. In transfer learning, DistilBERT, fine-tuned with Arabic, gives the best performance. Multilingual transfer languages do not exceed all monolingual transfer languages as expected. Furthermore, English is not the best transfer language across all models. Moreover, SentiBERT outperformed MultiBERT across all transfer languages. Generally, meta learning marginally surpassed transfer learning.

2 Related work

In a bid to further deal with this problem, researchers in succession rolled out various advanced BERT-based (Devlin et al., 2018) models. For example, Alexandros Potamias et al. (2019) designed Recurrent CNN RoBERTa (RCNN-RoBERTa), a hybrid neural architecture built on the RoBERTa architecture, which is further enhanced with the employment and devise of a recurrent convolutional neural network for text-based sarcasm detection tasks and reported an accuracy of 79% on the SARC dataset (Nimala et al., 2021). Moreover, Pant and Dadu (2020) used an ensemble of RoBERTa and ALBERT and got even higher accuracy at 85%. With an increasing number of tweets containing pictures being posted online, several scientists have started to take image feature extraction into consideration. Cai et al. (2019) introduced a Hierarchical Fusion Model that recognizes text features, image features, and image attributes as three modalities, then rebuilds features of those modalities and deeply merges them into one feature instead of simply concatenating them. Inspired by the idea of self-attention mechanism (Vaswani et al., 2017), Pan et al. (2020) focused on designing an inter-modality attention to capture both intra and inter-modality incongruity for multi-modal sarcasm detection.

Except for the superior models described above, Bamman and Smith (2015) achieved advances in accuracy in the detection of this complicated phenomenon by using additional attributes of the author, the audience, and the past tweets of the author as opposed to purely linguistic features. Similarly, Rajadesingan et al. (2015) figured out the sarcasm detection task by leveraging behavioral traits identified in the past tweets of users to construct a behavioral framework. Nonetheless, few experiments have been carried out on sarcasm detection via multilingual learning techniques. Because of this, as well as the deficiency of image sarcasm datasets, this paper probed into the effectiveness of multilingual meta learning and transfer learning in the binary text-based sarcasm classification task.

3 Experiments

3.1 Model selection

When it comes to models, it is well known that fine-tuned Multilingual BERT (Pires et al., 2019) for NLP tasks in a source language by researchers yields outstanding transferability in another target language without having to observe any supervised data. However, this large, pre-trained model may be source-consuming. In this case, Sanh et al. (2019) proposed a method to pre-train a smaller model, called DistilBERT, which can then be fine-tuned with excellent outcomes on NLP tasks similar to the larger version of it. There is also a sentiment analysis model invented by further fine-tuning Multilingual BERT with product reviews in six languages (English, Dutch, German, French, Spanish, and Italian), which could help with the sarcasm detection task because sarcasm judgment is part of sentiment analysis, and the product review corpus used to fine-tune the model could contain negative sarcastic comments. In this work, these three models were fine-tuned in both meta learning and transfer learning, and all of them were downloaded from HuggingFace[1]. The sizes of these models are presented in Table 1.

Model	Layers	Hidden Size	Heads
Multi	12	768	12
Distil	6	768	12
Senti	6	768	12

Table 1: Sizes of selected models.

3.2 Meta learning

3.2.1 Train and Test tasks

Under normal circumstances, the meta learning process consists of the train task and the test task, each of which should encompass numerous tasks across different languages. As an example, Gu et al. (2018) gathered datasets in 18 languages within the train task. This paper only managed to collect annotated sarcasm datasets in 5 languages

[1] https://huggingface.co/models

because of the rarity of the data source. As a result, Arabic, English, Korean, Spanish, and Turkish datasets were included in the train task. Finally, the Chinese dataset was selected for few-shot learning and composing the test task.

3.2.2 Algorithm selection

Finn et al. (2017) proposed an algorithm called Model-Agnostic Meta Learning (MAML) to upgrade the weights during the training process. Despite its successful employment in NLP applications, such as few-shot text classification (Liu et al., 2020), it is pretty time-consuming due to its complex optimization process containing outer loop as well as inner loop gradient updates, which renders it vulnerable to neural network architectures, often resulting in malfunctions during training (Antoniou et al., 2018). What's worse, each sub-task within the train task should be divided into support and query sets (Li et al., 2021), with the support set being used for gradient descent and the query set functioning to minimize the loss.

