DOCTORAL THESIS

Exploring Optimal Settings for Machine Translation of Irony with Application to Multilingual Irony Detection

皮肉文の機械翻訳における最適設定に関する考察及び
多言語皮肉検出への応用

by

Chia Zheng Lin

Advised by:
Michal Ptaszynski
Fumito Masui
Okumura Takashi

KITAMI INSTITUTE OF TECHNOLOGY
GRADUATE SCHOOL OF ENGINEERING

September 2024

Acknowledgements

I wish to express my sincere appreciation to my supervisor, Associate Professor Michal Ptaszynski, for his dedicated guidance and enlightening inspiration in defining the path of my research. He convincingly guided and encouraged me to be professional and always emboldened me with great ideas and amazing thoughts. Without his persistent help and valuable advice, the study would not have been completed successfully.

I would also like to thank Professor Masui Fumito for his unfailing support and continuous encouragement throughout my years of study. I am gratefully indebted to his precious comments and insight through the process of researching and writing this thesis.

Finally, I must express my special regards to my parents and to my friends for being with me throughout the present and past. This accomplishment would not have been possible without them.

September 2024

ABSTRACT

In this paper, we investigate sarcasm and irony as seen through a novel perspective of machine translation. We employ various techniques for translation, comparing both manually and automatically translated datasets of irony and sarcasm. We first clarify the definitions of irony and sarcasm and present an exhaustive field review of studies on irony both from purely linguistic as well as computational linguistic perspectives. We also propose a novel evaluation metric for the purpose of evaluating translations of figurative language, with a focus on machine-translated irony and sarcasm. The constructed English and Chinese parallel dataset includes polarized content from tweets as well as forum posts, categorized by irony types. The preferred translation model, mBART-50, is identified through a thorough experimental process. Optimal translation settings and the best-finetuned model for irony are explored, with the most effective model being finetuned on both ironic and non-ironic data. We also experimented which types of irony are best suitable for training in this specific task - short microblogging messages or longer forum posts. Moreover, we compare the capabilities of a well fine-tuned mBART to a prompt-based method using the recently popular ChatGPT model, with the conclusion that the former still outperforms the latter, although ChatGPT without any training can be considered as a “good enough" ad hoc solution in the case of a lack of data for training. Finally, we verify if the translated data - either manually, or with an MT model - can be used as training data in a task of irony detection. We believe that the presented research can be expanded into languages other than the presented here Chinese and English, which together with the ability to detect various categories of irony, could contribute to deepening the understanding of figurative language, especially irony and sarcasm.

ABSTRACT IN JAPANESE (論文内容の概要)

この論文では、機械翻訳の新しい視点から見た皮肉と風刺について調査します。さまざまな翻訳技術を使用して、皮肉と風刺の手動および自動翻訳データセットを比較します。まず、皮肉と風刺の定義を明確にし、純粋に言語学的および計算言語学的観点からの皮肉に関する研究の徹底的な分野レビューを提示します。また、比喩的な言語の翻訳を評価するための新しい評価指標を提案し、機械翻訳された皮肉と風刺に焦点を当てます。構築された英語と中国語の並列データセットには、ツイートやフォーラム投稿からの極性のあるコンテンツが含まれ、皮肉のタイプごとに分類されています。徹底的な実験プロセスを通じて、推奨翻訳モデルであるmBART-50が特定されました。皮肉に最適な翻訳設定と最高に微調整されたモデルを探索し、最も効果的なモデルは皮肉と非皮肉のデータの両方で微調整されました。また、この特定のタスクでのトレーニングに最適な皮肉のタイプ（短いマイクロブログメッセージや長いフォーラム投稿）を実験しました。さらに、最近人気のあるChatGPTモデルを使用したプロンプトベースの方法と、よく微調整されたmBARTの能力を比較し、後者が依然として前者を上回るが、訓練データが不足している場合には、訓練なしでもChatGPTが「十分に良い」アドホックソリューションと見なすことができるという結論に達しました。最後に、手動またはMTモデルで翻訳されたデータが皮肉検出タスクのトレーニングデータとして使用できるかどうかを検証します。提示された研究は、ここで提示された中国語と英語以外の言語に拡大できると信じており、さまざまなカテゴリーの皮肉を検出する能力と相まって、比喩的な言語、特に皮肉と風刺の理解を深めることに貢献できると考えています。

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As digital communication becomes increasingly prevalent, online posts can be viewed in any corner of the world by users of various languages. Especially, social media and online forums are melting pots of varied expressions, including nuanced forms of irony. However, although social media platforms provide machine-translated versions of such online messages, inaccurate translations of specific posts can lead to misunderstandings, conflicts, or even potentially to cases of cyberbullying. Unfortunately, due to the brief and informal nature of digital communication, translating irony accurately is a challenge. Therefore, accurately translating ironic expressions becomes imperative.

This paper delves into the complexities of machine-translating of irony, emphasizing the distinct features of social media text and online forum posts. Building on our prior research in irony detection [29, 30, 31], In this paper, we examine the language intricacies of internet irony, shedding light on how this impacts wider digital communication between different cultural regions and languages. Specifically, we first thoroughly explore the previous studies on both machine translation as well as irony and sarcasm, delineating their definitions and conducting an exhaustive field review spanning linguistic and computational linguistic perspectives. Moreover, a novel evaluation metric tailored for assessing translations of figurative language, particularly machine-translated irony and sarcasm, is proposed. Secondly, we construct a comprehensive English-Chinese parallel dataset comprising content sourced from tweets and forum posts, meticulously translated and categorized according to irony types.

The identification of the optimal translation model is performed through an exhaustive experimental process. We explore the optimal translation settings for the models fine-tuned for irony translation, culminating in the identification of the most effective model, fine-tuned on both ironic and non-ironic data. Additionally, we empirically investigate which types of irony prove most effective for training in this specific task, scrutinizing short microblogging messages versus longer forum posts.

Furthermore, a comparative analysis between an optimally fine-tuned mBART model and a prompt-based approach utilizing ChatGPT is conducted, revealing the superior performance of the former. However, it is noteworthy that ChatGPT, despite lacking specific training for irony, demonstrates promising efficacy, serving as a potential ad hoc solution in data-scarce scenarios. Subsequently, we evaluate the viability of utilizing translated data, whether manually or through an MT model, as training data in the task of irony detection, verifying the applicability of machine translation models as a means for data augmentation in case of data-scarce scenarios, such as for low-resource languages. Such endeavors, coupled with the ability to discern various categories of irony, hold promise for deepening our comprehension of figurative language, particularly irony and sarcasm.

In the dynamic landscape of natural language processing, the interaction between machine translation and irony presents a novel and compelling challenge, particularly within the context of social media and internet forums. Leveraging the lessons from irony detection research, our aim was to go beyond the task of understanding of irony to the task of translating it into other languages while attaining to those cultural differences. This could also enhance machine translation in general to convey not only literal meaning but also the subtle nuances of figurative expressions, especially present in online exchanges containing ironic meanings.

The significance of this research goes beyond pure analysis of such language, as intercultural understanding emerges as a critical factor in translating irony effectively. We aim to bridge the gap between understanding and translating ironic content in the digital domain. Our work not only hints at broader implications, potentially contributing to practical application in areas like toxic language detection, but also highlights the significance of cultural understanding, paving the way for enhanced cross-cultural communication.

As we extend our focus to widen the coverage of irony-related studies into multiple languages, the presented study does not only narrow the linguistic gap but also fosters a better understanding of cultural differences involved in irony understanding in languages other than English. This sensitivity to cultural differences has the potential to enhance cross-cultural communication, and mitigate misunderstandings, thus improving the quality of online interactions in general.

1.1 Research Background

Our motivation for this research primarily stems from the need for improving machine translation systems, especially, how machine translation handles figurative expressions. However, since it would be impractical to attempt to solve the translations of all types of figurative expressions, in this paper in particular we focus on translating irony. Moreover, as online communication becomes increasingly vital, we have observed instances where current translation systems commonly applied on various social media platforms, struggle to accurately convey the nuances of everyday conversations, specifically figurative expressions, which in large part also include ironic expressions. This means that the inability to properly translate irony can directly lead to potential misunderstandings and communication gaps, and in effect, to conflicts among users, also developing into cases of cyberbullying. For example, Mckenna[100] showed that online communication can amplify aggression and the potential for humiliation. This has been shown to be particularly relevant in the context of teenage users [60].

With this research, we aim to address instances where machine translation falls short, especially in the interpretation of irony. The goal is to refine language models, improve precision, and reduce instances of miscommunication in cross-language interactions. Recognizing that language is deeply tied to culture, our motivation includes the commitment to make machine translation more culturally sensitive. We want to ensure that translations not only capture the literal meaning but also respect the cultural nuances embedded in ironic expressions. Beyond theoretical improvements, we are motivated by the tangible impact our work can have on real-world issues. The prevalence of cyberbullying and toxicity in online spaces underscores the need for machine translation systems that are better aware of the implicit intents of the users, thus contributing to creating a safer digital environment. Our goals align with a meticulous fine-tuning approach, leveraging a new dataset, and a novel evaluation metric, while being informed by the previous literature.

Thus, with regard to the previous studies described in the above sections, the specific goals of this paper are the following.

- Performance enhancement through Language Model fine-tuning:

  The primary goal of the research is to optimize language models for improved

performance in translating irony. We approach this by fine-tuning models through various experiments with different setups and combinations of data, especially, utilizing our new parallel dataset, as well as drawing insights from the rich literature on machine translation.

- Improved figurative language comprehension:

  We aim to deepen the understanding of figurative language, specifically in terms of irony and its types, by refining language models and creating high-quality datasets. The objective is to conduct a series of experiments delving deeper into categorizing irony, with the potential of teaching the models to discern the layers of meaning within different types of irony.

- Creation of new dataset for irony translation:

  We also aim to develop a new, dedicated dataset for the task of translation of ironic sentences. This dataset will serve as a valuable resource for training and evaluating language models, contributing to the robustness and real-world applicability of our research.

- Proposal of novel evaluation metric for Machine Translation:

  Proper evaluation of machine translation outputs has been one of the main problems in the field of machine translation (MT). Although novel methods are being proposed each year, none of them can be considered as fully satisfying. Therefore, our goal here was to validate a novel evaluation metric specifically designed for assessing the translation of ironic content. This metric aims to overcome the limitations of existing metrics, especially those that are overused yet known for their bad quality, such as BLEU, or CHRF, yet also acknowledging the situations where such simple methods happen to work better than the recently popular language model-based methods, like COMET, or BERTScore, thus providing a more accurate measure of language model performance in translating irony and other figurative content.

- Extensive analysis and comparison of various model types:

  Final goal is to conduct a comprehensive analysis and comparison of different types of language models, including various standard language models, and

the popular prompt-based models. This exploration aims to uncover insights into the strengths and weaknesses of each type, particularly in the context of translation of ironic content.

These specific goals maintain a strategic and focused approach to advancing the field of machine translation in handling figurative language, with an added emphasis on exploring and understanding the overall effectiveness of language models in the area of machine translation of figurative language.

1.2 Thesis Outline

Chapter 2 presents the literature review for all of the past works the topics related to this study, namely in machine translation, evaluation metrics in machine translation, as well as irony and sarcasm, in general, and its appearance in the field of computational linguistics, which mostly represents automatic detection of irony and sarcasm. Next, Chapter 3 presents the datasets applied in this study, which includes both previously created datasets, as well as the dataset created and especially for this study. Later, in Chapter 4 we present the applied methods, which include both language models, as well as all the evaluation metrics, including our proposed metric which we designed specifically for a more comprehensive comparison of machine translation models trained and applied in this paper. Afterward, in Chapter 5, we present all of the experiments performed for this study, which consists of seven different experiments we conducted to thoroughly understand the optimal conditions for irony translation. Lastly, Section 5.8 presents an additional experiment, in which we evaluate the applicability of translation models in data augmentation for other tasks, in this case, for irony detection. We also discuss the broader implications of this study, its future applicability, as well as potential ethical considerations in Chapter 6 of the paper, before concluding the paper in Chapter 7.

Chapter 2

Literature Review

In this chapter, we will discuss past studies in related fields, including figurative language, machine translation, and evaluation metrics.

2.1 Machine Translation

2.1.1 Early Machine Translation

The first appearance of machine translation can be dated back to 1939 when the Academy of Sciences of USSR had been approached by Petr Petrovic Troyanskii with proposals for mechanical translation. However, with somewhat fruitless discussions, the proposals were never worked upon [70]. Thenceforth in 1947, one year after the first computer, ENIAC machine, was developed, Warren Weaver [153] suggested the use of computers for language translation tasks, and hence marked the beginning of extensive research into machine translation.

Some of the machine translation-related early works include the Russian-English machine translation experimental system which was developed in collaboration with IBM and Georgetown University. The public demonstration of the experiment in 1954 ignited much controversy and public interest [70]. However, after years of research without a significant breakthrough, the Automatic Language Processing Advisory Committee (ALPAC) published a report in November 1966 ending the substantial funding of machine translation research in the United States [69]. Meanwhile, some other machine translation researchers continued their studies and

attempts despite the after-effect of the report.

Traditionally, approaches to machine translation were mainly divided into two types: rule-based methods and corpus-based methods [151]. Rule-based Machine Translation (RBMT) in which translation depends on linguistic information retrieved from dictionaries and manually written rules was the primary focus of research during the early times of the development of machine translation as a scientific field [55]. Such classical systems were further categorized into Direct systems (map input to output with basic rules), Transfer RBMT systems (employ morphological and syntactical analysis), and Interlingual RBMT systems (using an abstract representation) [105].

One of the earliest noticeable translation systems which implemented the RBMT approach was the commercial translation system launched in 1968 by SYSTRAN, one of the oldest machine translation companies that survived the translation system crisis caused by the ALPAC report [142]. Google and Yahoo were users of the SYSTRAN system even until 2006 and 2012 respectively. Later in 1977, the METEO system [141] was developed specially for weather forecast translation in Canada, implementing similar approaches. In the following year, Carbonell [25] argued that competent translation requires some reasonable depth of understanding of the source text and access to detailed contextual information, and therefore proposed an Interlingual RBMT system.

Thereafter, in 1984, Nagao[104] pointed out the inconsistency of RBMT systems between languages with different structures (ex: Japanese-English) and proposed a translation system that implemented a new example-based learning method. Example-based machine translation (EBMT) was then joined by Statistical Machine Translation (SMT) and later Neural Machine Translation (NMT) to be the corpus-based set of methods that became dominant after the 2000s.

The idea of Statistical Machine Translation (SMT) was proposed in 1988 by Brown et al.[19], in which they outlined an approach to automatic translation that utilizes techniques of statistical information extraction from large databases. Two years later in 1990, Brown et al.[20] presented a statistical approach to machine translation with their first experimental results translating sentences from French to English. Their translation model introduced the concepts of word alignment and they discussed the importance of alignments of sentence pairs. Later in 1993,

Brown et al.[21] described a series of five statistical models of the translation process and defined a concept of word-by-word alignment between such pairs of sentences. Each of the five statistical machine translation models was modified with their unique alignment probability distribution.

As one of the first word-based SMT approaches, Och and Ney[107] explored and compared methods for computing word alignments, utilizing both statistical and heuristic models. Their key finding was the superior performance of refined alignment models with a first-order dependence and a fertility model compared to simple heuristic models. On the other hand, Marcu and Wong[95] presented a joint probability model for SMT, which automatically learned word and phrase equivalents from bilingual corpora. Later that year based on their findings and proposals, Koehn et al.[79] proposed a new phrase-based translation model and decoding algorithm that enabled evaluation and comparison between several, previously proposed phrase-based translation models. Another phrase-based SMT model was presented by Chiang[32], where he used hierarchical phrases. The model was formally a synchronous context-free grammar but learned from a bitext without any syntactic information, and also achieved a relative improvement over the state-of-the-art phrase-based systems at the time.

2.1.2 Early Neural Machine Translation

One of the first ideas of Neural Machine Translation (NMT) was proposed in 2003 by Bengio et al.[17], where they addressed the challenge of statistical language modeling, aiming to learn the joint probability function of word sequences. They tackled the curse of dimensionality by proposing a method that learns distributed representations for words, which allowed each training sentence to inform the model about an exponential number of semantically neighboring sentences. Their approach, implemented with neural networks, outperformed state-of-the-art n-gram models, demonstrating significant improvements and the ability to leverage longer contexts for better generalization. Later in 2008, Collobert and Weston[37] introduced a single convolutional neural network architecture designed to generate multiple language processing predictions. They demonstrated that both multitask learning and semi-supervised learning contribute to enhanced generalization, leading to state-of-the-art performance in the addressed language processing tasks.

Three years later, Sutskever et al.[136] discussed the challenges in training Recurrent Neural Networks (RNNs) and highlighted recent advancements in Hessian-free optimization that have overcome these challenges. They focused on demonstrating the effectiveness of RNNs trained with Hessian-Free (HF) optimizer in the character-level language modeling task. After a long training, their multiplicative RNNs utilizing HF optimization outperformed the best previous method for character-level language modeling, in which this achievement was noted as the largest improvement in such tasks for that time. Next, Schwenk[129] introduced a method for estimating translation model probabilities in phrase-based statistical machine translation. The approach employed neural networks to directly learn continuous representations of phrase pairs, demonstrating the ability to infer meaningful translation probabilities for unseen phrase pairs and predict likely translations. From then, the adoption of neural networks in the field of machine translation started to be widespread.

In 2014, Bahdanau et al.[12] introduced a groundbreaking novel approach to neural machine translation, aiming to overcome limitations in the traditional encoder-decoder architecture. The proposed method allowed the model to automatically and softly search for relevant parts of a source sentence when predicting a target word, eliminating the need for explicit hard segments. This innovative approach, by allowing the model to soft search for relevant segments, addressing the limitations of fixed-length vectors, lead to a significant leap in translation performance. The paper not only demonstrated comparable results to state-of-the-art phrased systems in English-French translation but also provided a qualitative analysis confirming the model's alignments aligning well with human intuition. The attention mechanism introduced in this work has since become a cornerstone in the field of NMT, influencing subsequent research and inspiring numerous model enhancements. Its incorporation has notably improved the ability of NMT systems to capture long-range dependencies and enhanced translation quality.

While the previous paper [12] introduced the attention mechanism to dynamically align source and target sequences during translation, Sutskever et al.[137] took a broader perspective on sequence learning. Their work proposed a comprehensive encoder-decoder architecture capable of handling diverse sequence transduction tasks beyond machine translation. The model's capacity to encode variable-length input sequences into fixed-size vectors, followed by decoding into target sequences,

offered a unified framework applicable to tasks like language modeling, summarization, and more. The sequence-to-sequence model not only demonstrated competitive performance but also provided a scalable and flexible architecture for various applications. The simplicity and effectiveness of this approach have contributed to its widespread adoption and further inspired the development of more sophisticated models in the subsequent years.

Following the strides in neural machine translation, the work of Cho et al.[34] represents a significant contribution to the evolving landscape. This paper builds upon the neural network-based sequence-to-sequence paradigm, focusing on the application of recurrent neural networks (RNNs) for statistical machine translation. Their approach involved an encoder-decoder architecture utilizing RNNs, a departure from fixed-length vector representations. The model, equipped with the ability to capture sequential dependencies, particularly in the form of phrases, demonstrated enhanced performance in statistical machine translation tasks. The RNN Encoder-Decoder presented in the paper showcased the adaptability of recurrent neural networks in encoding and decoding variable-length sequences. By delving into the learning of phrase representations, they added a nuanced layer to the evolving narrative of neural machine translation. The utilization of RNNs for capturing contextual information and dependencies marked a progression in the field, contributing to the broader understanding of effective sequence-to-sequence learning. Later on, Cho et al.[33] further delved into the analysis of NMT properties, focusing on the RNN Encoder-Decoder and a novel Gated Recursive Convolutional Neural Network. Notably, the paper revealed that while NMT excels in short sentences without unknown words, its performance diminishes with longer sentences and an increased number of unknown words. Additionally, the proposed gated recursive convolutional network is highlighted for its ability to autonomously learn the grammatical structure of sentences.

Later, in the study by Luong et al.[94], they scrutinized various attention mechanisms, proposing effective strategies to address common challenges and limitations. The paper systematically explored different attention architectures, shedding light on aspects such as global vs. local attention, and providing insights into their impact on translation quality. By fine-tuning the attention mechanism, the authors aimed to improve the efficiency and effectiveness of NMT systems. Their

work not only deepened the understanding of attention mechanisms in NMT but also provided practical guidance on selecting and implementing attention strategies based on specific translation tasks. The nuanced insights presented by Luong et al. contribute to the ongoing optimization of attention mechanisms, making them a pivotal reference in the evolution of NMT architectures, along with the previously mentioned works [12, 137, 34].

Dong et al.[45] addressed the challenge of a machine translation model that could translate sentences from a source language to multiple target languages simultaneously. Drawing inspiration from the neural machine translation model, they proposed an extension to a multi-task learning framework, where the framework shared source language representation while allowing the modeling of different target language translations. Their experimental results demonstrated that the multi-task learning model outperforms individually learned models, achieving significantly higher translation quality in both situations across publicly available datasets.

Later, Wu et al.[161] introduced Google's Neural Machine Translation system (GNMT), which Google implemented in Google Translate. The system employed a deep LSTM network with attention mechanisms and residual connections. To improve efficiency, it employed low-precision arithmetic during inference and addressed rare words by using sub-word units. The system achieved competitive results on benchmarks and significantly reduced translation errors compared to previous Google's phrase-based system in human evaluations on simple sentences. On the other hand, Cheng et al.[28] proposed a semi-supervised approach for NMT that leveraged both labeled (parallel corpora) and unlabeled (monolingual corpora) data. Their method involved reconstructing monolingual corpora using an autoencoder, with translation models serving as the encoder and decoder. This approach extended the benefits to both target and source language corpora while their experimental results demonstrated significant improvements on a Chinese-English dataset over state-of-the-art SMT and NMT systems.

2.1.3 Modern Neural Machine Translation

Around 2017, Kaiser et al.[74] introduced a versatile deep-learning model that achieved impressive results across various domains without extensive tuning. This singular model, trained concurrently on various translation tasks, image recognition

tasks, a speech recognition corpus, and an English parsing task, incorporated essential building blocks such as convolutional layers, an attention mechanism, and sparsely-gated layers. The model's joint training approach proved beneficial for tasks with limited data, while larger tasks exhibited minimal performance degradation. Later that year, the paper by Vaswani et al.[148] introduced the Transformer, a novel network architecture for sequence transduction based solely on attention mechanisms. Unlike dominant models relying on recurrent or convolutional neural networks, the Transformer abandoned recurrence and convolutions, achieving superior quality, increased parallelizability, and significantly reduced training time. This innovation has had a profound impact on the field of natural language processing, especially in machine translation, where the Transformer model achieved state-of-the-art results. Its simplicity, effectiveness, and versatility across various tasks have made it a foundational and influential contribution to the deep learning community.

On the foundation of the Transformer, Gu et al.[62] introduced a non-autoregressive NMT model that produced outputs in parallel, reducing inference latency significantly. The model incorporates knowledge distillation, input token fertility as a latent variable, and policy gradient fine-tuning. This approach achieved a near state-of-the-art performance compared to autoregressive Transformer networks. Later, Wei et al.[155] addressed a potential limitation of non-autoregressive translation models by proposing an imitation learning framework for non-autoregressive machine translation. While maintaining impressive translation speed, the proposed model considers previous context during parallel decoding, resulting in comparable translation performance to its autoregressive counterparts.

