University of Bonn
CAISA Lab

Introduction to Natural Language Processing

(Winter Semester 2023/2024)

Default Project

Project Proposal of Team #1

Formality Style Transformation

Group Members:

Duc Manh Vu	-	50137408
Suraj Giri	-	50190564
Muslimbek Abduvaliev	-	50136555
Nijat Sadikhov	-	50186266
Akmalkhon Khashimov	-	50178353

November 28, 2023

1. Introduction and Motivation

Formal Writing is a structured approach that uses linguistic conventions. The main aspects of this approach are precise vocabulary selection, proper grammar and punctuation, and objective tone. Pieces of formal writing are suitable for delivering truths, accurate information, academic knowledge, and logical argumentation to audiences whose level is unknown, or to people to whom the writers want to show their reverence. Therefore, they can be found in scholarly articles, scientific literature, business documentation, research publications, legislative texts, and even in communication between intellectuals. The application of formal language helps individuals yield advantageous outcomes and opportunities. Formal writing is ubiquitous and advantageous and should be a skill every individual should acquire. However, learning it requires proper guidance from tutors with profound knowledge and pedagogical skills. Besides, these tutors must spend much time checking students' work and providing thorough feedback. Therefore, human guidance is laborious and time-intensive. Various supportive programs have been developed to assist students in self-elevating their writing skills. Not only can the mentioned issues be resolved, but this initiative also allows students to schedule their study time and pace proactively.

Automatic Formality Style Transfer is one of the features those supportive programs should possess. It refers to the automated task of paraphrasing text from an informal style to its formal counterpart while preserving its meaning. This task is an extension of the Text Style Transfer task. Initially, the research progress was hindered as there was no large-scale dataset. However, the advent of the Grammarly Yahoo Answers Corpus Dataset (GYAFC) has promoted this research area. The authors of this dataset also introduced three main approaches to solve this task: the Rule-based system [1], the Phrase-based Machine Translation approach [1], and the Neural Machine Translation approach [1]. They also employed Human judgments and automatic metrics to assess the performance of these methods on different criteria: Formality, Fluency, Meaning Preservation, and Overall Ranking. The appearance of this dataset has raised interest in this field, and several research directions have emerged: Data Augmentation [2], Explainability, Multilingual Approach [3], Metrics [4], and Modern Deep Architecture [5].

In this report, we will present some general information about our project on Formality Style Transfer. The next Section will demonstrate some detail about the GYAFC dataset. Then we will establish the goals in Section 3. and contrive our approach in Section 4.. The expectation from the outcomes of this work will be illustrated in Section 5.. Later in Section 6., we will introduce the metrics used to evaluate these outcomes. Section 7. will mention potential obstacles we may encounter during our research and implementation time. Finally in Section 8., we will depict how we will distribute the total workload to every member of the group and visualize the timeline of this project.

2. Available Dataset and Primitive Data Exploration

Grammarly's Yahoo Answers Corpus Dataset (GYAFC) [1] is the only large-scale public dataset available in this field. The researchers derived this dataset from the Yahoo Answers corpus (L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0[1].) According to the original publication, they collected nearly 53000 informal sentences in each subject: Entertainment &

[1] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11

Music (E&M) and Family & Relationships (F&R), together creating a train set. Workers from Amazon Mechanical Turk were hired to create corresponding formal sentences, and experts were responsible for the quality of this production. They also sampled approximately 3000 sentences as the validation set and 1500 as the test set in each topic, but only the best-performed workers were accredited to annotate these sets. For each sentence in these sets, the workers generated four separate corresponding transformations.

Regarding the manifestations of informality observed in this dataset, the authors have also classified them as follows: Word Usage, Contractions, Misspelling, Capitalization, Incompleteness, Fillers, Repetitions, Lowercase, and Abbreviation.

After acquiring the dataset, we conducted a thorough examination of the dataset and made specific observations regarding the formal text. We realized that various details arouse skepticism about the formality mentioned by the authors. We will demonstrate several examples below.

- Use of contractions: "don't," "It's," "Who's"...
- Improper capitalization: "KAVUNDAMANI SENTHIL VADIVELU VIVEK N.S.KRISHNAN CHANDRA BABU CHARLIE CHAPLIN" or "Kavundamani Senthil, Vadivelu Vivek, N.S Krishnan, Chandra Babu, Charlie Chaplin"...
- Inappropriate symbols or sequences of symbols: ":-)", ":)"...
- Informal use of punctuations: "!!!!!", "....."...

Consequently, these phases pose a severe obstacle in our data preprocessing phase.

3. Objectives and Scope

We aim to experiment with the Formality Style Transfer task within this project's scope using the GYAFC dataset. After reviewing the literature and exploring the dataset, we have posed several questions that the results of the experiments will manifest their answers.

- How good is the GYAFC dataset in quantity and quality?
- How can we preprocess the data to create robust features? Specifically, what is the optimal way to tokenize informal and formal texts?
- Will Deep Learning methods output remarkable performance?

