Style Transfer in Natural Language Processing and Machine Translation

Aiswarya Mangalanandan Sobhana, Alex Farrell-Webb, Alice Varley, Artemis Forshaw, Digvijay Ashok Hajare, Kevin John Mathew, Rahul Arun Nawale

Abstract

Text Style Transfer (TST) falls under the domain of Natural Language Processing (NLP) and can be applied to multilingual contexts. This paper reviews these areas with respect to translation, and has a focus on the application of TST concepts in multilingual contexts. It is much harder to apply machine translation techniques to low-resource languages and is not yet ready for unsupervised use. In the future, there is scope to move away from English-centric models and further develop and research multilingual style transfer specifically.

Key words: *Natural Language Processing, Machine Translation, Natural Language Understanding, Natural Language Understanding, Text Style Transfer*

Introduction

Linguistic researchers all over the world have been fascinated by the stylistic quotients and properties that text possesses. Nils Erik Enkvist remarked in his book, *Style and Text*, that text style is a 'concept that is as common as it is elusive' and suggested that style may be described as a variation of linguistics while maintaining the actual meaning to the text (Hu et al, 2022, p. 1). To give a more practical example, the 'formality' of sentences will change based on the context in which they are used; examples include a conversation between colleagues such as 'Let's all meet up this Saturday evening!', or a professional email such as 'We will arrange a meeting this Saturday evening'.

Many academics and researchers in the field of computer science have recently become interested in studying text style and the transfer of style between languages. One of the main topics that are under investigation is Text Style Transfer (TST), which is an increasingly popular branch of Natural Language Generation (NLG) and aims to control certain attributes such as formality, politeness etc. that are present in the generated text while preserving the meaning. The majority of TST research so far uses freely accessible parallel corpora to accomplish text style transfer tasks (Xu et al, 2019, p. 1). The parallel corpora for TST consists of sentences equivalent in meaning with different styles which can be in different languages but with the same semantics. However, parallel data sets are scarce in many real-world TST applications, such as generation of dialogues or subtitles in different languages. Because of this, a completely new set of TST algorithms have been developed (Hu et al, 2022, p. 1).

This paper aims to review the literature on the advances and current research surrounding TST, as well as exploring its applications in Machine Translation. We begin our discussion with an overview on Natural Language Processing (NLP) and Natural Language Generation (NLG). In **Section 1**, we give a brief introduction on Machine Translation along with its history, related research areas, limitations and its proposed solutions. **Section 2** provides the preliminary information about TST, its history and an overview about the plethora of research that is happening in the domain. This section also includes the explanation of the TST process performed with conversational texts as input data. Finally, in **Section 3** we conclude that TST is at the forefront of research concerning Machine Translation, although it is still affected by limitations that are common throughout the field.

Section 1 - Machine Translation

Machine translation (MT) is a type of NLP that translates text from one language to another without the need for a human translator. Therefore, to understand MT, it is necessary to first explore some concepts in NLP.

1.1 Natural Language Processing (NLP)

1.1.1 Overview of Natural Language Processing

Natural Language Processing (NLP) is a domain that first came into the limelight when Artificial Intelligence (AI) technology started gaining popularity, which in itself is a vast domain. This discipline is focused on the creation of computational models which enable the computers to understand how humans speak and communicate. The use of NLP models enables the computer to comprehend the semantic associations between pairs of words (for instance, the semantic associations between the terms 'shirt and horse' and 'shirt and dog' are relatively similar) (Lake and Murphy, 2021, p. 3). This discipline is huge and includes other tasks such as converting text to speech, translating natural languages and so on.

Natural Language Processing combines a lot of domains such as computational linguistics, deep learning, machine learning and rule-based modelling of human language with some statistics to enable computers to understand and comprehend the human language, which can be in any form such as voice data or plain text. NLP is a technique which has been around for the past half a century but is now gaining much more importance due to the availability of powerful computers all around the globe (Brownlee, 2017). NLP can be further subdivided into Natural Language Generation (NLG) and Natural Language Understanding (NLU).

1.1.2 Natural Language Understanding

Natural language understanding (NLU) is a subset of NLP which focuses on syntactic and semantic analysis to understand the meaning behind a sentence. NLU also provides the structure of data, which specifies how words and phrases are connected to each other. While humans do this naturally, the combination of these is required by a machine to understand and comprehend the intended meaning of various kinds of texts in a sentence. For this purpose, various rules, techniques, and models are used to find the objective behind that text. Such processes are also commonly used in data mining to understand consumer attitudes.

1.1.3 Natural Language Generation

While NLU focuses on computer understanding and comprehending inputted information, NLG enables the computers to output text. Based on some kind of input data, NLG generates text which sounds human. This text can be further converted to speech using other text-to-speech algorithms or services. NLG can also summarise text from inputted documents whilst ensuring the integrity of the information remains intact (Yadav et al, 2022, pp. 4-5).

During its early stages, NLG was used to generate text using pre-set templates. However, recent and ground-breaking advancements such as recurrent neural networks have helped in creating text which can be processed in real time (Wen and Young, 2020, p. 2). Considering how NLU works, NLG applications need to consider language rules based on morphology, the lexicons in the content, the syntax and semantics on how to phrase a sentence properly.