Around that time, Delvin et al.[43] introduced BERT or Bidirectional Encoder Representations from Transformers, a language representation model designed for pre-training deep bidirectional representations from unlabeled text. Unlike previous Transformer-based models, BERT considers both left and right context in all layers, allowing fine-tuning with minimal task-specific modifications. Its bidirectional approach enhances its ability to capture contextual nuances, enabling a more comprehensive understanding of language semantics. This feature contributed to BERT's exceptional performance, highlighting its versatility across a wide range of natural language processing tasks. BERT achieved state-of-the-art results on eleven natural language processing tasks, showcasing its simplicity and empirical

power. The bidirectional nature of BERT proved to be a key factor in surpassing benchmarks, demonstrating its efficacy in capturing complex language patterns and context-dependent information.

Later, Yang et al.[167] introduced XLNet as a generalized autoregressive pretraining method addressing the limitations of BERT. It enabled bidirectional context modeling by maximizing the likelihood over various permutations of factorization order. Unlike BERT, XLNet maintains an autoregressive formulation, mitigating the pretrain-finetune discrepancy. Additionally, it incorporated ideas from Transformer-XL [39] into pretraining. Empirical results demonstrated that XLNet outperformed BERT across 20 tasks, often by a significant margin. Later, Xu et al.[163] addressed optimization challenges in training deep Transformer translation models by exploring modifications to the official implementation, specifically altering the computation order of residual connection and layer normalization. They also delved into the subtle differences in computation order and introduced a parameter initialization method leveraging the Lipschitz constraint. Meanwhile, Liu et al.[90] went deeper into exploring the application of very deep Transformer models for NMT. Through a simple and effective initialization technique that stabilizes training, the authors demonstrated the feasibility of constructing standard Transformer-based models with up to 60 encoder layers and 12 decoder layers.

Furthermore, on improving the Transformer and NMT, Banar et al.[13] presented a novel Transformer-based approach for character-level NMT, comparing its speed and quality to both subword and character-level Transformers, with previous character-level models. The proposed architecture, trainable on a single GPU, was 34% faster than the character-level Transformer while achieving at least comparable results. Dougal and Lonsdale[46] investigated the integration of vetted terminology into NMT to enhance translation quality, ensuring consistency with an authorized multilingual terminology collection. Their approach involved leveraging the long short-term memory (LSTM) attention mechanism in contemporary NMT systems to accurately identify and align semantic entities in source and target languages and injecting appropriate terminology into the corresponding alignments during the decoding process. Additionally, they introduced a new translation metric, sensitive to approved terminological content, for evaluating the impact of terminology injection.

The following year He et al.[66] introduced a fast and accurate approach to translation memory (TM)-based NMT within the Transformer framework. The proposed model employed a simple architecture, utilizing a single bilingual sentence as its TM, resulting in efficient training and inference. Meanwhile, the work of Dankers et al.[41] addressed the evaluation of compositional generalization in neural networks for NLP, focusing on its application to NMT. The findings revealed that models trained on more data tend to be more compositional. However, unexpected levels of compositionality were observed, suggesting the need for nuanced modulation in models. The study highlighted the importance of reevaluating compositionality assessment methods and advocated for benchmarks that use real data to capture the complexities of meaning composition in natural language.

On the other hand, focusing on the training data, the study by Edunov et al.[47] explored methods to enhance neural machine translation using monolingual data through back-translations of target language sentences. It found that back-translations obtained via sampling or noised beam outputs are most effective, providing a stronger training signal than beam or greedy search. Lample et al.[82] also explored the possibility of machine translation without parallel data, pushing the boundaries of low-resource language pairs. The proposed model mapped sentences from monolingual corpora in two languages into a shared latent space, learning translation without labeled data. The model achieved good results without using any parallel sentences during the training, representing a significant advancement in addressing translation challenges with minimal linguistic supervision.

Meanwhile, Artetxe et al.[9] addressed the challenge of training NMT systems solely from monolingual corpora, as opposed to relying on large parallel corpora. They proposed an innovative approach using phrase-based SMT to narrow down the performance gap. Leveraging SMT's modular architecture, the method combined a phrase table induced from monolingual corpora through cross-lingual embedding mappings with an n-gram language model. Later year they also addressed the deficiencies in unsupervised machine translation by enhancing both NMT and SMT systems trained solely on monolingual corpora [8]. The improvements included leveraging subword information, introducing a theoretically sound unsupervised tuning method, and implementing a joint refinement procedure. They used the enhanced SMT system to initialize a dual NMT model, further refined through

on-the-fly back-translation. Their combined approach outperformed the previous state-of-the-art in unsupervised machine translation, achieving promising results and improvements. Later year, Conneau et al.[38] showcased the effectiveness of scaling up pretraining for multilingual language models, introducing XLM-R, a Transformer-based model trained on one hundred languages with over two terabytes of CommonCrawl data. XLM-R outperformed multilingual BERT on various cross-lingual benchmarks, demonstrating significant improvements, especially for low-resource languages. It achieved competitive results with strong monolingual models without sacrificing per-language performance. The authors emphasize key factors influencing their gains, including positive transfer, capacity dilution trade-offs, and performance variations in high and low resource languages at scale.

In tackling contextual data, Saunders et al.[127] introduced a novel approach to NMT training by incorporating document-level metrics with batch-level documents. Unlike previous work focusing on sentence-level metrics, their method addressed the importance of document-level metrics in the training objective. They combined data and model architecture approaches, enabling the use of document-level evaluation metrics in NMT training, which yielded performance gains over sequence Minimum Risk Training (MRT) and maximum likelihood training, demonstrating improved robustness for document-level metrics. Later, Wu et al.[159] published a study focusing on context-aware NMT, emphasizing effective encoding and aggregation of contextual information. They claimed that while BERT has proven effective in natural language understanding, its application in context-aware NMT was under-explored. Hence, they investigated three common methods to leverage BERT for encoding contextual information in NMT. Through experiments on five translation tasks, the approach of concatenating all contextual sequences into a longer one and then encoding it with BERT yielded the best results.

Later year, Dodapaneni et al.[44] presented a survey discussing the significant developments in Multilingual Language Models (MLLM), including multilingual BERT, XLM, and XLM-R, which have proven effective in pretraining for a diverse range of languages. Their survey encompassed research areas such as expanding MLLM to cover more languages, establishing comprehensive benchmarks for evaluating MLLM across various tasks and languages, analyzing MLLM performance in monolingual, zero-shot cross-lingual, and bilingual tasks, understanding uni-

versal language patterns learned by MLLM, and enhancing MLLM capacity to improve performance on both seen and unseen languages. Meanwhile, the study by Eo et al.[49] addressed Quality Estimation (QE), a task predicting machine translation quality without reference translations. While studies at the time have focused on performance improvement using data augmentation with large-scale multilingual pretrained language models (mPLM), this work aimed to conduct a pure performance comparison among various mPLM. Additionally, the study introduced EQ using the multilingual BART model, previously unused and performed comparative experiments with cross-lingual language models (XLM), multilingual BERT, and XLM-RoBERTa.

2.1.4 Latest Neural Machine Translation

Prompt-based methods are a type of method based on a single pretrained large language model, and operate by providing explicit instructions or queries in the form of prompts to guide the model's responses. Models large enough to allow prompting on a useful level, have gained attention in recent years for their remarkable performance in zero-shot and few-shot learning scenarios, where they exhibited the ability to generalize across various tasks given specific prompt instructions. The concept of prompt-based models has evolved over time, and it is challenging to pinpoint a single first model as the field has seen iterative development. However, one of the pioneering models has been OpenAI's GPT3 (Generative Pre-trained Transformer 3) [22, 108], which with its 175 billion parameters, has demonstrated impressive capabilities in natural language understanding and generation. Published by Brown et al. in 2020, this work highlighted the substantial improvements achieved in task-agnostic, few-shot performance by scaling up language models. GPT-3 performed well in tasks such as translation, question-answering, cloze task, and on-the-fly reasoning. However, challenges still exist, and some datasets reveal limitations of GPT-3's few-shot learning capabilities.

Later year, Webson and Pavlick[154] presented a study with over 30 prompt templates for natural language inference (NLI) experimentally tested, challenging the common belief that prompts facilitate faster learning in a manner similar to human comprehension of task instructions. Surprisingly, models demonstrated comparable learning speeds with intentionally irrelevant or misleading prompts as

they did with instructively good prompts. This held true even for large models with 175 billion parameters and instruction-tuned models trained on hundreds of prompts. The findings raised questions about the extent to which the impressive progress seen in prompt-based approaches reflected a genuine understanding of task instructions akin to human learning. Meanwhile, Jim et al.[72] introduced FewVLM, a prompt-based, low-resource learning method for vision-language tasks, addressing challenges posed by large and slow models. It employed a sequence-to-sequence transformer model, pretrained with prefix and masked language modeling. FewVLM outperformed larger models on the Visual Question Answering (VQA) task and achieved results comparable to much larger models. The analysis underscored the impact of diverse prompts on zero-shot performance and the effectiveness of noisy prompts in quick learning with ample training data.

In the following year, Kavumba et al.[76] examined few-shot prompt-based models designed to reduce data requirements for task-specific language models. Despite the intention to mitigate dataset-specific superficial cues, the study on MNLI, SNLI, HANS, and COPA benchmarks revealed that these models still exploit such cues. While they performed well on instances with superficial cues, their effectiveness diminished on instances lacking them, often marginally outperforming random accuracy or generally underperforming. Xu et al.[164] investigated the vulnerability of prompt-based learning paradigms to attacks during pre-training. By injecting or searching for triggers in plain text, the study demonstrated that these triggers can significantly compromise the performance of prompt-based models on downstream tasks. Their findings highlighted the universal vulnerability of this learning paradigm, with experiments showcasing the transferability of adversarial triggers among language models. Intriguingly, conventional fine-tuning models did not show vulnerability to such triggers. Lastly, they concluded their study by proposing potential mitigation strategies for these attacks.

In the meantime, Lang et al.[83] demonstrated the efficacy of co-training in enhancing prompt-based learning with unlabeled data. Co-training improved the original prompt model while enabling the learning of a smaller, task-specific downstream model. Their approach adapted to scenarios with partial access to prompt models or prohibitive fine-tuning costs. Results showed significant performance improvement, particularly on challenging datasets with a considerable

gap between prompt-based learning and fully-supervised models. The work of Li et al.[88] also addressed challenges in NMT such as fragility and limited style flexibility. Introducing prompt-driven neural machine translation, the approach leveraged prompts to enhance translation control and flexibility. Empirical results showcased the method's effectiveness in prompt responsiveness and translation quality. Lastly, human evaluation highlighted the flexibility of prompt control and efficiency in human-in-the-loop translation.

Zhang et al.[169] did a case study about prompting large language models for machine translation by exploring the under-explored realm of prompting for machine translation, offering a systematic examination of prompting strategies. Using GLM-130B [168], a bilingual (English and Chinese) pre-trained language model with 130 billion parameters, as a testbed, the research emphasized the impact of prompt example quantity and quality on translation. Findings revealed correlations between certain prompt features and performance, highlighted the effectiveness of pseudo-parallel prompt examples from monolingual data and demonstrated improved performance through knowledge transfer from different settings. Lastly, the paper concluded with an analysis of model outputs and discussions on the remaining challenges in prompting for machine translation. Meanwhile, Mayer and Brandt[35] investigated automated classification using prompt-based learning with transformer models for a domain-specific task. They applied zero-shot and few-shot approaches, comparing with fine-tuning and human ratings. Their study suggested a collaborative workflow involving both machine and human raters, emphasizing the potential of prompt-based learning for democratizing the use of AI. At the same time, Wu and Hu[160] outlined the Lan-Bridge Translation systems' participation in the Eighth Conference of Machine Translation (WMT2023) General Translation shared task, specifically between English and Chinese. Acknowledging the impact of large-scale models on various industries, especially in document-level machine translation, they adopted a research approach centered around advanced models like GPT-3.5 and GPT-4 [109]. Their team conducted prompt-based experiments to enhance document-level machine translation and aimed for optimal human evaluation results. Their paper concluded by presenting their final outcomes submitted to WMT2023.

Additionally, over the years, several comprehensive reviews offered insights into

the past, present, and future of NMT surrounding its challenges and applications [78, 170, 103].

2.2 Evaluation Metrics in Machine Translation

Apart from studies on NMT, a significant range of studies focused on developing automatic evaluation measures for NMT systems.

Effective evaluation metrics are essential in assessing the performance of machine translation systems. Here we explore and introduce various evaluation metrics and the current state of evaluation metrics for machine translation. In the beginning, ALPAC[6] proposed the very idea of evaluating the quality of machine translations. Years later, White et al.[157] published about the ARPA MT Evaluation Methodologies, which aimed to establish a framework for measuring and tracking the advancements of MT systems within the ARPA-sponsored research program. This ongoing effort, initiated in 1991, focused on refining evaluation metrics and processes to assess the effectiveness and quality of MT outputs.

In 2002, Papineni et al.[111] introduced a metric designed to automatically evaluate the quality of machine-generated translation. The metric, Bilingual Evaluation Understudy (BLEU), measured the precision of n-grams in the output against reference translations. BLEU was not the first MT evaluation method, but it had become a widely adopted metric in the field of machine translation for its simplicity and effectiveness in providing a quick and quantitative assessment of translation performance.

Many years later, various studies started to realize and emphasize the importance and challenges of MT evaluation metrics [134, 52, 99, 85]. More recently, Han[63] provided an overview of the evolution of MT since the 1950s and emphasized the crucial role of evaluation in not only assessing translation quality but also providing feedback for improvement. The paper discussed the history of machine translation evaluation (MTE), the classification of research methods, and recent progress in human evaluation, automatic evaluation, and meta-evaluation.

Finally in 2021, Kocmi et al.[77] discussed the prevalent practice of relying on automatic metrics as the primary means to assess and compare the quality of MT systems. They presented the largest collection of human judgments to the time and focused on pairwise rankings, a prevalent evaluation task. The study investigated the accuracy of metrics in predicting translation quality rankings, emphasizing the limitations of relying solely on the BLEU metric. In the end, the authors released 2.3 million sentence-level human judgments for 4,380 systems, promoting

a more comprehensive evaluation approach in machine translation research and development. In the meantime, Marie et al.[97] conducted the first large-scale meta-evaluation of MT at the time of publication by analyzing 769 research papers published from 2010 to 2020. Their study revealed concerning trends in automatic MT evaluation practices over the past decade. Notably, an increasing number of evaluations rely solely on differences in BLEU scores without statistical significance testing or human evaluation. Despite claims of 108 metrics being superior to BLEU, recent MT papers often compare metric scores without ensuring comparable training, validating, and testing data. The paper highlighted the lack of standardized metric score reporting tools in the MT community. To address these issues, the authors proposed guidelines for improved automatic MT evaluation and introduced a meta-evaluation scoring method to assess credibility. Lastly, for further information on evaluation metrics used in this paper such as BLEU [111], SacreBLEU [114], ROGUE [89], TER [133], CharacTER [152], METEOR [42], CHRF [113], BERTScore [170], BLEURT [131], and COMET [121], please refer to Section 4.2.

2.3 Research on Irony and Sarcasm

2.3.1 Definition of Irony

The Oxford Dictionary defines irony as "The expression of one's meaning by using language that normally signifies the opposite typically for humorous or emphatic effect."[1] Similarly, Merriam-Webster Dictionary defines irony as "The use of words to express something other than and especially opposite of the literal meaning."[2]

One way of looking at language is by acknowledging the dichotomous distinction between literal and figurative language. Literal language uses words to convey meaning literally, commonly defined in terms of direct meaning. Searle[130] highlights the features of literal language to be context-free, and sententially direct, with literal meaning being sharply distinguishable.

On the other hand, figurative language uses words in a way that deviates from their conventionally accepted definitions in order to convey a more complicated meaning or heightened effect. In the twentieth century, literature researchers focused on topics like metaphor or metonymy for their roles in literary texts. Figurative language was argued to be an important aspect of a poetic text, which gives the text a special aesthetic value [40]. Merriam-Webster's Encyclopedia of Literature [101] classifies figurative language into five categories: (1) resemblance or relationship, (2) emphasis or understatement, (3) figures of sound, (4) verbal games, and (5) errors. With this regard, irony can be used as (1), (2), and (4), sometimes as (5) when errors are used purposefully. Therefore, irony can be considered as one of the most representative types of figurative language, next to metaphors or similes.

Defining a good translation, or a good translated text in target language in the context of machine translation involves several key criteria that collectively ensure the translation's quality and effectiveness. Some of the key criteria are accuracy, fluency, consistency, cultural appropriateness, contextual relevance and technical quality. For ironic text, cultural appropriateness is considered to be most challenging due to the understanding of irony in different culture.

[1] https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100011437

[2] https://www.merriam-webster.com/dictionary/irony

While most of the previous research in irony detection within the field of AI focused on binary classification between ironic or non-ironic contents, two types of irony have been widely distinguished in most of the previous linguistic and communication studies on irony: verbal irony and situational irony[135, 67, 15, 147]. This distinction has also been acknowledged in the Semantic Evaluation 2018 Workshop, a workshop in the form of a contest where multiple teams attempt to develop a Machine Learning method based on a unified dataset. Especially, Task B of that workshop, focused on multi-class irony classification. The task had the participants compete in predicting one out of four labels describing i) verbal irony realized though a polarity contrast, ii) verbal irony without a polarity contrast, iii) descriptions of situational irony, and iv) non-irony [147, 15].

Situational irony is an unexpected or incongruous event in a specific situation that fails to meet an expectation [132, 15] Shelly et al.[132] gave an example of a typically ironic situation regarding firefighters who left something cooking, had a fire in their kitchen while they were out to put down a fire alarm. As firemen are usually the ones who extinguish fire instead of starting it, this situation is quite unexpected and is considered ironic. This shows that situational irony is usually produced unintentionally and not planned. As indicated by Grant[61] 'Situational irony focuses on the surprising and inevitable fragility of the human condition, in which the consequences of actions are often the opposite of what was expected.".

According to "A glossary of literary terms" by Abrams[4], verbal irony is a statement in which the meaning that a speaker employs is sharply different from the meaning that is ostensibly expressed. An ironic statement usually involves the explicit expression of one's attitude or evaluation but with intended implications being very different, and often opposite, to the literal attitude or evaluation. Verbal irony is considered different from situational irony in that it is produced intentionally by the speakers.

On the other hand, sarcasm is described to be "a way of using words that are the opposite of what you mean in order to be unpleasant to somebody or to make fun of them" by the Oxford dictionary (https://en.oxforddictionaries.com/). As an attempt to explain the differences between irony and sarcasm, "A Dictionary of Modern English Usage" [51] concludes that sarcasm does not necessarily involve irony and irony has often no touch of sarcasm', although sarcasm is often expressed

with irony as a tool. Hence the relationship between verbal irony and sarcasm have been confused in many studies. Kreuz et al.[80] argued that sarcasm and irony are similar in that both are forms of a reminder, yet different in that sarcasm conveys ridicule of a specific victim whereas irony does not. Lee et al.[84] followed up with an indication that a ridicule of a specific victim plays a more important role in sarcasm than in irony. They also concluded that a sarcastic utterance brings to mind the expectation of a specific person who is identified by that expectation, whereas irony brings to mind the collective expectation of numerous people. In the same vein, Jorgensen [73] coined the term “Sarcastic Irony" which is typically used to complain to or criticize intimates, who are usually the target of the remarks. Attardo et al.[10] argues that sarcasm is an overtly aggressive type of irony and claims that there is no consensus on whether sarcasm and irony are essentially the same thing, with superficial difference, or if they differ significantly. Many studies also claim that there is no way to distinguish between the terms [144]. Barbieri et al.[15] points out another reason why researchers do not differentiate between the irony and sarcasm which is due to the observation of a shift in meaning between the two terms. Barbieri concludes that while research efforts on irony and sarcasm are expanding, a formal definition is still lacking in the literature. Hence, many researchers tend to not distinguish between the terms and consistently use either of them throughout their studies.

On the other hand, some other languages other than English, such as Japanese (where irony/sarcasm is called 皮肉 *hiniku*), have no distinction between irony and sarcasm. This is because the figurative function of irony, which in many languages is considered as a sophisticated figure of speech enriching conversation, is used in Japanese only in aggressive (sarcastic) context. In view of all that has been mentioned so far, one may suppose that the difference between verbal irony and sarcasm is not finite as most of the recent studies interpreted both the words as the same. Our previous studies [30, 31] confirmed sarcasm to be a type of irony, therefore we will use both of them interchangeably in the remaining of this paper without actual differentiation.

2.3.2 Irony and sarcasm detection

Irony and sarcasm detection have gained prominence in Natural Language Processing, with various studies exploring this field [23, 123, 57, 31]. Often used interchangeably, irony is recognized as a significant element in human communication, serving as a prominent and pervasive figurative language tool. Following the Semantic Evaluation 2018 international workshop Task 3: Irony Detection in English Tweets [146] which attracted submissions from 43 teams worldwide for the binary classification task A, deep learning algorithms were further investigated and optimized for irony detection tasks. The leading system submitted by team THU_NGN [158] employed a densely connected LSTM network with multi-task learning strategy. Another noteworthy system from team NTUA-SLP [16] used an ensemble of two bi-directional LSTM network-based models, achieving comparable results. The submissions showcased a variety of approaches including neural network-based methods and popular at that time classification algorithms like SVM, Random Forest, and Naïve Bayes [147]. Overall, the approaches with ensemble learners were the current trend to tackle the challenges in irony and sarcasm detection.

Later, Zhang et al.[171] introduced a sentiment-based transfer learning approach which used sentiment knowledge to improve the attention mechanism of recurrent neural models for capturing hidden patterns for incongruity. They proposed a sentiment transferred Bi-LSTM model which is designed to transfer deep features from sentiment analysis into irony detection for learning both explicit and implicit context incongruity. Their model achieved state-of-the-art performance at the time and proved using sentiment knowledge can be an effective approach to improving irony detection. The appearance of the attention-based Transformer model proposed by Vaswani et al.[148] has inspired various popular language representation models including the Bidirectional Encoder Representations from Transformer(BERT) [43], along with many irony detection related studies [139, 56, 54]. Potamias et al.[116] proposed a neural network methodology that builds on a pre-trained transformer-based network architecture which is further enhanced with the employment in a recurrent convolutional neural network (RCNN). They tested their model on the Semantic Evaluation 2018 Task 3 dataset and achieved results surpassing all of the submissions. Their model also achieved state-of-the-art performance for all

benchmark datasets, outperforming all other methodologies and published studies at that time.

Other research investigated the complexities of detecting irony and sarcasm in Twitter using Machine Learning and Feature Engineering techniques [29, 30, 31]. They first clarified the definitions of irony and sarcasm through a review of relevant studies, followed up by including experiments comparing different classification methods and data preprocessing techniques. Their experiments highlighted the rise of deep learning methods in classification tasks, emphasizing the importance of social media markers. To address complexity of data preprocessing, different types of data preprocessing were implemented and compared, with optimal results obtained from minimal manipulation. Notably, cross-testing sarcasm and irony datasets demonstrated a high degree of similarity. Additionally, referring Ptaszynski et al.[117], who mentioned the problem of sarcasm as one of the components in cyberbullying, they experimented with a cyberbullying dataset. They conclusions that applying a model trained on sarcasm to a cyberbullying dataset in the form of a zero-shot learning, yielded promising results, suggesting the prevalence of sarcasm in cyberbullying [118, 119] .