To measure the outcomes of our work and accomplish the mentioned goals, we need to outline our approach stepwise as follows:

- Exploratory Data Analysis: We aim to understand the attributes of the given dataset thoroughly.
- Data Preprocessing: We aim to find a method to preprocess and tokenize inputs and outputs and map the text representation to their corresponding numerical representation.
- Performance Comparison: For this step, we aim to apply different modern architectures for Machine Translation, which we will discuss in the upcoming sections. We will then assess the performance of every model by quantitative metrics.

4. Methodology

We will approach this problem entirely as an extension of the Neural Machine Translation (NMT) task, which is a sequence-to-sequence (seq2seq) mapping from the source to the target language [6]. One justification for our resolution is that both tasks share the goal of transforming a source sequence of tokens in one domain (language or style) to a target sequence of tokens in another domain without significantly altering the meaning of the original sequence and the length of the input and output sequences can vary. Furthermore, tokenization, word embedding, backbones, and result postprocessing are generally similar between these two tasks; therefore, we can utilize methods that have achieved remarkable outcomes in the NMT task.

Preprocessing and Word Embedding:
First, we will preprocess and tokenize the texts. We consider two methods to preprocess the informal texts:

- Genuinely tokenizing the texts without any intervention.
- Using the Rule-based Model mentioned in the dataset paper [1].

After transforming the texts into token sequences, we aim to use a pre-trained word embedding method to map tokens to numerical values. We can employ several available techniques: Google Word2Vec [7], Stanford GloVe [8], FastText [9], and others. Our ablation study will attempt to trial every configuration to explore the best technique.

Backbone Architecture:
Modern approaches to the NMT task have employed encoder-decoder deep architectures [6] and achieved remarkable accomplishments. The encoder processes the input sequences and compresses the information into intermediate representations that capture the crucial features of the input. The decoder then takes these values to generate the output sequences. Therefore, it would be advantageous for us to exert the abilities of these models to conduct our task. Let us depict brief information about the architectures that we will experiment with:

- **Recurrent Neural Networks (RNN)** [10, 11]: RNNs are designed to handle data sequences. For NMT, they process words in the input sentence sequentially, maintaining a 'memory' of previous inputs. However, RNNs struggle with long sequences due to issues like vanishing gradients.
- **Long Short-Term Memory (LSTM)** [12]: An advanced type of RNN, LSTMs are better at capturing long-range dependencies in text, making them more effective for complex translation tasks. They mitigate the vanishing gradient problem, allowing them to remember information over longer sequences.
- **Transformers** [13] : This architecture, introduced in the paper "Attention Is All You Need" by Vaswani et al., moves away from sequential processing and relies on a mechanism called 'attention' to weigh the influence of different parts of the input sentence. This allows for parallel processing and capturing complex relationships in the data, making Transformers very efficient and powerful for NMT tasks.
- **BERT2BERT:** This is a variation where both the encoder and decoder are based on the BERT (Bidirectional Encoder Representations from Transformers [14]) model. BERT is pre-trained on a large text corpus, making it adept at understanding language context. Using

BERT as both the encoder and decoder can enhance the model's ability to understand and generate text effectively, making it an intriguing choice for formality style transfer.

By experimenting with these architectures, we aim to measure their performance using the metrics mentioned in Section 6 and identify which would give desirable outcomes in this style transfer task.

5. Expected Results

By accomplishing this project, we can expect the development of an NLP model adept at converting informal text to formal style, effectively changing grammatical structures and vocabulary while preserving the original content and context. The project will offer insights into the linguistic features that differentiate formal and informal language. Additionally, it will provide quantitative evaluations of the model's performance, such as the accuracy and naturalness of the transformed text.

6. Evaluation Metrics

Since our team does not possess sufficient capabilities to evaluate the formality of the output sentence manually, we intend to employ only the automatic metrics mentioned in [the paper]. The metrics are categorized according to different criteria. We will attempt to give succinct descriptions of these criteria.

- **Formality** [15]: This metric measures how effectively the model can convert informal text into formal text. It measures the model's ability to identify and correctly apply the rules and nuances of formal language to the given informal text. This could be quantified by evaluating a set of transformed texts against standard criteria for formality and calculating the percentage of texts that meet these criteria. The authors in [the paper] used a statistical model to evaluate this criterion automatically.
- **Fluency** [16]: This metric refers to grammaticality, which ensures the structure and arrangement of words in the output of our Machine Translations. It involves maintaining the proper syntax, grammatical rules, and sentence structure while converting the text from an informal to a formal tone. A statistical model from [paper] can be employed to assess this aspect.
- **Meaning Preservation** [17]: This is critical in style transfer, as the goal is to change the style of the text without altering its underlying meaning or content. Meaning preservation can be assessed by comparing the semantic content of the original informal text with that of the transformed formal text. This comparison could be made using similarity measures that evaluate how well the critical information and meaning are retained after the transformation. Besides, He et al. introduced an automatic similarity measurement using a Convolution Neural Network.