To tackle this issue, Nichol and Schulman (2018) designed an upgraded algorithm called Reptile without calculating the second derivative gradient, which is more effective for the optimization process without weakening the performance. Furthermore, sub-tasks inside the whole train task no longer need to be split into different sets. In conclusion, the reptile algorithm serves to optimize the weights of models.

3.2.3 Training strategy

The code from GitHub[2] was employed to implement the task. It was originally designed for multi-language classification, but snippets of code were modified to be suitable for this experiment. For instance, the code is based on the BERT model. While segment_ids are not included in the DistilbertTokenizer, this part was excluded when fine-tuning the DistilBERT model.

The original optimizer Adam (Kingma and Ba, 2014) was replaced by AdamW, which is more computationally efficient (Loshchilov and Hutter, 2017). The original hyper parameters were used to acquire the highest scores.

Meta learning seeks to obtain optimal initial parameters rather than final parameters. After updating these initial parameters by observing a few target data samples, the model will be well adapted to the target task, implying that exploring zero-shot in this experiment is pointless.

3.3 Transfer learning

3.3.1 Transfer language selection

In the major transfer learning experiments, English was primarily selected as the default source language. However, Turc et al. (2021) reexamined the priority of English in zero-shot cross-lingual transfer, concluding that English is not always the optimal source language. Even if the source and target languages are not from the same language family tree, they can still possibly yield better results than English as the source language. Lin et al. (2019) stated that multilingual datasets are much better at transferring languages than a single language. As a result, in each different-shot experiment, the Arabic, English, Korean, and multilingual datasets were respectively utilized as the transfer languages to examine which one would be the best candidate. The Spanish and Turkish datasets were not chosen as the monolingual transfer languages to conduct the experiment, which will be elaborated on in section 4.2.

3.3.2 Fine-tuning strategy and testing

Kovaleva et al. (2019) corroborated that on many tasks, just the last few layers change the most after the fine-tuning process in transfer learning and Lee et al. (2019) found that fine-tuning all layers does not always help. In consideration of this corroborative evidence, only one classification layer was concatenated at the end of all pre-trained models, with the rest being frozen during the fine-tuning process in this work.

Gupta et al. (2020) discovered that zero-shot classification, without further fine-tuning on few-shot domains, performs equivalently to few-shot classifications. Hence, both zero-shot and few-shot classifications were performed to investigate which method could deliver higher accuracy. With regard to hyper parameters, the default ones suggested by HuggingFace were simply inherited, which proved to be optimal for fine-tuning. The same Chinese dataset applied to the meta learning was used to implement few-shot learning and testing.

[2] `https://github.com/mailong25/meta-learning-BERT`

Lan	Sarcastic	Non-sarcastic	Total
Ar	745	2,357	3,102
En	867	2,601	3,468
Ko	4,493	4,507	9,000
Es	90	869	959
Tu	100	100	200
Zh	436	3,578	4,014

Table 2: The distribution of all original datasets.

4 Dataset

4.1 Dataset source and description

Both Arabic and English datasets are provided by the shared task of iSarcasmEval: Intended Sarcasm Detection in English and Arabic[3]. The organizer of this task introduced a new data collection method where the sarcasm labels specifying the sarcastic nature of texts are provided by the authors themselves, thus eliminating labelling proxies (in the form of predefined tags, or third-party annotators). The Korean, Spanish, and Turkish datasets are all collected from Kaggle[4] and Github[5] [6]. The Chinese dataset named Ciron is provided by Xiang et al. (2020), which is from the posts of Weibo, a Chinese micro-blogging platform, and is classified into five fine-grained labels: 1 (not ironic), 2 (unlikely ironic), 3 (insufficient evidence), 4 (weakly ironic), and 5 (strongly ironic). Table 2 shows the distribution of all original datasets.

4.2 Dataset reconstruction

Meta learning merely needs a small number of training texts in various languages, as discussed in section 1. However, since this research only collected 5 languages, for the sake of compromise, each sub-task within the train task was reconstructed to simply include 200 texts, with the proportion of sarcastic texts and non-sarcastic ones being as close as possible to 1 to 1 (the Spanish dataset only includes 90 sarcastic texts). The total size of the train task is up to 1,000.