Tomas et al.[143] further addressed the prevalent use of irony in social networks, particularly in the context of multimodal communication involving text and images. The study introduced a transformer architecture designed for fusing textual and image information to enhance irony detection. The proposed model employed disentangled text attention with visual transformers, resulting in a notable improvement of up to 9% in F-score compared to previous similar works and state-of-the-art visio-linguistic transformers. Interestingly, their text-only version of the model exhibited a capacity to capture ironic nuances in many instances, even without utilizing visual information. This reveals linguistic patterns that offer contextual cues for irony detection, suggesting potential avenues for understanding irony in textual content without relying on additional visual elements. The study by Turban and Kruschwitz[145], which also implemented the SemEval 2018 Task 3 provided dataset, addressed the persistent challenge of achieving desirable performance in automatic irony detection. While the inherent complexity of the problem has hindered progress, recent advancements have focused on ensemble classifiers and automatic data augmentation. Their study explored both these directions in the

context of irony detection in social media, especially on Twitter. The findings indicate that transformer-based models, when employed in ensemble classifiers, show significant improvements in multi-class classification tasks compared to strong baselines. In binary classification task, the performance is comparable to state-of-the-art alternatives.

In 2023, Bettelli and Panzeri[18] explored the challenge of detecting irony in instant messaging, where shared context may be limited. The study focused on the role of emoji as potential cues to compensate for the lack of conventional contextual elements. The experiment consisted of a questionnaire administered to 156 participants, who were presented with WhatsApp[3] messages followed by congruent or incongruent emojis in relation to the evaluative positive or negative content. The findings revealed that evaluative incongruent items were perceived as more ironic. Notably, incongruent positive messages (criticisms) were more easily recognized as ironic compared to incongruent negative messages (ironic compliments), aligning with the asymmetry of affect hypothesis. This suggested that emojis can serve as important contextual cues in the detection of irony in instant messaging. Meanwhile, Frenda et al.[54] delved into the analysis of perspectivism to enhance the understanding of how individuals perceive pragmatic phenomena, specifically – irony. Their study introduced an analysis of irony perception in 11 perspectivist models, each trained on annotations from crowd-sourcing workers with diverse characteristics such as gender, age, and nationality. Given the sparsity of the dataset, the research examined texts classified as ironic and non-ironic by these models, unveiling linguistic patterns associated with irony across various perspectives. Notably, the paper provided evidence for distinct linguistic patterns perceived as ironic by specific perspectives, offering insights such as American and Australian-trained models being more inclined to classify texts as ironic when featuring negative sentiment, and younger annotators' models being particularly influenced by words related to immoral behaviors.

[3] `https://web.whatsapp.com`

2.3.3 Other studies on irony and sarcasm

While we introduced some irony and sarcasm detection-related past works above, which were in the main focus of AI-related research before, there have also been some more nuanced cross-field studies on irony and sarcasm and related, which are worth mentioning. The paper by Salameh et al.[126] explored the preservation of sentiment when translating Arabic social media posts into English, both manually and automatically. Their study assessed the loss in sentiment predictability by comparing the sentiment labels of original Arabic texts with those determined manually and automatically in English. Despite the significant reduction in human ability to recover sentiment in translations, the paper demonstrated that automatic sentiment systems perform competitively in analyzing the sentiment of English translations compared to the original Arabic sentiment analysis. This suggested that although translation affects human understanding of sentiment, automated systems can effectively capture sentiment information from the translated texts. The work of Peled and Reichart[112] introduced the novel task of sarcasm interpretation, defined as generating a non-sarcastic utterance conveying the same message as the original sarcastic one. They presented a dataset of 3,000 sarcastic tweets, each interpreted by five human judges. Framing the task as monolingual MT, the paper explored various MT algorithms and evaluation measures, and finally proposed SIGN, an MT-based sarcasm interpretation algorithm that targeted sentiment words, a key component of textual sarcasm. Despite similar n-gram based scores across interpretation models, SIGN's interpretations received higher human scores for adequacy and sentiment polarity.

When it comes to irony detection in Chinese, study that addressed the challenges in evaluating irony detection in this language was done by Li et al.[87]. They emphasized the difficulty in establishing a comprehensive definition and the absence of a gold standard for computational models. The authors presented preliminary results from experiments conducted on an irony detection system for Chinese. They analyzed examples of irony and related phenomena that posed challenges for NLP classifiers, and shed light on the complexities involved in developing effective models for detecting irony in Chinese. Xiang et al.[162], however highlighted the challenging nature of automatic Chinese irony detection and its significant impact on linguistic research, noting the scarcity of labeled benchmark datasets for this

task. The authors introduced Ciron, the initial Chinese benchmark dataset designed for irony detection with machine learning models. Ciron comprised of over 8,000 posts collected from Weibo, a Chinese microblogging platform. They first collected the data intentionally with no pre-conditions to ensure broader coverage. Their evaluation of seven machine learning classifiers demonstrated the utility of Ciron as a valuable resource for advancing Chinese irony detection research. Huang et al.[68] once again addressed the challenging task of Chinese multi-dimensional sentiment detection, emphasizing its significant impact on semantic understanding. They highlighted the limitations in past irony datasets, which focused on annotating sentiment types for entire ironic sentences without providing corresponding intensity measures for valence and arousal. Recognizing that contextual information is crucial for understanding ironic statements, they introduced the extended NTU irony corpus, called the Chinese Dimensional Valence-Arousal-Irony (CDVAI) dataset, which included valence, arousal and irony intensities on the sentence-level, along with valence and arousal intensities on the context-level. Finally, they analyzed annotation differences among human annotators and employed a deep learning model like BERT to assess prediction performances on the CDVAI dataset, addressing the nuanced dimensions of sentiment in Chinese ironic sentences and contexts.

For Arabic irony detection, Abbes et al.[1] addressed the underexplored area of identifying irony in user-generated Arabic social media content. To facilitate research in this domain, the authors created a new open domain Arabic corpus specifically annotated for irony detection. They collected irony messages from Twitter using irony-related hashtags and manually annotated them according to their working definition of irony. The annotation process highlighted the challenges stemming from the inherent limitations of interpreting Twitter messages and the complexity of the Arabic language and its dialects. Lastly, the resulting corpus was positioned as a valuable and freely available resource for the development of open domain systems aimed at automatically recognizing irony in Arabic and its dialects within social media text. On the other hand, AlMazrua et al.[5] addressed the challenge of detecting irony in sentiment analysis, emphasizing the difficulty in recognizing implicit and indirect phrases that convey the opposite meaning. They introduced Sa'7r, the Saudi irony dataset, comprising 19,810 tweets, with 8,089 labeled as ironic, which was collected using the Twitter API. The study involved

training various models for irony detection, including machine learning models and deep learning models, with several good results. They concluded that the dataset and the evaluation results contribute valuable insights and resources to the field of irony detection, particularly in the context of the Saudi dialect of Arabic.

2.3.4 Previous studies on Irony Translation

In this section we reviewed some irony translation related linguistic past studies. There were not many of related previous studies focusing on irony translation at written time, hence chosen below are some of the most related studies. Moneva et al.[125] acknowledged the historical diversity in approaches and definitions of irony, generally emphasizing a discrepancy in meaning, such as between what is said and meant or between attitudes like blame and praise while most literacy and pragmatic perspectives at the time highlighted the role of inference in interpreting irony, challenging the traditional communication models. They also noted that the relevance approach to communication, grounded in inference, had been influential in understanding irony for the past decades. However, there is a relative scarcity of suggestions regarding problems and recurring traits in translating irony.

A study by Chakhachiro et al.[26] questioned the suitability of existing irony classifications for translation studies. Most literary and pragmatic criteria at the time, while broad, are deemed impractical for translation, especially between distant languages like Arabic and English. Hence the author argued for a more objective linguistic analysis, emphasizing the identification of formal and rhetorical devices in ironic texts to enhance the translation process. Later, Babil[11] addressed the heightened role and responsibility of translators in interlingual communication, particularly when dealing with irony, a nuanced and elusive form of expression. They warned about irony, by its nature, invites multiple interpretations and carries a risk of misunderstanding. Lastly, their study explored how various factors such as types of irony, literary genres, and cultural norms may act as contextual constraints influencing the translator's choices.

Chapter 3

Applied Datasets

3.1 Introduction of Dataset selection

SemEval Twitter irony dataset As one of the most popular social media platform in the past decade and the natural characteristic of its functions and services, Twitter[1] provides a promising amount of ironic data. Therefore, a large portion of datasets implemented in works on irony detection in the past consisted of Twitter messages (tweets). The Twitter Application Programming Interface (API) officially provided by Twitter and the trending use of hashtags by Twitter users also contributed to the data collection and management for Twitter datasets.

In 2018, the Semantic Evaluation workshop had included a task specifically focusing on irony detection in English Tweets as one of their challenges. Being one of the first workshops in the natural language processing series to tackle on irony detection after the surge of deep learning approaches, the workshop received submissions from 43 teams worldwide for the binary classification task[147, 158, 149, 58]. The submissions represented a range of deep learning approaches and other popular classification algorithms.

Due to the fact that this dataset was one of the most popular ones in previous studies on irony detection[147, 158, 149, 58, 29, 30, 31], and that it is curated and provided by the Semantic Evaluation workshop, we decided to implement this

[1]At present, the platform that used to be known as Twitter does not exist. Twitter underwent drastic modifications and rebranding into "X" since April, 2022, after which it cannot be considered the same platform as Twitter anymore.

dataset in our current study which is also focuses on irony in natural language processing. We also applied the same preprocessing methods from previous studies including the neutral labelling of tweets' metadata (e.g. tagged users, URLs, and emojis) and the manipulation of ironic hashtags[29, 30, 31].

Sarcasm Corpus V2 In contrast with the SemEval dataset, the Sarcasm Corpus V2 dataset provides samples of logically structured long sentences with lower rate of internet slang. It is created from the Internet Argument Corpus V2 which consists of dialogues and forum posts collected from several popular political debate forums targeted at English speakers. There are three subsets of sarcastic data provided in the Sarcasm Corpus V2 which are categorized as hyperbole, rhetorical questions, and general sarcasm[110]. Some of the entries in hyperbole and rhetorical question subset samples are also existing in the general sarcasm subset. However, we decided to implement only the general sarcasm subset due to duplication in different categories and also being the largest subset of the Sarcasm Corpus V2 which included 69% of the samples in the dataset.

Dataset 1 (SemEval Twitter irony dataset) consists of short sentences with higher rate of social media slang and characteristics. Therefore the content of such tweets could be about anything with unlimited context. On the other hand, Dataset 2 (Sarcasm Corpus V2) covers logically structured long sentences and paragraphs focusing on debate of specific topics. Therefore, by combining in the research both datasets mentioned above, we can improve the scope of our study and assure diversity in our experiments.

3.2 Dataset 1: Ironic tweets

As the basis for the first dataset we used the SemEval irony dataset, which is a dataset provided by The Semantic Evaluation 2018 Task 3: Irony Detection in English Tweets[147]. The dataset was initially provided for all participating teams in the competitive workshop. It was constructed by searching Twitter for hashtags #irony, #sarcasm, and #not, which could occur anywhere in the tweet that was finally included in the dataset. Hashtag is keyword beginning with a hash symbol(#) followed by the relevant keyword or phrase in a tweet, used to categorize those tweets and help show them more easily in Twitter search. However, to minimize the noise introduced by collecting tweets automatically with ironic hashtags, all tweets were manually labelled using a fine-grained annotation scheme for irony[146]. All tweets were collected between 2014/12/01 and 2015/01/04 and represent tweets uttered by 2,676 unique users. The entire dataset was cleaned by removing retweets, duplicates and non-English tweets, and replacing XML-escaped characters (e.g. &). The dataset consists of 4,618 tweets (2,222 ironic and 2,396 non-ironic) that were manually labeled by three annotators using the Brat Rapid annotation tool with an inter-annotator agreement study set up to assure the reliability of the annotations[147] with Fleiss' Kappa used as metric for assessing the annotator agreement on categorical ratings, with Kappa value averaging 0.72.

Ratio of both irony and non-irony classes is 48.1% and 51.9%, respectively. Table 3.1 shows the general statistic of Dataset 1. The average token length and character length per entry for both classes are 16.754 and 16.459, and 76.901 and 79.096, respectively, while the standard deviation for both tokens and characters in both classes are 6.759 and 6.678, and 29,362 and 29,229, respectively. These show that both of the classes in this dataset are very balanced in term of tokens.

Out of all 37,228 and 39,433 tokens in both irony and non-irony class, we extracted 8,514 and 10,605 distinct tokens (no duplication, any letter cases), respectively. Then, similarly to the work of Eronen et al.[50], we calculated the lexical density and the dataset complexity of both classes in Dataset 1. We reached the results of 0.229 and 0.269 for lexical density (the higher the denser) of both classes. We can see the non-irony class has higher lexical density than the irony class This is due to the fact that the dataset was collected using fixed ironic hashtags (e.g. #not, #sarcasm, and #irony), therefore the amount of similar vocabulary (e.g.,

Table 3.1: General statistic of Dataset 1 (Twitter irony dataset).

General statistic of Dataset 1	irony class	non-irony class
entries	2,222	2,396
all chars	170,874	189,515
all tokens	37,228	39,433
distinct tokens	8,514	10,605
tokens per entry (average)	16.754	16.459
tokens per entry (std. dev.)	6.759	6.678
characters per entry (average)	76.901	79.096
characters per entry (std. dev.)	29.362	29.229
ratio of lexical density	0.229	0.269
ratio of long tokens	0.782	0.792
ratio of dataset complexity	0.505	0.530
Number of entries with...		
#any_hashtags	2,222	1,199
#not	907	439
#sarcasm	812	78
#irony	505	62
@user_mentions	794	1,123
URLs	304	750
emojis	208	286
Number of all...		
#any_hashtags	3,741	3,658
#not	908	445
#sarcasm	813	78
#irony	505	62
@user_mentions	1,073	1,809
URLs	320	825
emojis	375	533

ironic hashtags) in the irony class is significantly higher than the non-irony class. On the other hand, the amount of user-mentions, URLs, and emojis have higher frequency in non-irony class.

As shown in the second half of Table 3.1, all 2,222 entries in the irony class contains ironic hashtags due to being collected and classified with the help of the 3 specific ironic hashtags. However even in the non-ironic class, there are 579 out of 2,396 entries which contains the ironic hashtags, where 439, 78, and 62 of them contain #not, #sarcasm, and #irony, respectively. From our observation, this comes from unusual situations, where some of the authors wrote their sentences with each words appended to a hashtag (e.g. "#I #do #not #want #to #sleep #yet"). Hence, tweets with similar structure may cause some problem to the machine translation for their hashtagged words which were originally not meant to be a hashtag. Therefore almost 25% of the non-ironic tweets contains unintentional ironic hashtags, where #not, being an adverb as a part of their sentence, has the highest frequency.

Finally, we measured the ratio of long tokens in both irony and non-irony classes as 1 minus the division of average tokens per entry divided by average characters per entry, with the results of 0.782 and 0.792 respectively. The long tokens ratio of both classes were very similar, with the non-irony class only slightly higher by 0.01. Lastly, we calculated and compared the dataset complexity (the higher the more complex)[50] between both irony and non-irony class. The non-irony class achieved a higher dataset complexity of 0.53 while the irony class reached dataset complexity of 0.505.

Additionally, some preprocessing was applied to the provided raw dataset for our implementation. Each tweet was first transformed into lowercase. Moreover, according to Post et al.[115], emojis can be sometimes untranslatable due to being a unique type of data which are single Unicode codepoint. Therefore, in our data, all emojis were replaced with a universal label ":emoji:". Furthermore, all user mentions (e.g. @user123) and URLs (e.g. `https://google.com`) appearing in the text were replaced with specific neutral labels, such as "@user@" and "_url_". It is because these metadata were not likely to be contributing to the translation, but when left unprocessed could create undesirable bias.

Table 3.2: General statistic of Dataset 2 (Forum debates).

General statistic of Dataset 2	sarcastic class	non-sarcastic class
entries	3,260	3,260
all chars	526,020	809,312
all tokens	112,383	170,165
distinct tokens	20,323	23,361
tokens per entry (average)	34.473	52.198
tokens per entry (std. dev.)	24.659	43.821
characters per entry (average)	161.356	248.255
characters per entry (std. dev.)	114.559	168.390
ratio of lexical density	0.181	0.137
ratio of long tokens	0.786	0.789
ratio of dataset complexity	0.484	0.464

3.3 Dataset 2: Sarcasm in forum debates

The Sarcasm Corpus V2 is a large scale, highly diverse corpus of sarcasm developed using combination of linguistic analysis and crowd-source annotation, and published by The Baskin School of Engineering at UC Santa Cruz[110]. It is an update to the Sarcasm corpus V1[93] and was created from the Internet Argument Corpus 2.0[2, 150]. The Internet Argument Corpus 2.0 is a collection of corpora for research in political debate on internet forums consisting of dialogues and posts from `4forums.com`, `createdebate.com`, and `convinceme.com`. The data was collected from Internet Argument Corpus using the pattern experiments similar to Sarcasm corpus V1 using AutoSlog-TS[124]. The collected data was then annotated with Mechanical Turk, which is a crowdsourcing platform to hire remotely located crowdworkers to perform discrete on-demand tasks operated under Amazon Web Services[2]. A qualifier consisting of manually selected questions was also created to filter out unreliable annotators. This dataset was also used as a benchmark in many previous sarcasm related works[59, 48, 92, 128].

The Sarcasm Corpus V2 contains data representing three categories of sarcasm: hyperbole, rhetorical questions, and general sarcasm which are being used in this research. The generic sarcasm subset consists of 3,260 posts per class (sarcastic

[2] `https://aws.amazon.com/`

and non-sarcastic, 6,520 in total) with a class ratio of 1:1. Comparing to Dataset 1 (tweets), Dataset 2 which is generally composed of forum posts and dialogues has a higher count of words (tokens) due to forum posts being generally much longer. Furthermore, there are no special content such as hashtags and tagged users which would require extra preprocessing.

The average post length in words in sarcastic and non-sarcastic class is 34.473 and 52.198, while average length in characters is 161.356 and 248.255, respectively. The sarcasm may have been an important factor in causing the non-sarcastic class to have more average tokens length than the sarcastic class. Instead of explaining thoroughly and clearly in a longer form of post, often the authors of sarcastic posts have chosen to express themselves in a sarcastic manner with the intention of leaving their audiences to figure out their contents. This also caused the sarcastic class to have a higher lexical density of 0.181, comparing to the 0.137 of non-sarcastic class. However, both sarcastic and non-sarcastic class have a similar ratio of long tokens of 0.786 and 0.789, respectively. Finally, we measured the dataset complexity[50] of both classes and reached the results of 0.484 for sarcastic class, and 0.464 for non-sarcastic class. Surprisingly, both classes had similar dataset complexity even with huge differences in average tokens and lengths in characters.

3.4 Dataset 3: English-Chinese Parallel Combined Dataset

Dataset 3 we developed in this research is a new dataset containing all entries from both Dataset 1: SemEval Twitter irony dataset and Dataset 2: Dataset of sarcasm in forum debates. There are four subsets in Dataset 3, namely, tweet subset and forum post subset, which are equivalent to Datasets 1 and 2, as well as their corresponding training and test subsets. Details of the combined dataset were represented in Table 3.3.

As the new Dataset 3 inherited the characteristics of both datasets, it contains diverse entries ranging from short tweets to long forum debate posts, or, from random social media posts full of unintelligible internet slang to logically structured topic-focused debating paragraphs. By applying previous works' results which supported the claim that sarcasm is in most cases a type of irony[29, 30, 31], we

unified both irony and sarcasm class from two different datasets and finalized the class name to be irony. The reasoning behind this approach is, stems from the fact that although in theory irony and sarcasm are sometimes defined differently from the linguistic point of view, in practice, studies on irony or sarcasm detection usually mix the two and release data under the label of “sarcasm", while in fact it is annotated as “irony". This has also been confirmed experimentally before[29, 30, 31].

For further understanding into the irony of the new Dataset 3, we decided to further classify the entries into four classes with labels being:

- 0: **Non Irony**,
- 1: **Self-contained Irony**, which refers to an ironical statement that has sufficiently long context for the irony to be understood,
- 2: **Contextual Irony**, where the additional context for certain words or phrases is needed in order to understand the ironic meaning.
- 3: **Ambiguous Irony**, which is originally meant to be ironic but highly unintelligible to most audiences.

Then, all of the irony class entries were manually labelled by three annotators who had expertise in figurative language related field and had some experience in annotating natural language processing-related datasets.

To create the parallel translation counterpart in simplified Chinese for references in evaluation, the entire dataset was first machine-translated into simplified Chinese using Google Translate, a multilingual neural machine translation service developed by Google. The translations were then manually checked and corrected by four students-Chinese native speakers who were also second-language speakers of English, with each of them being in charge of one fourth of the whole dataset. Next, we performed a second phase of quality check by two other students fluent in both languages. All of the translators possess a degree in Computer Science from two different universities respectively. One of them was female, and the remaining five are male, with ages ranging from 20 to 30.

Table 3.3: Types of irony in subset of Dataset 3 (combined dataset)

Subset	Subset 1 (Twitter)		Subset 2 (Forums)	
Labels	train	test	train	test
Non Irony	1,923	473	2,608	652
Self-contained Irony	1,801	289	2,608	652
Contextual Irony	78	14	0	0
Ambiguous Irony	31	8	0	0
Total	3,834	784	5,216	1,304

Chapter 4

Applied Methods

4.1 Language Models

4.1.1 mBART

mBART is a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective [86] and presented by Liu et al.[91]. It consists of an encoder and an autoregressive decoder, and is one of the first methods for pretraining a complete sequence-to-sequence model by denoising full texts in multiple languages, while other approaches have focused only on the encoder, decoder, or reconstructing parts of the text. mBART uses a standard sequence-to-sequence Transformer architecture [148], with 12 layers of encoder and 12 layers of decoder with model dimension of 1024 and 16 heads. It tends to two type of noises in their noising function before the Transformer model is trained to recover the text. The first replaces phrases with a mask token, and the latter permutes the order of sentences within each instance. It is initially pretrained on a CC25, a subset of 25 languages extracted from the CommonCrawl (CC) corpus [156].

Based on the mBART-CC25, Tang et al.[138] demonstrated that pretrained models can be extended to incorporate additional languages without loss of performance by doubling the number of languages in mBART-CC25. They released the mBART50 model which pretrained in 50 different languages and created the ML50 benchmark, covering low, mid, and high resource languages, to facilitate

reproducible research by standardizing training and evaluation data.

The mBART models have been implemented in various kinds of works. Chakrabarty et al.[27] did a study on neural poetry translation with mBART50 fine-tuned on different poetic and non-poetic datasets. Their results show that multilingual fine-tuning on poetic text significantly outperforms non-poetic text even in thirty five times large difference in data size, and also outperforms bilingual fine-tuning on poetic data. Abdelghaffar et al.[3] proposed a method to adapt multilingual MT models to a low resource language using the mBART50 model. They fine-tuned their model with only a closely related high resource language and obtained better performances comparing to their previous attempts. Lai et al.[81] also implemented mBART for multilingual text style transfer and achived competitive performance without monolingual task-specific data. Both mBART and mBART50 are included in the list of provided pre-trained models in some shared tasks in the evaluation campaign of the 19th International Conference on Spoken Language Translation (IWSLT) [7].

4.1.2 Helsinki-NLP-opus-en-zh

Helsinki-NLP is the Language Technology Research Group at the University of Helsinki. They focused on NLP for morpghologically-rich languages, cross-lingual NLP, and NLP in the humanities. They developed and published more than 1,440 models in the Huggingface repository for different languages pairs. We implemented the Helsinski-NLP-opus-en-zh model in our study for comparing between language models focused on pair languages and multilingual models. It is a transformer-based duo-lingual translation model specially trained for translating English to Chinese[1].