Apart from the above criteria, we propose using some metrics previously used in Natural Language Style Transfer as below:

- **Bilingual Evaluation Understudy (BLEU)** [18]: BLEU is a metric commonly used to evaluate the quality of machine-generated translations by comparing them to one or more

reference translations provided to our model. It measures the overlap of n-grams between the candidate translation and the references, finally providing a score indicating the similarity between the reference and the translation.

- **Translation Edit Rate Plus (TERp)** [19]: TERp is the evaluation metric for the machine translation task, which measures the matching flaw between machine-generated translations and human-created translation. Calculation of this metric entails all Translation Edit Rate (TER) operations with three additional operations measuring correlations among stems, synonyms, and phases.
- **Paraphrase In N-gram Changes (PINC)** [20]: It provides a quantitative measure of how extensively the paraphrased text diverges from its source by computing the percentage of n-grams that appear in the candidate sentence but not in the source sentence. It provides clear insights into the level of linguistic transformation and the degree of paraphrasing performed.

7. Challenges and Limitations

We have been aware of the potential difficulties that our approach would provoke. We will demonstrate them and give their corresponding initiatives to ameliorate them.

- **Text processing**: This is a crucial task we must carefully conduct as the quality of the outcomes will be impaired if we process the texts improperly. Unlike standardized language, informal texts are not universal, and they can appear in various forms, which would pose a dilemma over the methods to process them. We should refer ourselves to the method in [1] or advice from the tutor. Regarding the formal references, we should also promote a processing method to alleviate the issues mentioned in Section 2..
- **Model Choices**: Various advanced architectures suitable for the text generation task have emerged recently. However, trialing all of them with different configurations are time-consuming. Besides, we have deliberated between fine-tuning the pre-trained weights or training models from scratch. The pre-trained weights have captured the knowledge in specific domains, so reusing these configurations would be beneficial in relevant downstream tasks [21]. For this reason, we will intend to employ several remarkable pre-trained language models and fine-tune them with the given dataset.
- **Computational Resources**: Training or fine-tuning deep architectures would require devices with sufficient computational capabilities. We suggest experimenting on the Google Colaboratory platform.

8. Task Assignment and Timeline

In our project approach, we allocate specific tasks to ensure a comprehensive and efficient workflow:

- **Data preparation (1-2 weeks):** We are responsible for carefully selecting and preprocessing the dataset. This involves cleaning the text, standardizing formats, and ensuring a balanced representation of various informal expressions for more accurate model training.
- **Model training (3-4 weeks):** Our team develops and trains the NLP models. This includes selecting appropriate algorithms, setting up training protocols, and iteratively refining the

model based on initial results.

- **Testing (1-2 weeks):** We conduct testing of the model using a portion of the dataset reserved for this purpose. The aim here is to evaluate the model's ability to accurately transform informal text into a formal style while maintaining the original content.
- **Validation (1 week):** Finally, we validate the model's performance against unseen data. This step is crucial to assess the model's generalizability and effectiveness in real-world scenarios.

Our scheduled plan is depicted in Figure 1.

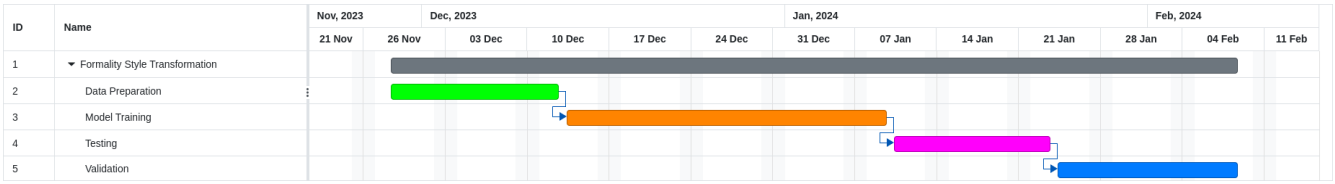

Figure 1: Project Timeline

Bibliography

[1] Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[2] Yi Zhang, Tao Ge, and Xu Sun. Parallel data augmentation for formality style transfer. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online, July 2020. Association for Computational Linguistics.

[3] Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online, June 2021. Association for Computational Linguistics.

[4] Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Huiyuan Lai, Antonio Toral, and Malvina Nissim. Thank you BART! rewarding pre-trained models improves formality style transfer. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online, August 2021. Association for Computational Linguistics.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Neural and Information Processing System (NIPS)*, 2013.

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[9] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[10] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[11] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics.

[15] Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74, 2016.

[16] Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, 2014.

[17] Hua He, Kevin Gimpel, and Jimmy Lin. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1576–1586, 2015.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[19] Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the*

Fourth Workshop on Statistical Machine Translation, StatMT '09, page 259–268, USA, 2009. Association for Computational Linguistics.

[20] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[21] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. Pre-trained models: Past, present and future. *CoRR*, abs/2106.07139, 2021.