For more direct comparison to the meta learning experiment, the identical datasets inside the train task of meta learning were used as multilingual transfer languages in transfer learning. As for monolingual transfer languages, the total size of each transfer language dataset was downsized to 1,000, with sarcastic texts and non-sarcastic ones evenly distributed as well. Because there are insufficient sarcastic texts in the Spanish and Turkish datasets, they were not chosen as the monolingual transfer languages to conduct the experiment.

This study concentrates solely on the binary classification task. As a result, the Ciron Chinese dataset was turned into a binary dataset, with text marked as labels 4 and 5 merged into 1, text marked as labels 1 and 2 merged into 0, and text marked as label 3 omitted. Due to the extreme imbalance in the original dataset, a new dataset with 436 sarcastic texts and 436 non-sarcastic texts was reconstructed, 80 of which were uniformly sampled for few-shot learning, with the remaining being for final testing. Table 3 and Table 4 show the reconstructed datasets for meta learning and transfer learning, respectively.

4.2.1 Dataset preprocessing

The Ciron Chinese dataset includes hashtags that indicate special meaning instead of labeling texts as sarcasm, unlike tweets. There is no need to remove them. They may include semantic features. Subramanian et al. (2019) put forward that emojis are widely used as emotion signals, which has great potential to advance sarcasm detection. Accordingly, Bashmal and AlZeer (2021) decoded the emojis and their emotions by using the emoji descriptions data online. In this paper, the emoji module was used to convert emojis in tweets and posts into English textual descriptions, which were then translated into other languages by the Google Translate API automatically.

HuggingFace provides the built-in tokenizers for each BERT-based model that can handle case changing, vocabulary construction, word segmentation, punctuation segmentation, and so on.

5 Results and Analysis

All results were measured by total accuracy and the marco average F1 metrics. Table 5 summarizes the meta learning results. It reveals that the Multilingual BERT exceeds the performance of others with a total accuracy and a macro avg F1 of 57.8%. Most likely, the more diverse languages exist in the train task, the more complicated model architecture is required to extract multiple features from

[3] https://sites.google.com/view/semeval2022-isarcasmeval

[4] https://www.kaggle.com/mikahama/datasets

[5] https://github.com/SpellOnYou/korean-sarcasm

[6] https://github.com/teghub/Turkish-Irony-Dataset

Lan	Sarcastic	Non-sarcastic	Total
Ar	100	100	200
En	100	100	200
Ko	100	100	200
Es	90	110	200
Tu	100	100	200
Total-train-task	490	510	1,000
Zh-test-task	396	396	792
Zh-few-shot-learning	40	40	80

Table 3: Reconstructed datasets for meta learning.

Lan	Sarcastic	Non-sarcastic	Total
Ar	500	500	1,000
En	500	500	1,000
Ko	500	500	1,000

Ar+En+Ko+Es+Tu	100(ar)+100(En)+100(Ko)
	+90(Es)+100(Tu)
		100(ar)+100(En)+100(Ko
		)+110(Es)+100(Tu)
			1,000

Zh-test	396	396	792
Zh-few-shot-learning	40	40	80

Table 4: Reconstructed datasets for transfer learning.

Metrics	**Multi**	Distil	Senti
accuracy	**0.578**	0.574	0.559
marco avg F1	**0.578**	0.571	0.536

Table 5: Results of few-shot meta learning across three pre-trained BERT-based models.

languages as much as possible. The DistilBERT performs pretty close to the Multilingual BERT, despite having a more concise architecture. Besides, SentiBERT tends to be the worst model in meta learning, indicating that meta learning may demand train and test tasks be in exactly the same domain, otherwise the effectiveness of meta learning may worsen.

In zero-shot transfer learning, every fine-tuned model tends to label all texts as non-sarcastic, which reveals that zero-shot does not generalize well in transferring sarcasm features from other languages to Chinese sarcasm. Sarcasm expressions may vary in different languages and datasets are sourced from different media, which could cause the failure of the sarcasm feature transfer in zero-shot transfer learning.