4.1.3 MT5

MT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (T5)[2] is a multilingual variant of T5 that was pre-trained on a new Common-Crawl-based dataset covering 101 languages [166]. The original T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. The developers of the model

[1] `https://blogs.helsinki.fi/language-technology/`
[2] `https://huggingface.co/K024/mt5-zh-ja-en-trimmed`

[120] proposed reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings, in contrast to BERT-style models that can only output either a class label or a span of the input. However, T5's text-to-text framework allows to use the same model, loss function, and hyperparameters on any NLP tasks. We implemented the official large MT5 model and a Chinese-Japanese-English pretrained MT5 model. The Chinese-Japanese-English MT5 model was pretrained on most wikimedia, wikititles, ted2020, and news commentary data between the three languages.

4.1.4 ByT5

ByT5 is a pretrained byte-level Transformer model based on the T5 architecture introduced by Xue et al.[165]. Comparing to other widely-used pretrained language models which operate on sequences of tokens corresponding to word or subword units, token-free models that operate directly on raw text have many benefits. Some of the benefits include processing text in any language out of the box, being more robust to noise, and minimizing technical debt by removing complex and error-prone text preprocessing pipelines.

4.1.4.1 ChatGPT

ChatGPT is a system that consists of a language model developed by OpenAI [22, 108] and further fine-tuned for chat-like abilities with instruction tuning. The GPT stands for "Generative Pre-trained Transformer" and is a type of machine-learning model that generates human-like responses to text-based prompts or questions. The largest version of GPT model at said time was trained on a dataset of over 45 terabytes of text data which includes a wide range of sources such as books, websites, and other forms of digital text. Specifically, for our purposes, we used a free version of ChatGPT, which allows for the use of the GPT-3.5 model. Unfortunately, OpenAI does not reveal the details of either the GPT models, or the ChatGPT as a system. Although, releasing an extensive technical report on the model OpenAI[109], it is not verifiable how the model works, or what exact requests are being sent between the ChatGPT interface and the end model. Therefore, the model as such cannot be fairly used in any research. Moreover, OpenAI can

also stop the access to the model and its API at any time, making any and all studies using this model not replicable. Therefore, making ChatGPT the center of any study will make the study vulnerable to being completely not replicable, thus by definition, making it pseudoscientific. For this reason, we did not make ChatGPT at the center of our study. However, we do acknowledge, that ChatGPT has become a widely used tool, and as such can be useful. Therefore, we used it at a later stage of the research. Specifically, we used ChatGPT to translate our test set from English to Chinese using a zero-shot prompt-based approach and to compare those translations with our best-performing model.

4.2 Evaluation Metrics

Kocmi et al.[77] categorized automatic metrics for machine translation (MT) into two types, namely, string-based metrics and pretrained-model-based metrics. String-based metrics compare the coverage of various substrings between machine-translated texts and provided references, and is largely dependent of the quality of the references. Pretrained-model-based metrics use pretrained language models to evaluate the quality of machine-translated texts given the source texts, references, or both, and the performance is largely influenced by their training data.

4.2.1 String-based Metric

4.2.1.1 BLEU

BLEU, or (BiLingual Evaluation Understudy) [111] is a metric for automatically evaluating quality of machine-translated text. It measures the similarity of the machine-translated text to a set of high quality reference translations. The BLEU metric ranges from 0 to 1, however in many works it is expressed as percentages rather than decimal. It was noted that with more reference translations for each translated sentences, the score will be higher.

BLEU score is mathematically defined as follows.

$$\text{BLEU} = \underbrace{\min\left(1, \exp\left(1 - \frac{\text{reference-length}}{\text{output-length}}\right)\right)}_{\text{brevity penalty}} \underbrace{\left(\prod_{i=1}^{4} precision_i\right)^{1/4}}_{\text{n-gram overlap}}$$

with

$$precision_i = \frac{\sum_{\text{snt} \in \text{Cand-Corpus}} \sum_{i \in \text{snt}} \min(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{\text{snt'} \in \text{Cand-Corpus}} \sum_{i' \in \text{snt}'} m^{i'}_{cand}}$$

where

m^i_{cand} is the count of i-gram in candidate matching the reference translation,

m^i_{ref} is the count of i-gram in the reference translation, and

w_t^i is the total number of i-grams in candidate translation.

BLEU score is mathematically defined by two parts, the brevity penalty and the n-gram overlap. The brevity penalty penalizes generated translations that are too short compared to the closest reference length with an exponential decay. It compensates for the fact that BLEU score has no recall term. N-gram overlap counts how many unigrams, bigrams, trigrams, and four-gram match their n-gram counterpart in the reference translations.

4.2.1.2 SacreBLEU

BLEU have been the dominant evaluation metric in machine translation field for past decade. However, there are also many works which pointed out the shortcomings of the BLEU metric [24, 122]. Different from other works which mostly concerned on the shortcomings of BLEU for the evaluation quality, Post[114] focused on the relatively narrower problem of the reporting of BLEU scores. He summarized the problem as follows. BLEU is not a single metric, but requires a number of parameters, preprocessing schemes have a large effect on BLEU scores, and lastly papers which implemented BLEU vary in the hidden parameters and unreported schemes applied. Hence, Post[114] developed a new tool, SacreBLEU, to facilitate the problem of reporting BLEU scores.

SacreBLEU is a measure that aims to solve the BLEU problems by combining the original reference implementation [111] with other useful features, such as automatically downloads WMT test sets and processes them to plain text, produces short version string that facilitates cross-paper comparisons, and computes scores on detokenized outputs using WMT standard tokenization. It provides hassle-free computation of shareable, comparable, and reproducible BLEU scores. SacreBLEU is also tuned to adapt the implementation of WMT (Conference on Machine Translation). The common Python implementation of SacreBLEU can automatically download common WMT test sets and processes them to plain text. It also computes scores on detokenized outputs using WMT standard tokenization, and produce the same values as the official script used by WMT. SacreBLEU supports different tokenizers for BLEU including the support of Japanese and Chinese. Lastly, it produces a short version string that facilitates cross-paper comparisons [114].

4.2.1.3 ROGUE

ROGUE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics to automatically determine the quality of a machine-translated text [89]. It counts the number of overlapping units such as n-gram, word sequences, and word pairs between the machine-translated texts and reference texts provided. ROGUE provides the following types of metrics:

- ROGUE-1: Overlap of unigram
- ROGUE-2: Overlap of bigrams
- ROGUE-L: Longest Common Subsequence (LCS) based statistics
- ROGUE-W: Weighted LCS-based statistics
- ROGUE-S: Skip-bigram based co-occurrence statistics
- ROGUE-SU: Skip-bigram plus unigram-based co-occurrence statistics

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}\in\{\text{Reference N-grams}\}} \min(\text{Count}_{\text{match}}(\text{gram}), \text{Count}_{\text{candidate}}(\text{gram}))}{\sum_{\text{gram}\in\{\text{Reference N-grams}\}} \text{Count}_{\text{Reference}}(\text{gram})}$$

4.2.1.4 TER

Translation Error Rate (TER) is an error metric for machine translation that measures the number of edits required to change a system output into one of the references. It was introduced in 2006 by Snover et al.[133] as a new and intuitive measure for evaluating machine-translation output that avoids the knowledge intensiveness of more meaning-based approaches, and the labor-intensiveness of human judgements. The implementation of TER is also presented in SacreBLEU, requiring a slightly different input format comparing to the original TERCOM implementation.

$$\text{TER} = \frac{\text{Number of edits}}{\text{Average number of reference words}}$$

4.2.1.5 CharacTER

CharacTER is a character-level metric inspired by TER. It calculates the character level edit distance while performing the shift edit on word level. It is defined as the minimum number of character edits required to adjust a hypothesis, until it completely matches the reference, normalized by the length of the hypothesis sentence [152].

$$\text{CharacTER} = \frac{\text{Number of character edits}}{\text{Number of characters in reference}}$$

4.2.1.6 METEOR

METEOR, or Metric for Evaluation of Translation with Explicit ORdering (METEOR) is an automatic metric for machine translation evaluation introduced in 2014 [42]. METEOR is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translation. It focuses on the harmonic mean of unigram precision and recall, with recall weighting higher than precision. Unigrams can be matched based on their surface forms, stemmed forms, and meanings.

METEOR score can be calculated as shown below:

$$\text{METEOR} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P} \times (1 - \gamma \cdot \left(\frac{ch}{m}\right)^{\theta})$$

where

P is Precision

R is Recall

ch as chunks which is the number of contiguous sequences of matched words

m refers to the total number of matched words

$\gamma and \theta$ are tunable parameters that control the impact of the penalty.

4.2.1.7 CHRF

In 2015, Popovic[113] proposed the use of character n-gram F-score for automatic evaluation of machine-translated text. He developed CHRF which calculates the F-score averaged on all characters and word n-grams, and arithmetic mean is applied for n-gram averaging. CHRF is one of the latest string-based metrics and recommended over using BLEU by many metric comparison related works [98, 77, 102]. CHRF has been implemented in many works alongside with BLEU and other evaluation metrics [140, 96, 14].

The general formula of CHRF score is shown below:

$$CHRF\beta = (1 + \beta^2)\frac{CHRP.CHRR}{\beta^2.CHRP + CHRR}$$

where CHRP and CHRR represent character n-gram precision and recall arithmetically averaged over all n-grams.

4.2.2 Pretrained Model-based Metrics

4.2.2.1 BERTScore

BERTScore is an automatic evaluation metric for text generation proposed in 2020 by Zhang et al.[172]. It is a pretrained-model-based metric specifically designed to be simple, task agnostic, and easy to use. BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence using contextual embeddings from BERT. Taking advantages of BERT's semantic and syntactic abilities, BERTScore seeks to avoid flaws of older metrics which largely relying on surface-level features such as n-grams [65]. As a measure of goodness, BERTScore computes precision, recall, and F1 score for each sentences.

4.2.2.2 BLEURT

BLEURT is a learned evaluation metric based on BERT that can model human judgements with a few thousand possibly biased training examples proposed in 2020 by Sellam et al.[131] for Natural Language Generation. It takes a pair of sentences as input, a reference and a candidate, and it returns a score that indicates

to what extent the candidate is fluent and conveys the meaning of the reference. BLEURT is a regression model based on BERT and RemBERT [36].

4.2.2.3 COMET

COMET is a neural framework for training multilingual machine translation evaluation models which obtains new state-of-the-art levels of correlation with human judgements proposed in 2020 by Rei et al.[121]. Their framework leverages recent breakthroughs in cross-lingual pretrained language modeling resulting in highly multilingual and adaptable MT evaluation models that exploit information from both the source input and a target-language reference translation in order to more accurately predict MT quality. Unable to discard BLEU completely, many of recent works in machine translation applied both BLEU and COMET as their evaluation methods [27, 64, 106]. However, Marie[96] observed that the absolute COMET scores are meaningless on their own and have to be combined with another metric score to fully estimate the translation quality.

4.3 COMMET: Combined Metric for Machine Translation

4.3.1 Recent trends in MT evaluation methodology leading to the proposal of COMMET

There has been an extensive research done on benchmarking and improving the existing metrics for MT evaluation.

Mathur et al.[98] revisited the findings of the metrics task at WMT19 and developed a method for thresholding performance improvement under an automatic metric against human judgements. They recommend not to use CHRF instead of BLEU or TER for evaluation of machine translation. Mathur et al.[99] also presented the results of WMT20 Metrics Shared Task in machine translation evaluation. They concluded the performance of several metrics on some language pairs. For the results on the Chinese to English and English to Chinese translation, BLEU ranked as the best metric lower only than other 3 systems.

Kocmi et al.[77] published an extensive evaluation of automatic metrics for machine translation. In the paper, they corroborate how reliable metrics are in contrast to human judgements on their collection of judgements reported in the literature. Based on their findings, they recommended COMET as the main automatic metric for best practices. The authors also suggested using a string-based metric for unsupported languages and as a secondary metric. CHRF is recommended and BLEU is highly rejected as they presented indirect evidence that the overuse of BLEU negatively affects Machine Translation development.

Marie et al.[97] presented the first large-scale meta-evaluation of machine translation. They annotated machine translation evaluation conducted in 769 research papers published from 2010 to 2020. They reported an increasing number of machine translation evaluations blindly rely on BLEU scores for conclusion while there are at least 108 metrics claiming to be better than BLEU. Lastly, a guideline for machine translation evaluation was proposed and they recommended not to exclusively rely on BLEU.

In 2022, Bapna et al.[14] shared their findings in an effort to build practical machine translation systems capable of translating across over one thousand languages.

Additionally they compared between BLEU and CHRF for machine translation evaluation. They deemed BLEU to be unsuitable for their system and chosen CHRF instead. Later, Karpinska et al.[75] created DEMETR, a diagnostic dataset for evaluating the sensitivity of machine translation metrics to 35 different linguistic perturbations. They found that learned metrics perform substantially better than string-based metrics on DEMETR.

The results of WMT22 Metric Shared Task was presented along with further emphasis on retiring BLEU [53]. The authors summarized the results of the shared task on automated machine translation evaluation and presented an extensive analysis on the performance of metrics on three main language pairs: English to Russian, English to German, and Chinese to English. The results demonstrated the superiority of learning-based pretrained metrics over string-based metrics.

Furthermore, Lee et al.[85] conducted another survey on evaluation metrics for machine translation and analyzed their strengths and weaknesses. They pointed out that traditional metrics (e.g., BLEU, TER) show poor performance in capturing semantic similarity between machine and human translations. Lastly, they confirmed that the learned metrics show the best performance and have the highest correlation with human evaluation but require training data.

The above meta studies, also backed by our initial experiments, conclude that there is no single best automatic evaluation metric for machine translation. Although some metrics tend to be better than others, there could still be situations, where the most criticized BLEU outperforms novel metrics based on pretrained language models. This finding led us to the proposal of a novel flexible COMMET metric, instantly applicable for any MT research.

4.3.2 Proposal of the COMMET metric

In order to prevent biased results from few dominant automatic evaluation metrics and due to no state-of-the-art, or widely used optimal automatic evaluation metric for figurative language translation, we decided to implement multiple metrics described above in one concise metric to use in our following experiments.

Specifically, we propose COMMET, or Combined Metric for Machine Translation. It is a new metric for MT, which combines several MT metrics with a designated weight for each of the implemented metrics.

The core idea behind the proposed metric came from he following consensus regarding the state-of-the-art in MT evaluation metrics.

1. There exists no ideal single automatic evaluation metric for machine translation.
2. Some metrics are more reliable then others.
3. It is inefficient and unrealistic to interpret multiple metrics for every instance in a case-by-case basis.
4. Even methods widely criticized as not adequate (BLEU, ROGUE), sometimes for some language pairs in some situations provide better assessment then the new metrics.
5. Some metrics work better for some language pairs rather then the others.
6. Although recent developments with metrics based on pretrained language models are promising, the evaluation metrics still fall far behind human assessment.

Based on the above consensus, we propose an evaluation metric of the following characteristics. Specifically, the proposed COMMET metric:

1. Combines multiple other metrics.
2. Allows for weighting of the combined metrics depending on their reliability, based on their usual performance.
3. Combining multiple metrics in one score allows for smoothing out inconsistencies among multiple metrics.
4. Using an average of multiple metrics, allows for preserving the general tendencies in scores, and works as a fail-safe for situations, where typically highly reliable one metric would lead to incorrect evaluation.
5. Both the number and types of metrics applied in the overall combined metric, as well as the the weights of each metric can be adjusted to the needs of each specific language pair and other conditions, allowing for flexibility and robustness.

6. As humans also evaluate translations in a multifaceted way [71], automatic evaluation method should also combine multiple metrics, each of which takes various aspects of translation into consideration.

The combined metric is calculated by dividing the sum of all multiplications of included metric scores with their assigned weights by the total number of the metrics included, according to equation 4.1.

Based on the related studies as well as the characteristics laid out above, we concluded a pool of automatic evaluation metrics applicable in our situation, along with their weights. The weight of metric is designed to be flexible and replaceable in different scenario and settings. Due to our experiments and discussions which will deal with English to Chinese translation outputs, we decided to emphasize our weight reference on the results reported by Kocmi et al.[77].

Kocmi et al. provides an extensive comparison of multiple automatic translation, which is crucial for establishing a well-rounded evaluation framework. Their study is also a recent contribution from WMT21 where it reflects the recent state of the field which adds credibility and timeliness for references. Therefore by implementing weight derived from their study, we can ensure the weights to be grounded in empirical evidence and reduces the risk of arbitrary or subjective weighting decisions.

They published extensive comparisons of accuracies for various language pairs in a form of rankings. They also specifically looked separately at metrics performance for languages which writing system is based on logograms (e.g., Chinese ideograms), rather than alphabet. Thus, the specific set of metrics and their weights was settled as in Table 4.1. Additionally, all scores were applied in their normalized form, reported as from 0 (lowest) to 1 (highest). Moreover, scores reporting errors were considered as 1-metric, e.g., 1-CharacTER, to unify the winning condition, namely, “the higher the better."

$$COMMET = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i) \tag{4.1}$$

where

x is the score of the metric,

y is the weight of the metric, and

Table 4.1: Metrics applied in COMMET in this paper, with their weights applied as accuracies reported by [77].

Metric	Weight
COMET	0.909
BERTScore	0.886
BLEURT	0.841
CHRF	0.886
SacreBLEU	0.795
CharacTER	0.705

n is the total number of included metrics.

Chapter 5

Experiments

5.1 Preliminary experiment 1: Choosing optimal translation model

In the following experiment, we compare between several language models in order to choose an optimal language model for our study.

5.1.1 Experiment setup

Five Transformer-based language models were implemented in this experiment. Pretrained MT5-K024, MT5-large, ByT5-large, and mBART-large-50 were chosen for their large pretrained size together with bilingual model opus-en-zh. Both training subsets of our new Dataset 3 were applied as the training set for all five models. All models were applied with their suggested learning rates, namely, mBART with a learning rate of 3e-5, MT5 and ByT5 with a learning rate of 1e-3, and Opus-en-zh with a learning rate of 5e-5. Each of the models was trained for 3 epochs. Implemented test sets were a combination of both testing subsets from Dataset 3. In the evaluation, we used the proposed COMMET metric with its all compound metrics, with additional ROGUE and METEOR for reference.

Table 5.1: Comparison between different language models on different evaluation metrics. Highest scores in **bold** font.

model	ROUGE	METEOR	SacreBLEU	CHRF	CharacTER	BLEURT	BERTScore	COMET	COMMET
mbart-large-50	**0.305**	**0.261**	**0.484**	**0.441**	**0.766**	**0.145**	**0.877**	**0.519**	**0.448**
opus-enzh	0.226	0.127	0.005	0.096	0.450	0	0.564	0	0.151
mt5-K024	0.168	0.178	0.292	0.278	0.706	0	0.759	0	0.275
mt5-large	0.121	0.102	0.091	0.156	0.438	0	0.734	0	0.195
byt5-large	0.001	0.029	0.016	0.013	0.496	0	0.554	0	0.144

5.1.2 Results and discussion

In Table 5.1 we can see the results for all included models in several evaluation metrics including our proposed combined metric. The results from mBART-large-50 are highly dominant compared to others presumably due to it being mainly developed for machine translation from the start. The MT5-K024 which was pretrained for three languages including Japanese, Chinese, and English has the second highest score. Disappointingly, opus-en-zh which was finetuned for English and Chinese did not score better than MT5-K024. As for MT5-large and ByT5-large which were pretrained on large data but not specifically finetuned for English to Chinese translation, their score was very low in comparison to others. Also for BLEURT and COMET, we standardize all score below 0 to 0, where most models with bad result maintaining 0 score for specified metrics.

In conclusion, we decided to implement mBART-large-50 as our main language model for all following experiments.

5.2 Preliminary experiment 2: Comparing between datasets with and without hashtags

In this experiment, we determine the optimal preprocessing method for ironic hashtags by implementing two versions of Dataset 3 subset 1 containing ironic tweets. Since Dataset 1 SemEval Twitter irony dataset was originally collected by searching Twitter for ironic hashtags (#irony, #sarcasm, #not), all of the ironic entries contain at least one of the ironic hashtags. However, according to previous works on Irony Detection [29, 31], the removal of ironic hashtags could improve the performance of the models and the integrity of the eperiment (since the model could not “cheat" by looking at a limited set of specific keywords). Therefore, in this experiment, we conducted a similar experiment and compared a version of the dataset with and without ironic hashtags. By the comparison, we also expected to gain more insight into the impact of ironic hashtags in social media texts.

5.2.1 Experiment setup

The two versions of Dataset 3 subset 1 were used, namely, the original version with the hashtags, and the version with all ironic hashtags replaced with a hashtag (#). Both versions included their training and test sets respectively.

5.2.2 Results and discussion

From Table 5.2 we can see the results of cross-training and testing between Dataset 3 subset 1 (Twitter dataset) with ironic hashtags, and without ironic hashtags. The model that trained on a dataset without ironic hashtags and tested on a dataset without ironic hashtags scored the highest in this experiment. Both models trained on a dataset without ironic hashtags achieved better results compared to models trained on a dataset with ironic hashtags even when tested on a dataset with ironic hashtags. The symptom above questions whether training on datasets with ironic hashtags will be necessary due to the noises caused by excessive ironic hashtags.

Therefore, similarly to the previous works in irony detection [29, 31], these results also promote the removal of ironic hashtags. Nonetheless, the removal or

Table 5.2: Comparison between data with and without ironic hashtag. Highest scores in **bold** font.

Train	Test	ROUGE	METEOR	SacreBLEU	CHRF	CharacTER	BLEURT	BERTScore	COMET	COMMET
without #	without #	0.451	**0.517**	0.4604	0.5060	**0.2973**	**0.1804**	0.8729	**0.5396**	**0.407**
	with #	**0.532**	0.510	**0.4639**	**0.5205**	0.1657	0.1562	**0.8742**	0.5365	0.390
with #	without #	0.444	0.505	0.4520	0.4940	0.2030	0.1690	0.8700	0.5180	0.387
	with #	0.406	0.457	0.4345	0.4691	0.2182	0.1261	0.8650	0.4835	0.371

labeling of ironic hashtags in the dataset will filter out some tweets in which the author uses the hashtagged word as part of the main sentence (e.g. “#i #loves #sarcasm"). Despite that, the original and current purpose of using hashtags in social media is to categorize their posts with relevant keywords used within text in the form of hashtags. Hence the ironic hashtags included in data of the mentioned type will be counted as standard ironic hashtags for categorization. Therefore, we will be using the version of the dataset without ironic hashtags for the following experiments.

5.3 Experiment 1: Short tweets vs long forum posts

In this experiment, we compare the training and testing between shorter texts, represented by tweets and longer texts represented by forum posts. As the model used in the experiment, we implemented mbart-large-50 and finetuned it separately on different subsets of our Dataset 3.

The short tweets subset 1 contains mostly short random social media posts full of unintelligible internet slang while the long forum post subset 2 contains logically structured topic-focused long forum debate posts. The aim of comparing between these two subsets, was to observe how the language model handles highly differentiated texts in both length and grammar structure. This way we can also find out whether if it is generally better to train translation models for irony using longer texts with sufficient context, or if it is sufficient to train use only short texts for training if testing is also based on similarly short texts.

5.3.1 Experiment setup

The language model was fine-tuned on three different training sets separately, subset 1 (tweets), subset 2 (forum posts), and both subset 1 and 2. All three different fine-tuned versions of the model were then tested on both subset 1 and subset 2 respectively.

5.3.2 Results and discussion

From Table 5.3 we can see the results of the language model trained and tested on both subset 1 (tweets), subset 2 (forum posts), and both subsets. All results from different finetuning settings which tested on the forum posts subset were lower than the results for the tweets subset. Even the language model which only finetuned on the forum posts subset scored higher when testing on the tweets subset. This might be due to the the length and increased complexity of the forum posts subset. Finetuning on both the tweets subset and the forum posts subset improved the score compared to the model which only finetuned on tweets subset. However, the model which only finetuned on the forum posts subset achieved the highest

Table 5.3: Comparison between the results for short tweets and long form posts. Highest scores in **bold** font.

Train	Test	ROUGE	METEOR	SacreBLEU	CHRF	CharacTER	BLEURT	BERTScore	COMET	COMMET
tweet	tweet	0.46	0.45	0.435	0.469	**0.609**	0.226	0.865	0.580	0.446
	forum	0.19	0.13	0.423	0.370	0.335	0.137	0.861	0.410	0.374
forum	tweet	**0.69**	**0.53**	0.490	**0.554**	0.561	0.171	**0.891**	**0.623**	**0.461**
	forum	0.26	0.15	**0.510**	0.456	0.434	**0.250**	0.875	0.601	0.442
tweet+forum	tweet	0.41	0.46	0.444	0.479	0.608	0.233	0.869	0.606	0.454
	forum	0.24	0.14	0.489	0.429	0.366	0.210	0.880	0.561	0.416

scores in both test subsets. The results can be caused by more colloquial sentence structures and excessive unusual sentence components included in the tweets subset (user mentions, URLs, etc.).

5.4 Experiment 2: Exploring optimal translation settings

From the results of the previous experiment, we determined the best combinations for irony translation to be finetuning with forum posts. In this experiment, we focus on the model finetuned with forum posts, but explore further whether it is generally better to train on irony to translate irony, and train on non-irony to translate non-irony, or if mixing irony and non-irony texts in training has merit, and if it does, what type of texts it is best for (irony/non-irony or both, as well as tweets/forum posts, or both).

5.4.1 Experiment setup

In this experiment, we implemented 3 models which were finetuned on different subsets of the forum posts subset, namely, the irony subset, the non-irony subset, and both. We tested each of the models on 9 different variations of test sets, namely, either tweets subset, forums subsets, or both, and those containing ironic data, non-ironic data, or both.

5.4.2 Results and discussion

Table 5.5 shows all the results from the 27 variations of finetuned models tested on various test sets. Among them the highest combined score comes from the model finetuned on both ironic and non-ironic forum posts, and tested on non-irony forum posts, which was 0.4471 of the COMMET score. However, the lowest score among them, reaching 0.3995, was also the model finetuned on both ironic and non-ironic forum posts training subset, but tested on ironic tweets subset.

Therefore, finetuning with both ironic and non-ironic forum posts gives both the best and the worst results. However, tests only on tweets showed lower scores in general. Hence, from this discovery, we can conclude that the translation results depend highly on what one is trying to translate. This is because some texts are more difficult to translate due to the lack of sufficient context navigating the model, especially for figurative language. The average COMMET score for the model finetuned on both ironic and non-ironic forum posts reached 0.428 which is an

2.15% and 3.13% increase from the model finetuned only on non-ironic forum posts subset with an average COMMET score of 0.419 and the model finetuned only on ironic forum posts subset with an average COMMET score of 0.415.

Furthermore, from Table 5.4 we can observe the COMMET score results from Table 5.5 sorted in descending order. To better understand the results, we calculated average and standard deviation of those results, and grouped together mid-range scores (within the scope of average+/-standard deviation), as well as highest and lowest scores (above the threshold of avg+stddev, and below the threshold of avg-stddev).

Focusing on the test sets, we can see that the best results tested on ironic test set, 0.4210, is lower than the worst result tested on non-ironic test set, 0.422. Almost all results tested on ironic test sets were below the threshold, or in general on the lower spectrum of all results. All results from testing on ironic tweets subset occupied the lower spectrum with their highest score being 0.404, a 5.94% decrease from the highest score of 0.447 in Table 5.4. These results confirm the intuition as well as results from previous experiments, that irony is consistently more difficult to translate. Furthermore, the ironic tweets subset remains the worst for machine translation among all ironic data implemented.

We can also observe that there is not much difference between the results of testing on tweets subsets and forum posts subsets. Therefore the length of translated text (forum posts vs tweets) is not that important when compared to the content of the texts (irony vs non-irony).

The results of this comparison suggest to use for finetuning both ironic and non-ironic data. Two reasons we have chosen the model finetuned on both ironic and non-ironic forum data are, primarily, because it achieved the highest score among all three models while testing for all kinds of test data. The second reason is its highest ability to translate ironic data as we can see from 5.4. The highest score on testing on ironic test data, 0.421, is also a result of the model finetuned on both ironic and non-ironic forum data. Lastly, we also conclude that training only on ironic data is not optimal for translating irony. This suggests, that non-ironic texts in training also contribute to the quality of translation. However, this could also mean that simply adding more data, regardless if ironic or not, could increase the quality of translation of irony. We test this hypothesis in the next experiment.

Table 5.4: The table shows all the results from testing on 9 different test sets (forum posts, tweets, and both forum posts and tweets, with irony, non-irony, and both irony and non-irony), using 3 different finetuned models, trained on ironic forum posts, non-ironic forum posts, and all forum posts. The average of all results calculated with the COMMET score was 0.4207, with a standard deviation of 0.0134 splitting all results into 3 tiers, top-scoring, average, and low-scoring models. Highlighted in yellow is the same model finetuned with both ironic and non-ironic forum posts tested on various test sets. Highlighted in pink is the model finetuned with only ironic forum posts. From the test column, highlighted in orange are the results for tests on non-ironic forums, while highlighted in blue are tests only on ironic forum posts. The optimal model, or the one that both reached the best score overall, as well as the best core for ironic posts was highlighted in green.

	COMMET score	train	test	rank
	0.4471	forum-both	forum-non	1
	0.4417	forum-both	both-non	2
	0.4378	forum-both	tweet-non	3
Top scoring models (above threshold)	0.4371	forum-non	forum-non	4
	0.4352	forum-both	forum-both	5
	0.4329	forum-non	both-non	6
	0.4329	forum-non	tweet-non	7
	0.4303	forum-both	both-both	8
	0.4273	forum-irony	tweet-non	9
	0.4232	forum-irony	forum-non	10
	0.4226	forum-both	tweet-both	11
	0.4222	forum-irony	both-non	12
	0.4218	forum-non	tweet-both	13
	0.4212	forum-non	forum-both	14
Average models (Avg+/-StDev)	0.4210	forum-both	forum-irony	15
	0.4194	forum-non	both-both	16
	0.4173	forum-irony	tweet-both	17
	0.4156	forum-irony	forum-both	18
	0.4146	forum-irony	both-both	19
	0.4139	forum-both	both-irony	20
	0.4072	forum-irony	forum-irony	21
	0.4048	forum-irony	both-irony	22
	0.4044	forum-non	tweet-irony	23
	0.4027	forum-non	forum-irony	24
Low scoring models (below threshold)	0.4023	forum-non	both-irony	25
	0.4019	forum-irony	tweet-irony	26
	0.3995	forum-both	tweet-irony	27

Table 5.5: This table shows results from 27 different combinations of finetuning and testing, models finetuned on 3 versions of forum datasets, irony, non-irony, and both, and tested on 9 versions of test set, namely, forum, tweets, both, and irony, non-irony, and both.

Train	Test	ROUGE	METEOR	Sacre BLEU	CHRF	Charac TER	BLEURT	BERT Score	COMET	COMMET
	tweets-irony	0.654	0.491	0.456	0.510	0.442	0.097	0.866	0.480	0.402
	tweets-non	0.688	0.537	0.488	0.548	0.418	0.171	0.876	0.523	0.427
	tweets-both	0.675	0.519	0.475	0.534	0.428	0.141	0.872	0.506	0.417
	forum-irony	0.280	0.151	0.476	0.421	0.414	0.226	0.879	0.474	0.407
forum-irony	forum-non	0.213	0.132	0.490	0.428	0.413	0.205	0.885	0.575	0.423
	forum-both	0.247	0.141	0.485	0.425	0.414	0.216	0.882	0.524	0.416
	both-irony	0.401	0.261	0.472	0.442	0.423	0.185	0.875	0.476	0.405
	both-non	0.415	0.302	0.491	0.458	0.416	0.191	0.881	0.553	0.422
	both-both	0.407	0.283	0.483	0.452	0.419	0.188	0.879	0.518	0.415
	tweets-irony	0.708	0.497	0.455	0.534	0.421	0.114	0.866	0.475	0.404
	tweets-non	0.707	0.569	0.506	0.574	0.407	0.176	0.878	0.521	0.433
	tweets-both	0.707	0.540	0.486	0.558	0.413	0.151	0.874	0.503	0.422
	forum-irony	0.245	0.149	0.474	0.415	0.428	0.207	0.875	0.463	0.403
forum-non	forum-non	0.232	0.139	0.518	0.454	0.393	0.235	0.891	0.599	0.437
	forum-both	0.239	0.144	0.501	0.439	0.410	0.221	0.883	0.531	0.421
	both-irony	0.396	0.261	0.472	0.444	0.426	0.177	0.872	0.466	0.402
	both-non	0.432	0.320	0.516	0.484	0.399	0.210	0.886	0.567	0.433
	both-both	0.415	0.292	0.498	0.468	0.411	0.195	0.879	0.521	0.419
	tweets-irony	0.668	0.487	0.455	0.519	0.437	0.076	0.864	0.482	0.400
	tweets-non	0.706	0.552	0.504	0.558	0.401	0.208	0.881	0.543	0.438
	tweets-both	0.691	0.526	0.484	0.543	0.415	0.156	0.874	0.519	0.423
	forum-irony	0.279	0.160	0.492	0.438	0.396	0.252	0.885	0.519	0.421
forum-both	forum-non	0.246	0.146	0.531	0.467	0.376	0.257	0.896	0.629	0.447
	forum-both	0.263	0.153	0.515	0.456	0.386	0.255	0.891	0.574	0.435
	both-irony	0.404	0.266	0.486	0.457	0.409	0.195	0.878	0.508	0.414
	both-non	0.439	0.317	0.527	0.490	0.386	0.236	0.889	0.593	0.442
	both-both	0.423	0.293	0.510	0.477	0.397	0.218	0.885	0.557	0.430

5.5 Experiment 3: Is more data better? Experiments on all data

Previous experiments provided interesting insights regarding finetuning on forum posts datasets. We also concluded that the length of each training data sample does not the impact the results as much as could be anticipated. Hence, in this experiment, we compare the contribution of the ironic and non-ironic dataset from the perspective of both tweets and forum posts subsets applied together. In this experiment we analyze separately the impact of ironic and non-ironic training data in machine translation of irony and non-irony. We also test which is more beneficial for figurative (ironic) language translation. We also check whether finetuning on both tweets and forum posts will provide better results than finetuning only on forum posts.

5.5.1 Experiment setup

We finetuned the model separately on all Dataset 3 ironic and non-ironic training data. Next with two different versions of the model, we conducted tests on both ironic test data and non-ironic test data.

5.5.2 Results and discussion

From Table 5.6 we can observe the results of the models finetuned and tested on two types of data, ironic and non-ironic. The model which was finetuned on non-irony data and tested on non-irony data achieved the highest score for all of the evaluation metrics, which again supports the claim that irony is more difficult to translate then non-irony. Also, all results from testing on non-ironic data are around 2 3% higher than all results from testing on ironic data. This clearly shows that translating ironic data is much harder comparing to normal data. The results from training on both ironic and non-ironic data are better than results trained on only ironic data but lost to results trained on only non-ironic data. These indicates that training on more data is not always a better approach as the noises in ironic data will affect the performance of the irony translation. The highest result obtained in this table, 0.441 is also lower than in previous experiments due

Table 5.6: Comparing between ironic data and non-ironic data in training and testing.

Train	Test	ROUGE	METEOR	SacreBLEU	CHRF	CharacTER	BLEURT	BERTScore	COMET	COMMET
irony	irony	0.304	0.244	0.470	0.425	0.500	0.152	0.875	0.486	0.408
	non-irony	0.311	0.278	0.490	0.450	0.526	0.159	0.880	0.554	0.429
non-irony	irony	0.309	0.246	0.483	0.436	0.492	0.149	0.877	0.504	0.413
	non-irony	**0.315**	**0.285**	**0.517**	**0.474**	**0.539**	**0.163**	**0.885**	**0.564**	**0.441**
both	irony	0.297	0.242	0.468	0.420	0.490	0.132	0.873	0.479	0.401
	non-irony	0.313	0.278	0.495	0.450	0.524	0.157	0.880	0.552	0.430

to using in this experiment all data (both forums and tweets), rather than only forum posts.

5.6 Experiment 4: Qualitative analysis of ambiguous and contextual irony

After quantitative experiments, we also performed qualitative analysis of ambiguous and contextual irony extracted from our combined dataset. Since, compared to self-contained irony, there were much fewer examples of contextual and ambiguous irony, we qualitatively analyzed them one by one manually to find any characteristics in how the lack of sufficient context influences the quality of machine translation of irony.

5.6.1 Discussion on ambiguous irony

From Table 5.7 and Table 5.8 we can see the list of all ambiguous ironic data in our dataset, along with their human-translated (gold standard) and machine-translated (model prediction) counterparts, together with their COMMET score. Listed in order sorted decreasingly by the COMMET score in Table 5.9 and Table 5.10, with the maximum achievable COMMET score being 0.837 (maximum achievable for this setting due to various weights of compound metrics applied in COMMET score). The highest score available in Table 5.7 and Table 5.8 was 0.557, which was achieved by example 7, while the lowest achieved score was 0.134 by example 13. Overall we can observe that the average length of ambiguous ironic data is very short, with few exceptions of longer tweets. This is reasonable since the short length makes the context provided in the tweets limited, hence making them ambiguous. However, the shorter the sentences, the higher the potential influence of the neutral labels (_url_, @user@), which our model does not always translate, which lowers the COMMET score. This discrepancy comes from the fact that human translators translated almost all understandable words including labels, non-ironic hashtags, labeled URLs, user mentions, and emoji. This was the choice of human translators, however, as such does not have any influence on the ironic meaning of the sentence. Therefore, we decided to perform additional data processing and clean up all of those neutral labels since the goal was to only compare those parts of the sentence, that take part in conveying the meaning of the sentence, including the irony. For that, all neutral labels were removed including

ironic hashtags, URLs, user mentions, and emoji. There was also an additional problem with punctuation and spaces for the model translations, which were not identical to those used by human translations. Hence, we remove most of the excessive and unhelpful punctuation and spaces. The new set of post-processed ambiguous data was included in Table 5.9 and Table 5.10.

In the Tables, we have 2 instances with perfect COMMET scores, specifically, examples 1 and 2, where both the human and model achieved the same translation. However, this was only possible due to the extremely short length of sentences, being only four characters, or three words each. We can see from examples 3, 9, 14, and later, that there are unusual words or phrases that both the human annotators and the model could not translate (e.g., “decemberbessen"). Also, some of the predictions (translation by our model), although mostly semantically identical with the references (human annotated gold standard), obtained lower scores, due to using slightly different expressions in Chinese for the same English phrases. In example 27, we can see that the translation differences between the reference and prediction mostly comes from the difference in punctuation where, where humans used punctuation typically used in Chinese texts (full-width period), while the MT output simply left the original English periods untranslated. Moreover, in example 37, the main content is just one word, which also is a cultural slang “swag" (short from “something we all got", representing a popular merchandise in a subculture), both human and model provided their own translations which can be both back-translated as “trendy" or “fastidious" respectively. It is not possible to clearly decide which of those translations is more accurate, since there exists no definitive translation for the term. Across the tables, we can also notice that most of the comprehensible hashtags are translated by human annotators, but the model would leave most of the hashtags untranslated. For example, in sample 34 there were two hashtags combined without space, the model failed to translate the hashtag “#praisehim". Also, sample 26 which contains the hashtag “#addiction" which was not translated by the model, despite being a comprehensible word. However there were also examples where the words in hashtags were incomprehensible for human annotators, yet the model translated them. For example, sample 39 contained “#noa" and “#onuchapel", which were translated only by the model, into words meaning “no" and “chapel" in the target language (Chinese). This is most likely the

result of the model applying subwords to extract any possible meaning from the incomprehensible character strings. Some of the incomprehensible hashtagged words untranslated by both human translators and the model found in the ambiguous irony samples were “#instacraze", “#shelovesitwhenisendthatmany", “#inasong", “#myfav", etc. We can see most of them are either short forms of longer words and phrases or a combination of many words without spaces. As we can see from the fact that neither human translators nor the model were able to translate those, we can conclude that these words pose a challenge for irony translation, due to the fact that such original neologisms can contain much of the ironic meaning of the whole example.

We can further observe the ambiguity of the data after the cleaning shown in Tables 5.9 and 5.10. Unlike in contextual irony, which mostly contains some words or phrases requiring additional context for understanding, ambiguous irony is sometimes not easily interpretable even to human translators. Both tables listed only data annotated as irony by previous studies [67, 147]. In this research we did not re-evaluate the definitions of irony and sarcasm and annotations performed by previous research, due to the fact that this data has been already used in multiple previous studies and is considered a strong baseline. However, looking at such borderline examples as mentioned above, we can conclude that further study should thoroughly re-evaluate both the definitions and all annotations done previously to correct or, in the worst case, discard any of such unintelligible examples. In this regard, the present study can be useful as a baseline for specifying which of those samples should be re-evaluated in the first place, as being either ambiguous or lacking sufficient context to understand the ironic meaning. Good examples of such ambiguous statements ca also be seen in example 1 and 2, after the removal of the ironic hashtags, namely, “i hate it" and “this is fun." Since their meaning, and thus the translations for them are straightforward and universally accepted as non-ironic, it is not possible to distinguish whether these phrases were used in ironic meaning without the additional hashtag (#sarcasm, and #not). Furthermore, examples 7, 19, 25, 28, and 32 are a bit longer than other ambiguous samples in terms of length, however even the longer context contained in those samples could not help human translators understand the irony within. Some of the samples such as 19 can be misunderstood as contextual irony, however, even when all the named entities

contained in the example (such as “victoria’s secret fashion show") are decoded, this still does not make the whole example less ambiguous. From example 32 we can also see the translation differences between the humans and the model, where humans translated the number “600k" into its corresponding Chinese translation “60万" (which literal translation is “60 of ten thousand"), but the model translated it into full numerical form of “600,000" and both translations can be considered correct. Also, examples such as 24 where there were no content other than the hashtags and user mentions, would necessarily rely strongly on the translation of those mentioned specific meta-phrases. Finally, in example 24 (#shocking), both translations were correct, however, the human translator used a longer expression, which made the final score lower.

Table 5.7: This table shows the first 20 out of all 39 ambiguous irony examples including their order (# in the first column) adjusted to Table 5.9 and contains the original sentence in English (top row), human translated reference sentence (center row), and model translated prediction (bottom row), as well as model translation scores calculated with the COMMET metric.

#	English Human translation Model translation	COMMET score
1	i hate it #sarcasm _url_ 我讨厌它#讽刺 _网址_ 我讨厌它 #sarcasm _url	0.371
2	this is fun #not _url_ 这很有趣#不 _网址_ 这是有趣的 #not _url_	0.31
3	@user@ illridewithyou #not @用户@illridewithyou #不 @user@ illridewithyou #not	0.468
4	produce mobile apps #not _url_ _url_ 制作移动应用程序#不 _网址_ _网址_ 制作移动应用#not _url_ _url_	0.348
5	it's international human rights day #irony 今天是国际人权日#讽刺 这是国际人权日 #irony	0.422
6	i'm such a good friend ... #sarcasm _url_ 我真是个好朋友.....#讽刺 _网址_ 我真是个好朋友..... #sarcasm _url_	0.381
7	girl put your records on . tell me your favorite song . #irony #inasong #myfav 女孩把你的唱片放上。告诉我你最喜欢的歌。 #讽刺 #inasong #myfav 女孩把你的记录放在上。告诉我你最喜欢的歌曲。 #irony #inasong #myfav	0.557
8	@user@ she showed you ! #sarcasm @用户@她给你看了！ #讽刺 @user@ 她给你展示了! #sarcasm	0.254
9	decemberbessen #not #fail #footprint sorry ... _url_ decemberbessen #不 #失败 #脚印 抱歉... _网址_ decemberbessen #not #fail #footprint 对不起... _url_	0.349
10	testing water temperatures #excitingtimes #not 测试水温##兴奋的时刻 #不 测试水温#兴奋时间#不	0.535
11	@user@ hmm ... let me think about that #sarcasm @用户@ hmm ...让我想想#讽刺 @user@嗯.....让我想想那个 #sarcasm	0.329
12	@user@ i have a boyfriend ! (#not) * hair flip and walks away * @用户@我有男朋友了！ （#不）*甩头发然后走开* @user@我有男朋友! (#not) * 卷头发走开*	0.332
13	@user@ clever you ! #not @用户@你聪明！ #不 @user@ clever you! #不	0.134
14	@user@ sig-ni-fi-cant witter #not @用户@sig-ni-fi-cant witter #不 @user@sign-ni-fi-cant witter #not	0.429
15	why is isis an acronym for words in english ? #irony 为什么 isis 是英文单词的首字母缩写词? #讽刺 为什么 isis 是英语单词的缩写? #irony	0.396
16	cool . #sarcasm 凉爽的 。 #讽刺 凉爽的。 #sarcasm	0.41
17	the world is such a smiley place . :emoji: #not 世界就是这样一个笑脸的地方。 : 表情符号: #不 世界真是个笑脸的地方。 :emoji: #not	0.337
18	jordan fan here | | #sarcasm :emoji: - not a hater . _url_ 乔丹球迷在这里 | | #讽刺 :表情符号: - 不是仇恨者。 _网址_ 这里是约旦粉丝| | #sarcasm :emoji: - 不是仇恨者。 _url_	0.321
19	reading a victoria's secret fashion show recap with a plate of french fries in front of me :emoji: #irony 在我面前拿着一盘炸薯条阅读维多利亚的秘密时装秀回顾: 表情符号: #讽刺 在我面前拿着一 plate of french fries 阅读维多利亚的秘密时装秀总结:表情符号:#irony	0.322
20	@user@ your last retweet though #irony @用户@ 你最后一次转发 #讽刺 @user@ 你的最后一次retweet though #irony	0.189

Table 5.8: This table shows the latter 19 out of all 39 (21–39) ambiguous irony examples including their order (# in the first column) adjusted to Table 5.10, and contains the original sentence in English (top row), human translated reference sentence (center row), and model translated prediction (bottom row), as well as model translation scores calculated with the COMMET metric.