On the contrary, few-shot transfer learning indeed yields more powerful outcomes. The results are reported in Table 6. DistilBERT, fine-tuned with Arabic, delivers the highest total accuracy of 57.4% and a macro avg F1 of 56.3%, demonstrating that less complex BERT-based models can prevail over more complex ones. It is surprising that the Arabic approach outperforms others across DistilBERT and SentiBERT, given that Arabic and Chinese are from entirely different language families. More research is required to explain it. In addition, Korean exceeds the rest in Multilingual BERT, along with showing better performance than English across MultiBERT and SentiBERT. Historically, there were a certain number of loanwords from Chinese Pinyin in Korean, which could be the reason why Korean is a better transfer language than English. Another finding is that, except in SentiBERT, multilingual transfer languages tend not to be the top performers as anticipated. Because each language has its own complex sarcastic features, adding more languages could result in a lot more noise. In contrast to meta learning, SentiBERT gives better results in transfer learning than Multilingual BERT, proving that more general sentiment analysis data examples may contribute to sarcasm detection in transfer learning.

Meta learning appears to be better at extracting abstract sarcastic linguistic features from multiple languages than transfer learning. Meta learning is a broader mode that is devoted to the representation and acquisition of "meta-knowledge." This meta-knowledge is defined as generic knowledge

Model	Metrics	Ar	En	Ko	Ar+En+Ko+Es+Tu
Multi	accuracy	0.524	0.489	0.527	0.508
	marco avg F1	0.521	0.400	0.527	0.508
Distil	accuracy	**0.574**	0.572	0.569	0.561
	marco avg F1	**0.563**	0.557	0.557	0.548
Senti	accuracy	0.552	0.504	0.545	0.556
	marco avg F1	0.546	0.504	0.545	0.554

Table 6: Results of few-shot transfer learning across three pre-trained BERT-based models.

about a broad range of languages that can be acquired through training. It has a high level of representational ability that can generalize into other languages, whereas transfer learning directly applies what is learned from the source language to the specific target language.

6 Conclusion and Future work

To conclude, two strategies show good effectiveness in the Chinese sarcasm detection task. However, their scores are rather close, so more trials are needed to determine which strategy is more effective. In this work, BertTokenizer was employed to tokenize all languages. Arguably, it is significant to attempt different tokenization methods, like morphological segmentation for chronologically rich languages like Turkish or the Jieba tokenizer for Chinese, which can be segmented by meaning units rather than single tokens, to investigate if these can influence the final result. In the end, this research might as well be extended to improve the performance of the model by adding more multilingual sarcastic datasets from diverse media in the meta learning train task to fine-tune a model for multilingual low-resource languages sarcasm detection, not only for Chinese.

References

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2019. A transformer-based approach to irony and sarcasm detection. *arXiv e-prints*, pages arXiv–1911.

Antreas Antoniou, Harrison Edwards, and Amos Storkey. 2018. How to train your maml. *arXiv preprint arXiv:1810.09502*.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth international AAAI conference on web and social media*.

Laila Bashmal and Daliyah AlZeer. 2021. Arsarcasm shared task: An ensemble bert model for sarcasmdetection in arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 323–328.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

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*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. 2018. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

Aakriti Gupta, Kapil Thadani, and Neil O'Hare. 2020. Effective few-shot classification with transfer learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1061–1066.

Niels van der Heijden, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2021. Multilingual and cross-lingual document classification: A meta-learning approach. *arXiv preprint arXiv:2101.11302*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning. *arXiv preprint arXiv:1911.03090*.

Xiaoxu Li, Zhuo Sun, Jing-Hao Xue, and Zhanyu Ma. 2021. A concise review of recent few-shot meta-learning methods. *Neurocomputing*, 456:463–468.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*.

Zequn Liu, Ruiyi Zhang, Yiping Song, and Ming Zhang. 2020. When does maml work the best? an empirical study on model-agnostic meta-learning in nlp applications. *arXiv preprint arXiv:2005.11700*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Alex Nichol and John Schulman. 2018. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4.

K Nimala, R Jebakumar, and M Saravanan. 2021. Sentiment topic sarcasm mixture model to distinguish sarcasm prevalent topics based on the sentiment bearing words in the tweets. *Journal of Ambient Intelligence and Humanized Computing*, 12(6):6801–6810.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. *arXiv preprint arXiv:2003.02739*.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1383–1392.

Kartikey Pant and Tanvi Dadu. 2020. Sarcasm detection using context separators in online discourse. *arXiv preprint arXiv:2006.00850*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jayashree Subramanian, Varun Sridharan, Kai Shu, and Huan Liu. 2019. Exploiting emojis for sarcasm detection. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 70–80. Springer.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.

Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey. *arXiv preprint arXiv:2007.09604*.