21	should i buy these ? and become a full time hipster ? lol ! #not ... _url_ 我应该买这些吗? 成为全职潮人? 哈哈 ! #不 ... _网址_ 我应该买这些吗?并成为一个全职嘻哈明星?哈哈! #not... _url_	0.379
22	@user@ i know ... #sarcasm @用户@我知道... #讽刺 @user@我知道..... #sarcasm	0.411
23	reunited with my pump this morning . #yay #not :emoji: 今天早上与我的水泵重逢。 #耶 #不 :表情符号: 今天早上我和我的泵重聚了。 #yay #not :emoji:	0.218
24	@user@ # shocking #not @用户@#令人震惊#不 @user@ # 震惊的#不	0.356
25	just snap chatting my sister a few times .. #shelovesitwhenisendthatmany #not _url_ 只是和我姐姐聊了几次.. #shelovesitwhenisendthatmany #不 _网址_ 只是几遍我妹妹的闲谈.. #shelovesitwheninedthatmany #not _url_	0.338
26	#instacraze #not #addiction :emoji: at _url_ #instacraze #不 #上瘾 :表情符号: 在_网址_ #instacraze #not #addiction :emoji:at _url	0.265
27	@user@ have fun at that ... #sarcasm @用户@ 玩得开心..... #讽刺 @user@ 玩得开心... #sarcasm	0.392
28	aaaaaaand ... i left my work computer at home . now i have to go get it . #bestdayever #sarcasm aaaaaaand ...我把工作电脑忘在家里了。现在我得去拿了。 #最美好的一天 #讽刺 aaaaaaand... 我把工作电脑留在家里。现在我必须去拿它。 #最好的一天 #sarcasm	0.479
29	also sick names #not 还有病名#不 也病名#not	0.243
30	so they lost ... such a shame ... i was really rooting for them #not ... #lfc 所以他们输了..... 太可惜了..... 我真的很支持他们#不 ... #lfc 所以他们输了..... such a shame... 我真的很追随他们 #不..... #lfc	0.36
31	glad there's not a typhoon where we go on holiday in 4 weeks . #sarcasm #fml 很高兴 4 周后我们去度假的地方没有台风。 #讽刺#fml 很高兴 4 周后我们不会有台风去度假。 #sarcasm #fml	0.408
32	isn't this blaming the victim ? #sarcasm rt @user@ every year , 600k college students are injured while drunk . _url_ 这不是在指责受害者吗? #讽刺 rt @用户@ 每年，有 60 万大学生在醉酒时受伤。 _网址_ 这不是指责受害者吗? #sarcasm rt @user@ 每年, 600k 名大学生喝醉后受伤。 _url_	0.303
33	grandma's coming over , yay money . #lol #kidding #not 奶奶要过来了，钱啊。 #大声笑 #开玩笑 #不 奶奶来了,哇,钱。 #lol #kidding #not	0.167
34	such #irony . you still have to #praisehim ;) 这样的#讽刺。你仍然需要#赞美他 ;) 这样的 #irony 。你仍然必须 #赞美他 ;)	0.469
35	you're so sweet #not 你真可爱#不 你太甜了#no	0.167
36	luv this #not 喜欢这个#不 说这话#不	0.382
37	swag #not @user@ 赃物#不 @用户@ swag #not @user@	0.144
38	shakespeare is great #not :emoji: :emoji: 莎士比亚很棒 #不 :表情符号: :表情符号: shakespeare 很棒 #not :emoji: :emoji:	0.159
39	are they allowed to be touching each other's kneecaps like that ? ? #questionable #noa #onuchapel #sarcasm 他们可以像那样抚摸对方的膝盖骨吗? ? #有问题 #noa #onuchapel #讽刺 他们是否允许那样触摸彼此的膝盖?? #值得质疑 #noa #onuchapel #sarcasm	0.368

Table 5.9: This table shows the first 20 out of all 39 additionally preprocessed ambiguous irony examples (labels removed, punctuation reduced), including their order (# in the first column) sorted in descending order by the COMMET score, and contains the original sentence in English (top row), human translated reference sentence (center row), and model translated prediction (bottom row).

#	English Human translation Model translation	COMMET score
1	i hate it 我讨厌它 我讨厌它	0.837
2	this is fun 这很有趣 这很有趣	0.837
3	illridewithyou illridewithyou illridewithyou	0.702
4	produce mobile apps 制作移动应用程序 制作移动应用	0.682
5	it's international human rights day 今天是国际人权日 这是国际人权日	0.652
6	i'm such a good friend.. 我真是个好朋友. . . . . . 我真是个好朋友..	0.637
7	girl put your records on tell me your favorite song #inasong #myfav 女孩把你的唱片放上告诉我你最喜欢的歌 #inasong #myfav 女孩把你的记录告诉我你最喜欢的歌曲 #inasong #myfav	0.63
8	she showed you ! 她给你看了! 她给你看!	0.602
9	decemberbessen #fail #footprint sorry.. decemberbessen #失败 #脚印 抱歉... decemberbessen #失败 #足迹 抱歉。.	0.583
10	testing water temperatures #excitingtimes 测试水温#兴奋的时刻 测试水温#兴奋时间	0.58
11	hmm.. let me think about that hmm..让我想想 嗯..让我想想那个	0.578
12	i have a boyfriend !* hair flip and walks away * 我有男朋友了！ *甩头发然后走开* 我有　个男朋友!*卷头发然后走开*	0.544
13	clever you ! 你聪明! 你很聪明!	0.543
14	sig-ni-fi-cant witter sig-ni-fi-cant witter sign-ni-fi-cant witter	0.52
15	why is isis an acronym for words in english ? 为什么isis是英文单词的首字母缩写词? 为什么isis是英语单词的缩写?	0.516
16	cool 凉爽的 凉爽	0.514
17	the world is such a smiley place 世界就是这样　个笑脸的地方 这个世界真是个笑脸的地方	0.511
18	jordan fan here		- not a hater 乔丹球迷在这里		- 不是仇恨者 这里是约旦粉丝|	- 不是仇恨者	0.497
19	reading a victoria's secret fashion show recap with a plate of french fries in front of me 在我面前拿着　盘炸薯条阅读维多利亚的秘密时装秀回顾 在我面前阅读维多利亚的秘密时装秀总结和　盘法国炸薯条	0.495
20	your last retweet though 你最后　次转发 不过你最后　次再发帖	0.469

Table 5.10: This table shows the latter 19 out of all 39 (21–39) additionally preprocessed ambiguous irony examples (labels removed, punctuation reduced), including their order (# in the first column) sorted in descending order by the COMMET score continued from Table 5.9, and contains the original sentence in English (top row), human-translated reference sentence (center row), and model-translated prediction (bottom row)

21	should i buy these ? and become a full time hipster ? lol ! .. 我应该买这些吗? 成为全职潮人? 哈哈 ! .. 我应该买这些吗?并成为 个全职嘻哈明星?哈哈!..	0.44
22	i know.. 我知道... 我知道。	0.438
23	reunited with my pump this morning #yay 今天早上与我的水泵重逢 #耶 今天早上与我的泵重聚 #yay	0.434
24	#shocking #令人震惊 #震惊	0.424
25	just snap chatting my sister a few times. #shelovesitwhenisendthatmany 只是和我姐姐聊了几次.. #shelovesitwhenisendthatmany 只是我妹妹的几次闲谈。 #shelovesitwhenineddhatmany	0.414
26	#instacraze #addiction at #instacraze #上瘾 在 #instacraze #addiction at	0.408
27	have fun at that.. 玩得开心..... 玩得开心。.	0.383
28	aaaaaaand.. i left my work computer at home now i have to go get it #bestdayever aaaaaaand..我把工作电脑忘在家里了现在我得去拿了 #最美好的 天 aaaaaaand.. 我把工作电脑留在家里,现在我必须去拿它 #bestdayever	0.365
29	also sick names 还有病名 也叫病名	0.361
30	so they lost.. such a shame.. i was really rooting for them .. #lfc 所以他们输了..... 太可惜了..... 我真的很支持他们.. #lfc 所以他们输了.. such a shame.. 我真的很追随他们.. #lfc	0.36
31	glad there's not a typhoon where we go on holiday in 4 weeks #fml 很高兴 4 周后我们去度假的地方没有台风 #fml 很高兴我们不会在 4 周内去度假的台风 #fml	0.357
32	isn't this blaming the victim ? rt every year , 600k college students are injured while drunk 这不是在指责受害者吗? rt 每年，有 60 万大学生在醉酒时受伤 这不是责备受害者吗? 每年,600,000名大学生喝醉后受伤	0.352
33	grandma's coming over , yay money #lol #kidding 奶奶要过来了，钱啊 #大声笑 #开玩笑 奶奶来了,哇,钱 #lol #开玩笑	0.315
34	such you still have to #praisehim 这样的你仍然需要#赞美他 这样的你仍然必须 #praisehim	0.27
35	you're so sweet 你真可爱 你太甜了	0.25
36	luv this 喜欢这个 说这话	0.247
37	swag 时尚 挑剔	0.224
38	shakespeare is great 莎士比亚很棒 shakespeare 很棒	0.189
39	are they allowed to be touching each other's kneecaps like that ? ? #questionable #noa #onuchapel 他们可以像那样抚摸对方的膝盖骨吗? ? #有问题 #noa #onuchapel 他们是否允许那样触摸彼此的膝盖?? #值得质疑 #不 #大教堂	0.142

5.6.2 Discussion on contextual irony

In this section, we discuss contextual irony. We base our discussion on some of the contextual irony examples from the data. Table 5.11 shows the best-scoring 5 and the worst-scoring five examples of all from the contextual irony subset annotated in our combined dataset. Also, applying the experience from the analysis of ambiguous irony, we analyze the examples in their post-cleaning form. We can observe that the examples were categorized into contextual irony mainly due to their ironic content in the sentence being mostly dependent on a context wider than the one included in the sentence. For example, in sample 1 it is necessary to understand certain named entities and their context in culture, such as “skrillex" (a popular American DJ and musician) which people who do not know will also not understand the irony. The same can be seen in example 2 where the named entity is “mexico" but both human and model managed to translate it correctly due to it being a well-known country name, as well as in example 3, where the name of the two constables mentioned makes it unable to be translated without the knowledge of the whole story. Examples 4 and 5 have the same problem for both human and model translations. It is not possible to translate the named entities which are “w-wada" and “concordia wifi" without knowing the wider context associated with those keywords. Most of the remaining contextual irony samples share the same characteristics, namely, lack of sufficient knowledge about the wider context of the sentence, which is easier to identify than in ambiguous irony examples. Scoring-wise, for the worst five samples, we can see that the problem is mostly with the translation differences in the grammar and sentence structure used by humans and the model. Other problems include some translation discrepancies for the translations of named entities between humans and the model. In samples 88, 89, 90, 91 we can also see that the human translator translated the named entities contained in the sample whereas the model was unable to translate them. However, on the other hand, in sample 92, we can see that the model translated the named entity “zoloft" and the hashtag “#psychhumor" which the humans were not able to translate.

Table 5.11: This table shows the first five and last five out of all 92 contextual irony examples, further preprocessed (labels removed, punctuation minimized), including their order (# in the first column), and sorted by COMMET score. Each example contains the original sentence in English (top row), human-translated reference sentence (center row), and model-translated prediction (bottom row).

#	English Human translation Model translation	COMMET score
1	oh it's just skrillex no big deal .. i see him on the daily .. #skrillex #rave #edm ... 哦，这只是skrillex没什么大不了的..我每天都看到他..#skrillex#rave#edm... 哦,这只是skrillex没什么大不了的..我每天都看到他.#skrillex#rave#edm...	0.755
2	i'm getting excited about mexico now . 我现在对墨西哥感到兴奋。 我现在对墨西哥很兴奋。	0.682
3	the two constables - - veer pal singh yadav and avnish yadav - did not have any criminal record . - ndtv | | “这两名警员–veerpalsinghyadav和avnishyadav-没有任何犯罪记录。”-ndtv|| 两位警察–veerpalsinghyadav和avnishyadav-没有任何犯罪记录。-ndtv||	0.675
4	suddenly w-wada 突然w-wada 突然间w-wada	0.669
5	mvp goes to concordia wifi . mvp属于concordiawifi。 mvp访问concordiawifi。	0.661
...	...	...
88	loving how fulham looks like a playoff team . really 喜欢富勒姆看起来像一支季后赛球队。真的 热爱fulham看起来像预选团队的样子。真的	0.262
89	yet another rainy goshen concert night #surprised 又一个下雨天的歌珊音乐会之夜:表情符号::表情符号:#惊讶 又一个多雨的goshen音乐会之夜#惊讶	0.26
90	gael ' pro at presentations ' anderson #choke 盖尔'专业演讲'安德森#choke。 Gael'proatpresentations'anderson#choke。	0.26
91	just wait till cleveland comes again ! 等着克利夫兰再来吧！ 等cleveland回来再等一下!	0.255
92	abnormal professor walks in with the finals in a zoloft bag #psychhumor 不正常的教授带着zoloft袋子走进决赛#psychhumor。 异常教授带着动物座袋中的最后一场比赛走进来#心理	0.238

Table 5.12: This table shows five examples of self-contained irony with their COMMET score.

#	English Human translation Model translation	COMMET score
1	so i'm the school nurse today . and i have a fever. 所以我今天是学校护士。我发烧了。 所以今天我是学校护士。我发烧了。	0.728
2	how can u miss something u never had ? #randomthoughts #miss. 你怎么能想念从未有过的东西呢? #随机想法 #想念。 你怎么能错过你从未有过的东西? #随机想法 #错过。	0.495
3	planning a vacation instead of studying ... my priorities are in line. 计划假期而不是学习.....我的优先事项是一致的。 计划休假而不是学习.....我的优先事项排序一致。	0.406
4	thanks for the dick pic , really , thanks. 谢谢你的鸡巴照片，真的，谢谢。 感谢你的迪克照片,真的,谢谢。	0.37
5	most of us didn't focus in the #adhd lecture. 我们大多数人都没有专注于#过动症 讲座。 我们大多数人在#adhd 演讲中没有关注。	0.244

5.6.3 Comparison between normal, ambiguous and contextual irony

In this section, we compare the three types of irony classified in our dataset, namely self-contained, or normal irony, ambiguous irony, and contextual irony. Listed in Table 5.12 are five random examples of normal irony extracted from our dataset for comparison. All examples here show easily intelligible irony. In example 1 we can understand the irony from the context of the nurse having a fever themselves. The translation for this data is very similar for both model and humans where only the pronoun "I" is arranged in a different spot. Example 2 has a translation problem with the verb "miss" where both the model and human provided different translations. Example 3 is also easily understandable for its irony especially as the author tried to explain it themselves with the second sentence. Most of the translation differences here come from different wordings whereas both model and humans were generally not wrong in their translations. As for example 4, we can tell the irony from the double thanks gratitude from the author even also emphasized by the word "really". For this, the model failed to translate the vulgar word and instead translated the word semantically. As for example 5, in Table 5.12 we can understand the irony from the situation where the authors' actions could

not fit into the environment due to their nature (“ADHD" and “focus").

Next, we discuss the characteristics of contextual irony with comparison to self-contained irony. From Table 5.11 we can see that all ten examples presented share the same characteristic where the existence of at least one named entity, either regional or cultural limits the understanding of irony. This can be seen in example 1, where knowledge and information about “skillrex" is needed in order to understand the irony of the whole sentence. Same with examples 2 and 3 where the contextual components are the country name and some individual names. Sometimes, understanding of not widely known text such as "w-wada" in example 4 is also required to understand the conveyed irony.

Lastly, different from normal irony and contextual irony, ambiguous irony is the most difficult to identify. The ambiguity of examples from this group causes them to be not identified as irony even when the sentences are meant to be ironic. From Table 5.9 we can see examples 1 and 2 which are annotated as short ironic sentences but they can also be interpreted as non-ironic if no further information is provided. As we described in section 3.3, the characteristics of ambiguous irony is originally meant to be ironic but highly unintelligible to most audiences. Longer sentences such as example 30, 31, 32 or 33 from Table 5.10 also share the same characteristic. The irony is ambiguous and the examples can be interpreted as non-ironic. Thus, the only way to understand that the sentence is ironic is with the help of specific hashtags used to originally collect the dataset (#sarcasm, #irony, #not).

5.7 Additional Experiment: mBART vs ChatGPT

5.7.1 Experiment setup

In this section, we compare the translations of our finetuned model and the recently popular ChatGPT model available through API from OpenAI. Due to the recent rise in popularity for ChatGPT, we wanted to observe its performance in translating ironic texts. For comparison, we used our model which achieved the highest score in the previous experiments. It is the model that was finetuned on both ironic and non-ironic forum posts. For ChatGPT, we used a simple prompt like “Can you translate some ironic tweets from English into Simplified Chinese for us?" with all data manually pasted into ChatGPT for translation.

5.7.2 Results and discussion

Table 5.13 shows results in COMMET score for comparison between translations of our fine-tuned mBART model and ChatGPT. The data used for testing were all 784 instances from our Dataset 3 tweet subset which included non-ironic data and different types of ironic data introduced in the previous sections. Next Table 5.14 shows the comparison of translations of specific examples done by our fine-tuned mBART model, and ChatGPT, together with the gold standard translations from human translators, also including the source sentences in English and the different types of irony. We can see ChatGPT achieved much lower score than the mBART model, which suggests that fine-tuning of much smaller models is still more effective, and efficient than using prompting on huge language models. However, after carefully checking on the ChatGPT translations, our human translators and annotators reached a mutual agreement in which both models provided either similar or sufficient in quality and understandable translations. Due the fact that gold standard translations we produced were based on editing and improving translations from Google Translate, and having similarly processed data being used for training in our model, our fine-tuned model was able to output translations with sentence style and grammar structure similar to the references provided from human gold standard translation. Therefore, ChatGPT translations were comparable, and even better in some cases but due to being different in structure, or with similar

vocabulary, they did not achieve high scores when compared to the provided gold-standard translations. This can be observed in Table 5.14 for examples 5 and 8, wherein data 5 ChatGPT provided better translation but with a different structure, while mBART failed to translate the named entity, hence ChatGPT achieved higher score. However, in example 8, translation provided by ChatGPT can be also qualified as satisfactory, but due to different choice of words, it resulted in a lower score. Also, because of its size and robustness, ChatGPT sometimes translated most of the named entities or even rare vocabulary which neither Google Translate nor our translators were able to translate and left untouched in the gold standard data. For example, example 2 from Table 5.14, where “decemberbessen" is translated by ChatGPT but not human or our model. Also, in example 4 we can see that our model failed to translate the hashtag while ChatGPT succeeded.

We can also see that ChatGPT achieved better results for contextual irony when compared to our model due to the same reasons as above. Namely, since ChatGPT already is pretrained on much larger data, and thus can handle context more effectively, even if the context is missing from the sentence. There are also some examples where both mBART and ChatGPT performed poorly mainly because of excessive use of hashtags, social media slang, and abbreviations, or simply because of bad components from source sentences, hence the translations did not make sense. As for ambiguous irony, both mBART and ChatGPT achieved overall higher scores compared to other irony types. This may be due to ambiguous irony being too ambiguous to the point where both humans and the models cannot differentiate whether they were ironic. The scores on translations provided from both mBART and ChatGPT share the same characteristics where scores on all test data including ironic and non-ironic data are higher than scores on ironic data. This shows that even ChatGPT cannot translate ironic text well enough to reach human level and preserve the full meaning required for understanding of the irony expressed in the sentence. Lastly, we also observed that the problem with this comparison could be mainly in the evaluation measures and the lacking of references. Due to semantically similar translations but expressed with different vocabulary, the evaluation metric could not always provide fair scores for the lack of proper references in target language. For example, one of the most common occurrences was the word “can" which can be translated into both “能" and “可以", but due to the lack of references,

only the one translated exactly as in gold standard was favored.

Table 5.13: The table shows results in COMMET score for the comparison between the translations of our fine-tuned mBART model and ChatGPT on our Dataset 3 tweet subset with different types of irony.

Data Type	fine-tuned mBART	ChatGPT
All Data	0.443	0.369
All Irony	0.435	0.366
Normal Irony	0.410	0.364
Contextual Irony	0.370	0.397
Ambiguous Irony	0.520	0.448

Table 5.14: This table shows several examples of different types of irony with their scores for each model, and translations provided from: human translators as references (REF), our mBART model predictions (PRD), ChatGPT translations (GPT), along with the source sentences in English (SRC). The presented examples were from irony types: A - ambiguous, C - contextual, and N - normal.

#	type	class	data	mBART	ChatGPT
1	A	SRC	hmm.. let me think about that	0.52	0.37
		REF	hmm..让我想想		
		PRD	嗯.....让我想想		
		GPT	嗯.. 让我想想		
2	A	SRC	decemberbessen #fail #footprint sorry.	0.37	0.36
		REF	decemberbessen #失败 #印记 抱歉..		
		PRD	decemberbessen #失败 #足迹 抱歉		
		GPT	十二月的灰 #失败 #印记 抱歉..		
3	C	SRC	just wait till cleveland comes again !	0.26	0.52
		REF	等着克利夫兰再来吧！		
		PRD	等cleveland回来再等一下!		
		GPT	等着看克利夫兰再次出现！		
4	C	SRC	how dare charles barkley have an intelligent conversation about race #charlesbarkley	0.33	0.6
		REF	查尔斯巴克利怎么敢对种族问题进行明智的对话。#查尔斯巴克利		
		PRD	查尔斯·巴克利怎么敢有关于种族的聪明谈话。 #charlesbarkley。		
		GPT	查尔斯·巴克利怎么敢就种族问题进行聪明的对话。 #查尔斯·巴克利		
5	C	SRC	: al jazeera is pro anti - #gamergate because feminism , or something :	0.34	0.55
		REF	：半岛电视台支持反#gamergate，因为女权主义或其他原因：		
		PRD	: al jazeera 支持反击 - #gamergate 因为女权主义,或者什么:		
		GPT	：因为女权主义或其他原因，半岛电视台支持反对#gamergate		
6	N	SRC	best christmas eve news : i have a half sister #thanksdad	0.22	0.64
		REF	最好的平安夜新闻：我有一个同父异母的妹妹。#谢谢爸爸。		
		PRD	最好的圣诞老人新闻:我有半个妹妹#谢谢你		
		GPT	最好的平安夜消息：我有一个异父同母的妹妹。#谢谢爸爸		
7	N	SRC	thanks for pirating my music ! :d	0.61	0.59
		REF	感谢盗版我的音乐！ :d		
		PRD	感谢您盗版我的音乐! :d		
		GPT	谢谢你盗版我的音乐! :d		
8	N	SRC	perfect time to get sick	0.83	0.25
		REF	生病的完美时机		
		PRD	生病的完美时机		
		GPT	生病再合适不过了		
9	N	SRC	everyone wants the truth but nobody wants to be honest..	0.64	0.62
		REF	每个人都想要真相，但没有人愿意诚实.....		
		PRD	每个人都想要真相，但没有人愿意诚实..		
		GPT	每个人都想要真相，但没有人想要诚实...		

5.8 Irony classification with translated texts: indirect practical evaluation of translated irony

In this section, we implemented our model to check whether finetuning on ironic translation data improves the model for irony detection.

5.8.1 Experiment setup

Here we mainly apply 2 models, namely, raw mbart50 model (with no finetuning), and our best model which we finetuned on our Dataset 3 forum subset ironic and non-ironic training data for translation from English to Chinese, while Ambiguous and Contextual irony data are not included here. For the testing below, we also apply two sets of test data, namely, Dataset 3 forum subset test data both source data (original English forum posts) and reference data (human translated Chinese translations). For simplicity, we will just refer to them as English test set and Chinese test set.

5.8.2 Results and discussion

First, we finetuned both models mentioned above with classification data from our Dataset 3 forum training subset which included reference data (human-translated Chinese translations) and binary labels for irony (irony or non-irony). We named the finetuned models for classification of irony in Chinese as “our-cn" and “raw-cn". Then, we tested them both on English test set and Chinese test set. From Table 5.15 we can see the “raw-cn" performed slightly better than “our-cn" on classifying English ironic data, with F-score of 0.787 compared to 0.770. However, both models obtained the same F-score of 0.78 when tested on Chinese ironic data, with “raw-cn" obtaining higher recall but “our-cn" obtaining higher precision. This shows that a model finetuned on translations does not improve the performance of classifying English ironic data when finetuned on Chinese classification data.

Secondly, we finetuned another copy of both models (not finetuned for classification) with data from our Dataset 3 forum training subset, which included source data (original English forum posts) and binary labels for irony. These models were named “our-en" and “raw-en". Then, we tested them both on the same test

set as above (English and Chinese test sets). Here, we wanted to check whether finetuning on English data for irony classification will be better after finetuning on translations. Table 5.15 shows results for classification on both English and Chinese ironic test set with both raw model and the model fine-tuned previously on translations. Our model ("our-en") finetuned first for translation of irony and secondly for classification of irony in English, performed better than the raw model finetuned only with English ironic classification data "raw-en". Unlike for the model finetuned for irony classification in Chinese above, the translation finetuning actually improved both English and Chinese irony classification, with an F-score of 0.78 increased to 0.791 and 0.749 to 0.752 for English and Chinese test data, respectively.

Next, we finetuned our model (a new not fine-tuned copy of the model) with Chinese prediction data translated by our model along with binary labels for irony. This finetuned model was named as "our-cn2". Here we wanted to compare the effectiveness between Chinese translations from our model and human-translated Chinese data. We tested both "our-cn" and "our-cn2" models on the same test set as above (English and Chinese test sets). From the same Table 5.15 we can observe that the results from model finetuned on human translated data "our-cn" are slightly better than the model finetuned on only translated data "our-cn2". However, the results tested on English and Chinese irony classification data respectively reached an F-score of 0.77 and 0.78 for "our-cn" were just only slightly better than the results of "our-cn2" which were 0.762 and 0.778.

Lastly, we tested both results of 0.778 and 0.780 from "our-cn" comparing to "our-cn2" tested on Chinese dataset with McNemar's test to determine if there is a significant difference between two models. Figure below shows the contigency matrix of McNemar's test.

		Control		
		+	-	**Total**
Case	+	707	177	**884**
	-	166	254	**420**
	Total	**873**	**431**	**1304**

The test showed p-value of 0.5892 which in by conventional criteria, this difference is considered to be not statistically significant. Lastly based on the performance metrics and test results, there is no significant evidence to suggest that either "our-cn" or "our-cn2" performs significantly better. Therefore, both models can be considered comparable in terms of their classification performance on the given test set, which suggests that

both training on human translated data and training on our machine translated data can be considered equally sufficient in terms of data quality. This is result supports the proposed claim that is effective to use translation models for data augmentation, specifically for irony detection.

Furthermore, we finetuned our model with both English source data and Chinese translated data along with the binary labels for irony and named the model “our-cnen". A model trained this way could be considered as the combination of the first and second sets of our fine-tuned models “our-en" and “our-cn". We tested it on the same test sets for both English and Chinese irony classification. The results for English test data in F-score was 0.795 which is an improvement from both “our-cn" and “our-en". This is possibly primarily due to the increase in the number of training data samples, which often leads to better performance. However for Chinese test data, comparing to the result in F-score of 0.78 from “our-cn", finetuning on combined English and Chinese data obtained a worse result of 0.767. This shows that the finetuning on English ironic training data worsens the performance of a model in classifying Chinese ironic data, opposite to classifying English ironic data.

Lastly, we checked the effectiveness of our best classification model in detecting irony which is ambiguous and contextual, in which task we used the “our-cnen" model. We extracted all ambiguous and contextual irony data (mentioned in Table 5.9, 5.10, and 5.11) from Dataset 3 tweet subset to serve as the test data. All results were represented in Table 5.16. However, since for this task we use only samples containing irony, the results were reported using a simplified accuracy based on TP and FN. For classifying ambiguous irony, for being either ironic or not, our model got it right for 33 out of 39 of cases in English test data and 35 out of 39 for the Chinese test data. These reached an accuracy of 0.846 and 0.897 respectively which are considered as surprisingly good due to the ambiguity of ambiguous irony which is normally supposed to be more difficult to understand. For contextual irony classification, the model got achieved an accuracy of 0.913 (84 out of 92) for test data in English, and 0.815 (75 out of 92) for test data in Chinese. For English test data, the result was better than expected, but the test data in Chinese somehow scored somewhat lower. This shows the ironic component in Chinese contextual irony is harder to detect, or the irony as a whole is much

more difficult to be identified. Even so, the named entities which were the main obstacles for irony translation contained in the contextual irony data could in fact just be a minor reason for the lower-than-expected accuracy in irony classification.

Table 5.15: This table shows results of testing on both English and Chinese classification data using various models and combinations of test data.

model-train	test	F1	acc	prc	rec	TP	TN	FP	FN	total
our-cn	en	0.770	0.768	0.755	0.793	517	484	168	135	1304
raw-cn	en	0.787	0.777	0.752	0.827	539	474	178	113	1304
our-cn	cn	0.780	0.771	0.753	0.808	527	479	173	125	1304
raw-cn	cn	0.780	0.769	0.743	0.824	537	466	186	115	1304
our-en	en	0.791	0.789	0.783	0.799	521	508	144	131	1304
raw-en	en	0.780	0.780	0.780	0.781	509	508	144	143	1304
our-en	cn	0.752	0.760	0.777	0.729	475	516	136	177	1304
raw-en	cn	0.749	0.763	0.797	0.705	460	535	117	192	1304
our-cn	en	0.770	0.768	0.755	0.793	517	484	168	135	1304
our-cn	cn	0.780	0.771	0.753	0.808	527	479	173	125	1304
our-cn2	en	0.762	0.774	0.804	0.724	472	537	115	180	1304
our-cn2	cn	0.778	0.771	0.755	0.802	523	482	170	129	1304
our-cnen	en	0.795	0.791	0.781	0.810	528	504	148	124	1304
our-cnen	cn	0.767	0.768	0.770	0.764	498	503	149	154	1304

Table 5.16: This table shows results of both English and Chinese ambiguous and contextual irony data classification tested by our model pretrained on English and Chinese irony classification data

model-train	test	acc	TP	FN	total
our-cnen	amb-en	0.917	33	6	39
our-cnen	amb-cn	0.946	35	4	39
our-cnen	cont-en	0.955	84	8	92
our-cnen	cont-cn	0.898	75	17	92

Chapter 6

Discussion

6.1 Addressing research goals

In this section we address and conclude our research goals mentioned in Section 1.1.

- Performance Enhancement through Language Model Fine-Tuning:

  After many delicately designed experiments, we have our designated model to be able to provide machine translations of ironic text with quality comparable to state-of-the-art system.

- Improved Figurative Language Comprehension:

  With the help of intensive literature reviews and new classification types of irony in our new dataset, we get to experiment and analyze the characteristic of different irony types textually.

- Creation of New Dataset for Irony Translation:

  From both datasets we collected from reliable sources, short tweets and long forum posts, we developed a new dataset, inheriting and combined both their characteristics, and a parallel Chinese translations for all entries translated and quality checked by professionals fluent in both languages. Further classification of irony types are also monitored for part of our new dataset in Chapter 3.

- Proposal of Novel Evaluation Metric for Machine Translation:

  We have proposed a new metric, COMMET, or Combined Metric for Machine Translation, which combines several MT metrics with a designated weight for each of the implemented metrics. Further information in Section 4.2.

- Analysis and comparison of various model types:

  Using available resources we compared between different types of popular language model and with our chosen optimal language model gone through various finetuning, we get to understand the characteristics of most language models, including our mbart model, chatGPT, and others. Related sections to be in Section 5.1 and Section 5.8

6.2 Future Applicability

6.2.1 Extension of this study into other languages

The intricate nature of irony poses a challenge in translation. Successfully addressing this challenge in languages other than those studied in this paper (English and Chinese) would not only enhance our understanding of linguistic subtleties of irony, but also for figurative language in general. Accurate irony translation could also contribute to sentiment analysis, opinion mining, and broader natural language understanding.

6.2.2 Irony and sarcasm in the context of cyberbullying

The integration of machine translation models in understanding and translating irony holds profound implications, particularly in the context of language employed in cyberbullying, as already pointed out by Chia et al.[31]. As these negative online behaviors often exploit subtle linguistic nuances, an effective translation of irony has the potential to enhance the early detection and prevention of cyberbullying incidents, which should be studied in the future.

6.2.3 Energy efficiency in machine learning models

In response to the growing demand for energy-efficient machine learning models [50], leveraging classic ML models becomes imperative. These models, despite being efficient, require specific datasets for optimal performance. Therefore, efficiently translating diverse datasets, especially from languages with limited coverage, can extend the applicability of classic ML models.

6.2.4 Broader implications

Beyond the above direct implications, the effective translation of irony has the following broader implications. It can improve the accuracy of machine translation services, enhance cross-cultural communication, and refine sentiment analysis in social media monitoring. The nuanced understanding of irony achieved through efficient translation has the potential to reshape how machines process and generate human-like language, impacting various sectors from customer service to content creation, ultimately fostering a more nuanced and adaptable landscape for artificial intelligence and natural language processing.

6.3 Ethical Considerations

In the context of irony research in machine translation, informed consent was obtained from all of our annotators and translators, who were informed about this study's focus on the translated content, specifically addressing irony. Everyone involved was made aware of potential challenges and risks associated with the cultural sensitivity of translating irony. All of our implemented data are collected from trustful sources which also obtained appropriate approvals for the collection and use of the provided data. Our combined dataset is constructed from two datasets provided by two different sources. The first dataset mainly consists of tweets that were distributed during the irony detection workshop organized by Semantic Evaluation 2018 as Task 3 Irony Detection in English Tweets with 43 participated teams worldwide [147]. The second dataset includes ironic and normal forum posts collected as the Sarcasm Corpus, currently updated as version 2 available for public use and maintained by Baskin School of Engineering [110]. In

case of sensitive contents, we approached them with utmost care and ensured the translation was respectful and loyal to the provided source data.

In incorporation of Google Translate and ChatGPT as intermediary tools for translation of irony in this study, the principle of beneficence was carefully considered to prioritize the well-being of users and participants. Several measures were implemented to uphold ethical standards and mitigate any potential concerns. Specifically, for all translation tasks, participants were provided with the option to discontinue their interaction with our data, models, and even ChatGPT at any point, respecting the voluntary nature of their participation.

In the case of current and future models with better capability in translating irony, we acknowledge that there might be users who abuse them for automatic generation of contents applicable in cyberbullying or flooding the internet with ironic messages in many languages. Cases that exploit language models by finetuning more toxic and harmful data in order to provide offensive and blasphemous outputs are also being considered as a possible ethical concern. To mitigate that we release the data and the model under appropriate licences, e.g., the BigScience RAIL License v1.0[1]

[1] `https://huggingface.co/spaces/bigscience/license`

Chapter 7

Conclusions and Future Work

In this paper, we set out to explore sarcasm and irony through the perspective of machine translation, using various techniques and different combinations of training and testing datasets, human translated and annotated for a variety of uses.

Firstly, we reviewed and clarified the definitions of irony and sarcasm by discussing various studies in both linguistics and computational linguistic fields, following a wide array of reviews and summaries of related studies surrounding machine learning and natural language processing with the focus on machine translation and irony-related studies.

Next, we constructed our English-Chinese parallel dataset by manually translating and annotating two datasets originally collected from highly varied reliable sources for the purpose of irony classification. Specifically, one of the datasets in our collection consisted of data collected from tweets, which characterize in sentences short in length, and full of social media slang. While the other dataset contained logically structured long forum posts in the form of sentences and paragraphs extracted from online debates on specific topics. All of the data in our constructed dataset were both originally in English, which was then manually translated into Chinese, with a certain range of the dataset further classified into different types of irony. After preliminary experiments, we decided to implement the version of our dataset without ironic hashtags.

After the construction of the dataset, we compared different types of language models that were best suited for translating between English and Chinese. Before the comparison which led us to our preferred base model, namely, mBART-large-50,

we reviewed some of the most commonly used models and evaluation metrics for machine translation. To facilitate a more concise interpretation of translation results, we proposed a new combined metric for machine translation, COMMET, which is potentially more democratic and fair in the evaluation of machine translation models. The combined metric is calculated by first choosing multiple initial evaluation metrics usable for the particular study, and calculating the scores for each metric. Next, those score are multiplied by weights representing relative reliability of each metric. Finally, these weighted scores are summed and divided by the total number of metrics included.

Having the dataset and the evaluation metric ready, we first explored optimal translation settings and the best finetuned models for translating irony. Firstly, we experimented with different combinations of training data with the best result being finetuning only with the long forum posts. Next, we expanded the experiment with more combinations of finetuned models and test data. We found out that the best model is the one finetuned on both irony and non-irony forum posts, surpassing all other combinations. We also analyzed the difficulties of the models in translating ambiguous and contextual irony. Finally, we utilized ChatGPT in creating a new set of translations for ironic texts for comparison between ChatGPT and our best model, with the general conclusion that although our model reach much higher scores, zero-shot translations by ChatGPT are also acceptable, and the lower scores were caused by the model using sometimes different vocabulary. This highlights the imperfections of automatic evaluation measures for machine translation and the importance of human evaluation in machine translation studies. Lastly, we applied our translation model in the irony classification task by finetuning it additionally with classification datasets. Improvements were found in the results of models blending both translation and classification training, which proves translation training benefit models in classification, or in mode general terms, fine-tuning first on certain tasks can improve later fine-tuning, if the tasks are related to some extent (e.g., here: the general topic of irony).

Finally, we plan to extend this research the range of other applicable languages other than Chinese and English. Moreover, we will delve further into the potential studies of fine-grained irony types, specifically, ambiguous and contextual, to obtain a better assessment and analysis of irony.

Bibliography

[1] I. Abbes, W. Zaghouani, O. El-Hardlo, and F. Ashour. DAICT: a dialectal Arabic irony corpus extracted from Twitter. English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6265–6271, Marseille, France. European Language Resources Association, May 2020. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.768.

[2] R. Abbott, B. Ecker, P. Anand, and M. Walker. Internet argument corpus 2.0: an SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA), May 2016. URL: https://aclanthology.org/L16-1704.

[3] M. Abdelghaffar, A. E. Mogy, and N. Sharaf. Adapting large multilingual machine translation models to unseen low resource languages via vocabulary substitution and neuron selection, 2022.

[4] M. H. Abrams and G. G. Harpham. A glossary of literary terms. In Wadsworth Cengage Learning, 2009.

[5] H. AlMazrua, N. AlHazzani, A. AlDawod, L. AlAwlaqi, N. AlReshoudi, H. Al-Khalifa, and L. AlDhubayi. Sa'7r: a saudi dialect irony dataset. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 60–70, Marseille, France. European Language Resources Association, June 2022. URL: https://aclanthology.org/2022.osact-1.7.

[6] ALPAC. Language and machines: computers in translation and linguistics, Washington, DC, 1966. DOI: 10.17226/9547. URL: https://nap.nationalacademies.org/catalog/9547/language-and-machines-computers-in-translation-and-linguistics.

[7] A. Anastasopoulos, L. Barrault, L. Bentivogli, M. Z. Boito, O. Bojar, R. Cattoni, A. Currey, G. Dinu, K. Duh, M. Elbayad, C. Emmanuel, Y. Estève, M. Federico, C. Federmann, S. Gahbiche, H. Gong, R. Grundkiewicz, B. Haddow, B. Hsu, D. Javorský, V. Kloudová, S. M. Lakew, X. Ma, P. Mathur, P. McNamee, K. Murray, M. Nadejde, S. Nakamura, M. Negri, J. Niehues, X. Niu, J. E. Ortega, J. M. Pino, E. Salesky, J. Shi, M. Sperber, S. Stüker, K. Sudoh, M. Turchi, Y. Virkar, A. H. Waibel, C. Wang, and S. Watanabe. Findings of the iwslt 2022 evaluation campaign, 2022.

[8] M. Artetxe, G. Labaka, and E. Agirre. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics, July 2019. DOI: 10.18653/v1/P19-1019. URL: https://aclanthology.org/P19-1019.

[9] M. Artetxe, G. Labaka, and E. Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics, Oct. 2018. DOI: 10.18653/v1/D18-1399. URL: https://aclanthology.org/D18-1399.

[10] S. Attardo. Irony as relevant inappropriateness. In *Journal of Pragmatics*, 1999.

[11] O. Babîi. Variables as contextual constraints in translating irony. *Linguaculture*, 6(1):98–123, June 2015. DOI: 10.1515/lincu-2015-0039. URL: https://journal.linguaculture.ro/index.php/home/article/view/64.

[12] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ArXiv*, 1409, Sept. 2014.

[13] N. Banar, W. Daelemans, and M. Kestemont. Character-level transformer-based neural machine translation. *CoRR*, abs/2005.11239, 2020. arXiv: 2005.11239. URL: https://arxiv.org/abs/2005.11239.

[14] A. Bapna, I. Caswell, J. Kreutzer, O. Firat, D. van Esch, A. Siddhant, M. Niu, P. N. Baljekar, X. García, W. Macherey, T. Breiner, V. Axelrod, J. Riesa, Y. Cao, M. X. Chen, K. Macherey, M. Krikun, P. Wang, A. Gutkin, A. Shah, Y. Huang, Z. Chen, Y. Wu, and M. Hughes. Building machine translation systems for the next thousand languages. *ArXiv*, abs/2205.03983, 2022.

[15] F. Barbieri. Machine learning methods for understanding social media communication: modeling irony and emojis. In Department DTIC, 2017.

[16] C. Baziotis, N. Athanasiou, P. Papalampidi, A. Kolovou, G. Paraskevopoulos, N. Ellinas, and A. Potamianos. NTUA-SLP at semeval-2018 task 3: tracking ironic tweets using ensembles of word and character level attentive rnns. *CoRR*, abs/1804.06659, 2018. arXiv: 1804.06659. URL: http://arxiv.org/abs/1804.06659.

[17] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155, Mar. 2003. ISSN: 1532-4435.

[18] G. Bettelli and F. Panzeri. *Intercultural Pragmatics*, 20(5):467–493, 2023. DOI: doi:10.1515/ip-2023-5001. URL: https://doi.org/10.1515/ip-2023-5001.

[19] P. Brown, J. Cocke, S. D. Pietra, V. D. Pietra, F. Jelinek, R. Mercer, and P. Roossin. A statistical approach to language translation. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 1*, COLING '88, pages 71–76, Budapest, Hungry. Association for Computational Linguistics, 1988. ISBN: 963 8431 56 3. DOI: 10.3115/991635.991651. URL: https://doi.org/10.3115/991635.991651.

[20] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85, June 1990. ISSN: 0891-2017.

[21] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. URL: https://aclanthology.org/J93-2003.

[22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. arXiv: 2005.14165 [cs.CL].

[23] C. Burfoot and T. Baldwin. Automatic satire detection: are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore. Association for Computational Linguistics, Aug. 2009. URL: https://aclanthology.org/P09-2041.

[24] C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics, Apr. 2006. URL: https://aclanthology.org/E06-1032.

[25] J. G. Carbonell, R. E. C. Ford, and A. V. Gershman. Knowledge-based machine translation, 1978.

[26] R. Chakhachiro. Analysing irony for translation. *Meta*, 54(1):32–48, 2009. DOI: https://doi.org/10.7202/029792ar.

[27] T. Chakrabarty, A. Saakyan, and S. Muresan. Don't go far off: an empirical study on neural poetry translation, 2021.

[28] Y. Cheng, W. Xu, Z. He, W. He, H. Wu, M. Sun, and Y. Liu. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics, Aug. 2016. DOI: 10.18653/v1/P16-1185. URL: https://aclanthology.org/P16-1185.

[29] Z. L. Chia, M. Ptaszynski, and M. Fumito. Exploring machine learning techniques for irony detection. *Proceedings of the Annual Conference of JSAI*, JSAI2019:2A4E203–2A4E203, 2019. DOI: 10.11517/pjsai.JSAI2019.0_2A4E203.

[30] Z. L. Chia, M. Ptaszynski, M. Fumito, G. Leliwa, and M. Wroczynski. A study in practical solutions to sarcasm detection with machine learning and knowledge engineering techniques. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, 2020.

[31] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski. Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Information Processing and Management*, 58(4):102600, 2021. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102600. URL: https://www.sciencedirect.com/science/article/pii/S0306457321000984.

[32] D. Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics, June 2005. DOI: 10.3115/1219840.1219873. URL: https://aclanthology.org/P05-1033.

[33] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics, Oct. 2014. DOI: 10.3115/v1/W14-4012. URL: https://aclanthology.org/W14-4012.

[34] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics, Oct. 2014. DOI: 10.3115/v1/D14-1179. URL: https://aclanthology.org/D14-1179.

[35] S. L. Christian W. F. Mayer and S. Brandt. Prompt text classifications with transformer models! an exemplary introduction to prompt-based learning with large language models. *Journal of Research on Technology in Education*, 55(1):125–141, 2023. DOI: 10.1080/15391523.2022.2142872. eprint: https:

//doi.org/10.1080/15391523.2022.2142872. URL: https://doi.org/10.1080/15391523.2022.2142872.

[36] H. W. Chung, T. Févry, H. Tsai, M. Johnson, and S. Ruder. Rethinking embedding coupling in pre-trained language models. *CoRR*, abs/2010.12821, 2020. arXiv: 2010.12821. URL: https://arxiv.org/abs/2010.12821.

[37] R. Collobert and J. Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, Helsinki, Finland. Association for Computing Machinery, 2008. ISBN: 9781605582054. DOI: 10.1145/1390156.1390177. URL: https://doi.org/10.1145/1390156.1390177.

[38] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

[39] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, 2019. arXiv: 1901.02860. URL: http://arxiv.org/abs/1901.02860.

[40] B. Dancygier and E. Sweetser. *Figurative language / Barbara Dancygier, University of British Columbia, Vancouver ; Eve Sweetser, University of California, Berkeley.* eng. Cambridge textbooks in linguistics. Cambridge University Press, New York, 2014. ISBN: 9781107005952.

[41] V. Dankers, E. Bruni, and D. Hupkes. The paradox of the compositionality of natural language: a neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics, May 2022. DOI: 10.18653/v1/2022.acl-long.286. URL: https://aclanthology.org/2022.acl-long.286.

[42] M. Denkowski and A. Lavie. Meteor universal: language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics, June 2014. DOI: 10.3115/v1/W14-3348. URL: https://aclanthology.org/W14-3348.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics, June 2019. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[44] S. Doddapaneni, G. Ramesh, A. Kunchukuttan, P. Kumar, and M. M. Khapra. A primer on pretrained multilingual language models. *CoRR*, abs/2107.00676, 2021. arXiv: 2107.00676. URL: https://arxiv.org/abs/2107.00676.

[45] D. Dong, H. Wu, W. He, D. Yu, and H. Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics, July 2015. DOI: 10.3115/v1/P15-1166. URL: https://aclanthology.org/P15-1166.

[46] D. K. Dougal and D. Lonsdale. Improving NMT quality using terminology injection. English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4820–4827, Marseille, France. European Language Resources Association, May 2020. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.593.

[47] S. Edunov, M. Ott, M. Auli, and D. Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association

for Computational Linguistics, Oct. 2018. DOI: 10.18653/v1/D18-1045. URL: https://aclanthology.org/D18-1045.

[48] C. I. Eke, A. A. Norman, and L. Shuib. Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and bert model. *IEEE Access*, 9:48501–48518, 2021.

[49] S. Eo, C. Park, H. Moon, J. Seo, and H. Lim. Comparative analysis of current approaches to quality estimation for neural machine translation. *Applied Sciences*, 11(14), 2021. ISSN: 2076-3417. DOI: 10.3390/app11146584. URL: https://www.mdpi.com/2076-3417/11/14/6584.

[50] J. K. K. Eronen, M. Ptaszynski, F. Masui, A. Smywinski-Pohl, G. Leliwa, and M. Wroczynski. Improving classifier training efficiency for automatic cyberbullying detection with feature density. *CoRR*, abs/2111.01689, 2021. arXiv: 2111.01689. URL: https://arxiv.org/abs/2111.01689.

[51] H. W. Fowler. A dictionary of modern english. In Oxford University Press, 1926.

[52] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, Dec. 2021. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00437. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00437/1979261/tacl_a_00437.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00437.

[53] M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, and A. F. T. Martins. Results of WMT22 metrics shared task: stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics, Dec. 2022. URL: https://aclanthology.org/2022.wmt-1.2.

[54] S. Frenda, S. M. Lo, S. Casola, B. Scarlini, C. Marco, V. Basile, and D. Bernardi. Does anyone see the irony here? analysis of perspective-aware model predictions in irony detection. In *ECAI 2023 Workshop on Perspectivist Approaches to NLP*, 2023. URL: https://www.amazon.science/publications/does-anyone-see-the-irony-here-analysis-of-perspective-aware-model-predictions-in-irony-detection.

[55] A. Garg and M. Agarwal. Machine translation: A literature review. *CoRR*, abs/1901.01122, 2019. arXiv: 1901.01122. URL: http://arxiv.org/abs/1901.01122.

[56] B. Ghanem, J. Karoui, F. Benamara, P. Rosso, and V. Moriceau. Irony detection in a multilingual context, Apr. 2020. DOI: 10.1007/978-3-030-45442-5_18.

[57] A. Ghosh and T. Veale. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California. Association for Computational Linguistics, June 2016. DOI: 10.18653/v1/W16-0425. URL: https://aclanthology.org/W16-0425.

[58] A. Ghosh and T. Veale. IronyMagnet at SemEval-2018 task 3: a Siamese network for irony detection in social media. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 570–575, New Orleans, Louisiana. Association for Computational Linguistics, June 2018. DOI: 10.18653/v1/S18-1093. URL: https://aclanthology.org/S18-1093.

[59] D. Ghosh, A. R. Fabbri, and S. Muresan. Sarcasm analysis using conversation context. *Computational Linguistics*, 44:755–792, 2018.

[60] D. Ging and J. O'Higgins Norman. Cyberbullying, conflict management or just messing? teenage girls' understandings and experiences of gender, friendship, and conflict on facebook in an irish second-level school. *Feminist media studies*, 16(5):805–821, 2016.

[61] D. Grant, C. Hardy, C. Oswick, and L. L. Putnam. The sage handbook of organizational discourse. In SAGE knowledge, 2004.

[62] J. Gu, J. Bradbury, C. Xiong, V. O. K. Li, and R. Socher. Non-autoregressive neural machine translation. *CoRR*, abs/1711.02281, 2017. arXiv: 1711.02281. URL: http://arxiv.org/abs/1711.02281.

[63] L. Han. An overview on machine translation evaluation, 2022. DOI: 10.48550/ARXIV.2202.11027. URL: https://arxiv.org/abs/2202.11027.

[64] L. Han, G. Erofeev, I. Sorokina, S. Gladkoff, and G. Nenadic. Examining large pre-trained language models for machine translation: what you don't know about it. *ArXiv*, abs/2209.07417, 2022.

[65] M. Hanna and O. Bojar. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics, Nov. 2021. URL: https://aclanthology.org/2021.wmt-1.59.

[66] Q. He, G. Huang, Q. Cui, L. Li, and L. Liu. Fast and accurate neural machine translation with translation memory, Jan. 2021. DOI: 10.18653/v1/2021.acl-long.246.

[67] C. V. Hee, E. Lefever, and V. Hoste. Guidelines for annotating irony in social media text, version 2.0, 2016.

[68] S.-W. Huang, W.-Y. Chung, Y.-H. Wu, C.-C. Yu, and J.-L. Wu. A dimensional valence-arousal-irony dataset for Chinese sentence and context. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 147–154, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Nov. 2022. URL: https://aclanthology.org/2022.rocling-1.19.

[69] J. Hutchins and D. G. Hays. 11 alpac : the (in) famous report, 2015.

[70] J. Hutchins and E. Lovtskii. Petr petrovich troyanskii (1894-1950): a forgotten pioneer of mechanical translation. *Machine Translation*, 15(3):187–221, 2000. ISSN: 09226567, 15730573. URL: http://www.jstor.org/stable/40009018 (visited on 08/30/2022).

[71] A. L. Jakobsen. Investigating expert translators' processing knowledge. *Knowledge systems and translation*, 1732189, 2005.

[72] W. Jin, Y. Cheng, Y. Shen, W. Chen, and X. Ren. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *CoRR*, abs/2110.08484, 2021. arXiv: 2110.08484. URL: https://arxiv.org/abs/2110.08484.

[73] J. Jorgensen. The functions of sarcastic irony in speech. In *Journal of Pragmactics 26-5*. Elsevier, 1996.

[74] L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit. One model to learn them all. *CoRR*, abs/1706.05137, 2017. arXiv: 1706.05137. URL: http://arxiv.org/abs/1706.05137.

[75] M. Karpinska, N. Raj, K. Thai, Y. Song, A. Gupta, and M. Iyyer. Demetr: diagnosing evaluation metrics for translation, 2022. DOI: 10.48550/ARXIV.2210.13746. URL: https://arxiv.org/abs/2210.13746.

[76] P. Kavumba, R. Takahashi, and Y. Oda. Are prompt-based models clueless? *NLP*, 29(3):991–996, 2022. DOI: 10.5715/jnlp.29.991.

[77] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes. To ship or not to ship: an extensive evaluation of automatic metrics for machine translation, 2021. DOI: 10.48550/ARXIV.2107.10821. URL: https://arxiv.org/abs/2107.10821.

[78] P. Koehn and R. Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics, Aug. 2017. DOI: 10.18653/v1/W17-3204. URL: https://aclanthology.org/W17-3204.

[79] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133, 2003. URL: https://aclanthology.org/N03-1017.

[80] R. J. Kreuz and S. Glucksberg. How to be sarcstic: the echoic reminder theory of verbal irony. In *Journal of Experimental Psychology*. American Psychological Association, 1989.

[81] H. Lai, A. Toral, and M. Nissim. Multilingual pre-training with language and task adaptation for multilingual text style transfer. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 262–271, Dublin, Ireland. Association for Computational Linguistics, May 2022. DOI: 10.18653/v1/2022.acl-short.29. URL: https://aclanthology.org/2022.acl-short.29.

[82] G. Lample, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043, 2017. arXiv: 1711.00043. URL: http://arxiv.org/abs/1711.00043.

[83] H. Lang, M. N. Agrawal, Y. Kim, and D. Sontag. Co-training improves prompt-based learning for large language models. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11985–12003. PMLR, 17–23 Jul 2022. URL: https://proceedings.mlr.press/v162/lang22a.html.

[84] C. J. Lee and A. N. Katz. The differential role of ridicule in sarcasm and irony. In *Journal of Metaphor and Symbol*, 1998.

[85] S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, and H. Lim. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4), 2023. ISSN: 2227-7390. DOI: 10.3390/math11041006. URL: https://www.mdpi.com/2227-7390/11/4/1006.

[86] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. DOI: 10.48550/ARXIV.1910.13461. URL: https://arxiv.org/abs/1910.13461.

[87] A. Li, E. Chersoni, and R. Xiang. On the "easy" task of evaluating chinese irony detection, 2019.

[88] Y. Li, Y. Yin, J. Li, and Y. Zhang. Prompt-driven neural machine translation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2579–2590, Dublin, Ireland. Association for Computational Linguistics, May 2022. DOI: 10.18653/v1/2022.findings-acl.203. URL: https://aclanthology.org/2022.findings-acl.203.

[89] C.-Y. Lin. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics, July 2004. URL: https://aclanthology.org/W04-1013.

[90] X. Liu, K. Duh, L. Liu, and J. Gao. Very deep transformers for neural machine translation. *CoRR*, abs/2008.07772, 2020. arXiv: 2008.07772. URL: https://arxiv.org/abs/2008.07772.

[91] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210, 2020. arXiv: 2001.08210. URL: https://arxiv.org/abs/2001.08210.

[92] Y. Liu, Y. Wang, A. Sun, Z. Zhang, J. Guo, and X. Meng. A dual-channel framework for sarcasm recognition by detecting sentiment conflict, 2022.

[93] S. Lukin and M. Walker. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40, Atlanta, Georgia. Association for Computational Linguistics, June 2013. URL: https://aclanthology.org/W13-1104.

[94] M. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015. arXiv: 1508.04025. URL: http://arxiv.org/abs/1508.04025.

[95] D. Marcu and W. Wong. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 133–139, USA. Association for Computational Linguistics, 2002. DOI:

10.3115/1118693.1118711. URL: https://doi.org/10.3115/1118693.1118711.

[96] B. Marie. An automatic evaluation of the wmt22 general machine translation task. *ArXiv*, abs/2209.14172, 2022.

[97] B. Marie, A. Fujita, and R. Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. *CoRR*, abs/2106.15195, 2021. arXiv: 2106.15195. URL: https://arxiv.org/abs/2106.15195.

[98] N. Mathur, T. Baldwin, and T. Cohn. Tangled up in BLEU: reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics, July 2020. DOI: 10.18653/v1/2020.acl-main.448. URL: https://aclanthology.org/2020.acl-main.448.

[99] N. Mathur, J. Wei, M. Freitag, Q. Ma, and O. Bojar. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics, Nov. 2020. URL: https://aclanthology.org/2020.wmt-1.77.

[100] P. Mckenna. The rise of cyberbullying. *New Scientist*, 195(2613):26–27, 2007. ISSN: 0262-4079. DOI: https://doi.org/10.1016/S0262-4079(07)61835-1. URL: https://www.sciencedirect.com/science/article/pii/S0262407907618351.

[101] I. Merriam-Webster and K. Kuiper. *Merriam-Webster's Encyclopedia of Literature*. Academic OneFile. Merriam-Webster, 1995. ISBN: 9780877790426. URL: https://books.google.co.jp/books?id=ttSufPA-AwIC.

[102] J. Mirzakhalov, A. Babu, D. Ataman, S. Kariev, F. Tyers, O. Abduraufov, M. Hajili, S. Ivanova, A. Khaytbaev, A. Laverghetta Jr., B. Moydinboyev, E. Onal, S. Pulatova, A. Wahab, O. Firat, and S. Chellappan. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, Nov. 2021. DOI: 10.18653/v1/2021.emnlp-main.475. URL: https://aclanthology.org/2021.emnlp-main.475.

[103] S. Mohamed, A. Elsayed, Y. Hassan, and M. Abdou. Neural machine translation: past, present, and future. *Neural Computing and Applications*, 33:1–13, Dec. 2021. DOI: 10.1007/s00521-021-06268-0.

[104] M. Nagao. A framework of a mechanical translation between japanese and english by analogy principle. In *Proc. of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, Lyon, France. Elsevier North-Holland, Inc., 1984. ISBN: 0444865454.

[105] S. Nirenburg, V. Raskin, and A. Tucker. On knowledge-based machine translation. In *Proceedings of the 11th Coference on Computational Linguistics*, COLING '86, pages 627–632, Bonn, Germany. Association for Computational Linguistics, 1986. DOI: 10.3115/991365.991549. URL: https://doi.org/10.3115/991365.991549.

[106] A. Nowakowski. Approaching english-polish machine translation quality assessment with neural-based methods. *ArXiv*, abs/2209.11016, 2022.

[107] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. DOI: 10.1162/089120103321337421. URL: https://aclanthology.org/J03-1002.

[108] OpenAI. Chatgpt [computer software]. https://openai.com/, 2021.

[109] OpenAI. Gpt-4 technical report, 2023. arXiv: 2303.08774 [cs.CL].

[110] S. Oraby, V. Harrison, L. Reed, E. Hernandez, E. Riloff, and M. Walker. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics, Sept. 2016. DOI: 10.18653/v1/W16-3604. URL: https://aclanthology.org/W16-3604.

[111] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics, 2002. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135.

[112] L. Peled and R. Reichart. Sarcasm SIGN: interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1690–1700, Vancouver, Canada. Association for Computational Linguistics, July 2017. DOI: 10.18653/v1/P17-1155. URL: https://aclanthology.org/P17-1155.

[113] M. Popović. ChrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics, Sept. 2015. DOI: 10.18653/v1/W15-3049. URL: https://aclanthology.org/W15-3049.

[114] M. Post. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771, 2018. arXiv: 1804.08771. URL: http://arxiv.org/abs/1804.08771.

[115] M. Post, S. Ding, M. Martindale, and W. Wu. An exploration of placeholding in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 182–192, Dublin, Ireland. European Association for Machine Translation, Aug. 2019. URL: https://aclanthology.org/W19-6618.

[116] R. A. Potamias, G. Siolas, and A. Stafylopatis. A transformer-based approach to irony and sarcasm detection. *CoRR*, abs/1911.10401, 2019. arXiv: 1911.10401. URL: http://arxiv.org/abs/1911.10401.

[117] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. In the service of online order: tackling cyber-bullying with machine learning and affect analysis. *International Journal of Computational Linguistics Research*, 1(3):135–154, 2010.

[118] M. Ptaszynski, J. K. K. Eronen, and F. Masui. Learning deep on cyberbullying is always better than brute force. In *LaCATODA@ IJCAI*, pages 3–10, 2017.

[119] M. Ptaszyński, G. Leliwa, M. Piech, and A. Smywiński-Pohl. Cyberbullying detection–technical report 2/2018, department of computer science agh, university of science and technology. *arXiv preprint arXiv:1808.00926*, 2018.

[120] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. arXiv: 1910.10683. URL: http://arxiv.org/abs/1910.10683.

[121] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. *CoRR*, abs/2009.09025, 2020. arXiv: 2009.09025. URL: https://arxiv.org/abs/2009.09025.

[122] E. Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, Sept. 2018. DOI: 10.1162/coli_a_00322. URL: https://aclanthology.org/J18-3002.

[123] A. Reyes, P. Rosso, and D. Buscaldi. From humor recognition to irony detection: the figurative language of social media. *Data Knowledge Engineering*, 74:1–12, 2012. ISSN: 0169-023X. DOI: https://doi.org/10.1016/j.datak.2012.02.005. URL: https://www.sciencedirect.com/science/article/pii/S0169023X12000237. Applications of Natural Language to Information Systems.

[124] E. Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, pages 1044–1049, Portland, Oregon. AAAI Press, 1996. ISBN: 026251091X.

[125] M. A. Ruiz Moneva. Searching for some relevance answers to the problems raised by the translation of irony. *REVISTA ALICANTINA DE ESTUDIOS INGLESES*, 14:213–247, Jan. 2001. DOI: 10.14198/raei.2001.14.14.

[126] M. Salameh, S. Mohammad, and S. Kiritchenko. Sentiment after translation: a case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado. Association for Computational Linguistics, May 2015. DOI: 10.3115/v1/N15-1078. URL: https://aclanthology.org/N15-1078.

[127] D. Saunders, F. Stahlberg, and B. Byrne. Using context in neural machine translation training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online. Association for Computational Linguistics, July 2020. DOI: 10.18653/v1/2020.acl-main.693. URL: https://aclanthology.org/2020.acl-main.693.

[128] E. Savini and C. Caragea. Intermediate-task transfer learning with bert for sarcasm detection. *Mathematics*, 2022.

[129] H. Schwenk. Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of COLING 2012: Posters*, pages 1071–1080, Mumbai, India. The COLING 2012 Organizing Committee, Dec. 2012. URL: https://aclanthology.org/C12-2104.

[130] J. R. Searle. Literal meaning, 1978.

[131] T. Sellam, D. Das, and A. P. Parikh. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696, 2020. arXiv: 2004.04696. URL: https://arxiv.org/abs/2004.04696.

[132] C. Shelley. The bicoherence theory of situational irony. In *Cognitive Science 25*. Elsevier, 2001.

[133] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation, Jan. 2006.

[134] P. Stanchev, W. Wang, and H. Ney. Towards a better evaluation of metrics for machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 928–933, Online. Association for Computational Linguistics, Nov. 2020. URL: https://aclanthology.org/2020.wmt-1.103.

[135] E. Sulis, D. I. H. Fariaz, P. Rosso, and V. Patti. Figurative messages and affect in twitter: differences between #irony, #sarcasm and #not. In *Knowledge-Based Systems*. Elsevier, 2016.

[136] I. Sutskever, J. Martens, and G. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1017–1024, Bellevue, Washington, USA. Omnipress, 2011. ISBN: 9781450306195.

[137] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Montreal, Canada. MIT Press, 2014.

[138] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. Multilingual translation with extensible multilingual pretraining and finetuning, 2020. DOI: 10.48550/ARXIV.2008.00401. URL: https://arxiv.org/abs/2008.00401.

[139] Y. Tay, A. T. Luu, S. C. Hui, and J. Su. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics, July 2018. DOI: 10.18653/v1/P18-1093. URL: https://aclanthology.org/P18-1093.

[140] B. Thompson and M. Post. Automatic machine translation evaluation in many languages via zero-shot paraphrasing, Online, Nov. 2020.

[141] B. Thouin. The meteo system, 1982.

[142] D. P. P. Toma. Systran as a multilingual machine translation system, 1977.

[143] D. Tomás, R. Bueno, G. Zhang, P. Rosso, and R. Schifanella. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, 14:1–12, Oct. 2022. DOI: 10.1007/s12652-022-04447-y.

[144] O. Tsur, D. Davidov, and A. Rappoport. Icwsm – a great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2010.

[145] C. Turban and U. Kruschwitz. Tackling irony detection using ensemble classifiers. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources*

and Evaluation Conference, pages 6976–6984, Marseille, France. European Language Resources Association, June 2022. URL: https://aclanthology.org/2022.lrec-1.754.

[146] C. Van Hee, E. Lefever, and V. Hoste. Monday mornings are my fave :) #not exploring the automatic recognition of irony in English tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2730–2739, Osaka, Japan. The COLING 2016 Organizing Committee, Dec. 2016. URL: https://aclanthology.org/C16-1257.

[147] C. Van Hee, E. Lefever, and V. Hoste. SemEval-2018 task 3: irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics, June 2018. DOI: 10.18653/v1/S18-1005. URL: https://aclanthology.org/S18-1005.

[148] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[149] T. Vu, D. Q. Nguyen, X.-S. Vu, D. Q. Nguyen, M. Catt, and M. Trenell. NIHRIO at SemEval-2018 task 3: a simple and accurate neural network model for irony detection in Twitter. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 525–530, New Orleans, Louisiana. Association for Computational Linguistics, June 2018. DOI: 10.18653/v1/S18-1085. URL: https://aclanthology.org/S18-1085.

[150] M. Walker, J. F. Tree, P. Anand, R. Abbott, and J. King. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA), May 2012. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/1078_Paper.pdf.

[151] H. Wang, H. Wu, Z. He, L. Huang, and K. Ward Church. Progress in machine translation. *Engineering*, 2021. ISSN: 2095-8099. DOI: https://doi.org/

10.1016/j.eng.2021.03.023. URL: https://www.sciencedirect.com/science/article/pii/S2095809921002745.

[152] W. Wang, J.-T. Peter, H. Rosendahl, and H. Ney. CharacTer: translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics, Aug. 2016. DOI: 10.18653/v1/W16-2342. URL: https://aclanthology.org/W16-2342.

[153] W. Weaver. Translation. In W. N. Locke and A. D. Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA, 1949/1955. Reprinted from a memorandum written by Weaver in 1949.

[154] A. Webson and E. Pavlick. Do prompt-based models really understand the meaning of their prompts? *CoRR*, abs/2109.01247, 2021. arXiv: 2109.01247. URL: https://arxiv.org/abs/2109.01247.

[155] B. Wei, M. Wang, H. Zhou, J. Lin, and X. Sun. Imitation learning for non-autoregressive neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1304–1312, Florence, Italy. Association for Computational Linguistics, July 2019. DOI: 10.18653/v1/P19-1125. URL: https://aclanthology.org/P19-1125.

[156] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. Ccnet: extracting high quality monolingual datasets from web crawl data, 2019. DOI: 10.48550/ARXIV.1911.00359. URL: https://arxiv.org/abs/1911.00359.

[157] J. S. White, T. A. O'Connell, and F. E. O'Mara. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches, Columbia, Maryland, USA, Oct. 1994. URL: https://aclanthology.org/1994.amta-1.25.

[158] C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, and Y. Huang. THU_NGN at SemEval-2018 task 3: tweet irony detection with densely connected LSTM and multi-task learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56, New Orleans, Louisiana. Association for Computational Linguistics, June 2018. DOI: 10.18653/v1/S18-1006. URL: https://aclanthology.org/S18-1006.

[159] X. Wu, Y. Xia, J. Zhu, L. Wu, S. Xie, and T. Qin. A study of bert for context-aware neural machine translation. *Mach. Learn.*, 111(3):917–935, Mar. 2022. ISSN: 0885-6125. DOI: 10.1007/s10994-021-06070-y. URL: https://doi.org/10.1007/s10994-021-06070-y.

[160] Y. Wu and G. Hu. Exploring prompt engineering with GPT language models for document-level machine translation: insights and findings. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169, Singapore. Association for Computational Linguistics, Dec. 2023. DOI: 10.18653/v1/2023.wmt-1.15. URL: https://aclanthology.org/2023.wmt-1.15.

[161] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.

[162] R. Xiang, X. Gao, Y. Long, A. Li, E. Chersoni, Q. Lu, and C.-R. Huang. Ciron: a new benchmark dataset for Chinese irony detection. English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5714–5720, Marseille, France. European Language Resources Association, May 2020. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.701.

[163] H. Xu, Q. Liu, J. van Genabith, and J. Zhang. Why deep transformers are difficult to converge? from computation order to lipschitz restricted parameter initialization. *CoRR*, abs/1911.03179, 2019. arXiv: 1911.03179. URL: http://arxiv.org/abs/1911.03179.

[164] L. Xu, Y. Chen, G. Cui, H. Gao, and Z. Liu. Exploring the universal vulnerability of prompt-based learning paradigm, 2022. arXiv: 2204.05239 [cs.CL].

[165] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. Byt5: towards a token-free future with pre-trained byte-to-byte models. *CoRR*, abs/2105.13626, 2021. arXiv: 2105.13626. URL: https://arxiv.org/abs/2105.13626.

[166] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. Mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934, 2020. arXiv: 2010.11934. URL: https://arxiv.org/abs/2010.11934.

[167] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019. arXiv: 1906.08237. URL: http://arxiv.org/abs/1906.08237.

[168] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, P. Zhang, Y. Dong, and J. Tang. Glm-130b: an open bilingual pre-trained model, 2023. arXiv: 2210.02414 [cs.CL].

[169] B. Zhang, B. Haddow, and A. Birch. Prompting large language model for machine translation: a case study, 2023. arXiv: 2301.07069 [cs.CL].

[170] J. Zhang and C. Zong. Neural machine translation: challenges, progress and future. *CoRR*, abs/2004.05809, 2020. arXiv: 2004.05809. URL: https://arxiv.org/abs/2004.05809.

[171] S. Zhang, X. Zhang, J. Chan, and P. Rosso. Irony detection via sentiment-based transfer learning. *Information Processing Management*, 56(5):1633–1644, 2019. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2019.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0306457318307428.

[172] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: evaluating text generation with bert, 2019. DOI: 10.48550/ARXIV.1904.09675. URL: https://arxiv.org/abs/1904.09675.