NLG forms a large part of the process of MT. It works in the most straightforward manner when the input and output languages are clearly specified in advance of the operation. Garje and Kharate (2013) via Phadke and Devane (2017, p. 881) note that most works which are focused on bilingual models '[translate] text from one natural language to another in a unidirectional way'. It follows logically that languages with a larger corpus of text, in other words a larger database of translations to draw from, will result in higher quality translations.

1.1.4 Language guessing

For multilingual translation, the language identification (or language guessing) aspect of NLU is also extremely important. The first task of the MT software is to identify its source language, which is what differentiates multilingual translation from single source translation.

When it comes to source texts in more than one language, Babhulgaonkar and Sonavane (2020, p. 401) note that a translation model trained on a simple source-language to target-language data pairs would be insufficient, as only one of the languages in the source text would be processed. Any other languages would be discarded by the model as 'noise', degrading the quality of the translation overall. To combat this, a multilingual MT software will essentially contain several different MT 'engines' that function separately according to their source-target language pair, and it is the job of the language guessing functions to correctly identify the language being translated and assign it to its relevant engine.

1.2 Machine Translation

1.2.1 History of Machine Translation

While earlier versions of automatic translation had been floated around, the first documented idea of using computers to automatically translate text are noted by Hutchins (2010, p.1) to have come from Andrew Booth and Warren Weaver in 1947. Computers as a tool were brand new and already their potential use in the field of translation was being considered. However, most of this work remained purely theoretical until 1952 when the first Machine Translation conference was held at MIT by Yehoshua Bar-Hillel (Hutchins, 2010, p. 2), and then later, in 1954, when he and IBM demonstrated the first MT system. Hutchins described the system as a 'carefully selected sample of 49 Russian sentences [that] was translated into English, using a very restricted vocabulary of 250 words and just 6 grammar rules.' (2010, p. 2). This may seem rudimentary now, but at the time was revolutionary, and a major driver of interest in the further development of research in this area.

Research continued at a regular pace, utilising every technological advancement of the era to improve. However, it was the impetus of the Cold War between the USA and the Soviet Union that led to an important form of MT that is still the fundamental form used today: Statistical Machine Translation (SMT). Described as a 'corpus based approach' (Hutchins, 2010, p. 12), this is a system which breaks down the source text into sequences of words, dubbed 'phrases' upon which probabilistic functions are performed to determine the most appropriate translation. Hutchins summarised the essential format of this method of generating a 'translation model' (2010, p. 11) as follows:

> "first to align phrases, word groups and individual words of the parallel texts, and then to calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in the other language" (2010, p. 11).

1.2.2 Statistical Machine Translation

Statistical Machine Translation performs best when provided with a large corpus of text to reference when making decisions about the best possible translation for any given phrase, or set of words. Consequently, the system works best when used in languages for which there exists already a large bilingual corpus (known as a language model).

However, not all languages around the world have such a large corpus to pull from. Such languages are described as low-resource languages. As a result, recent research has begun to head towards MT algorithms based on a monolingual corpus instead. Unsupervised Neural Machine Translation is one such area of research.

1.3 Neural Machine Translation

1.3.1 Overview

Neural Machine Translation (NMT) works by processing language through a single neural network (Stahlberg, 2020, p. 343), treating the input text as continuous data rather than discrete parcels of data (Tan et al, 2020, p. 5). NMT models have the capacity to process the context of words, considering syntax, semantics and morphology during its selection of output, and is particularly useful for sentiment analysis, which is key to style transfer (Stahlberg, 2020, p. 343). As a data-driven approach, NMT uses probabilities to handle inputs and outputs, selecting what it calculates to be the best results (Tan et al, 2020, p. 6).

There have been papers exploring the use of neural networks since the early 2010s, initially as part of larger SMT models (Stahlberg, 2020, p. 343). NMT is relatively recent compared to other MT forms; it has only begun to gain traction since around 2016 when Google unveiled their own GNMT (Google Neural Machine Translation) software (Phadke, 2017, pp. 882-883). NMT has been found to be more effective than SMT, and so has been adopted as the default in online industries (Ding et al, 2017, p. 1150)

1.3.2 Limitations of Neural Machine Translation

However, while unsupervised NMT is where most cutting edge research is being conducted, it has not yet overtaken previous forms of translation in overall accuracy. In fact, as Kim, Graca and Ney (2020, pp. 38-39) argue, unsupervised NMT often performs worse than both semi-supervised and supervised NMT. Because of this Kim argues that UNMT can not be recommended for NMT models.

Even NMT more generally has its flaws and limitations. Any data within an input classified as ‘noise’, whether natural or artificial, can significantly impact the performance of NMT systems (Tan et al, 2020, p.17). To compound the issue, NMT systems are also complex in their internal workings, processing language data through vectors and matrices, leading them to be difficult for humans to parse. This makes the debugging of NMT systems particularly challenging, and research to address this has been in progress for the past decade (Ding et al, 2017, p. 1150).

The mere running of NMT systems can also be challenging. GPU hardware can handle NMT at an acceptably high speed, but GPU costs considerably more than CPU. Current research is looking into how to increase the efficiency of NMT systems to reduce the reliance on expensive GPU hardware (Stalhberg, 2020, p. 365). There is still much research to be done in this area to address these hardware constraints.

Section 2 - Style Transfer

2.1 Overview

The difficult task of changing some elements of a given text, such as those of mood, sentiment, tense, voice, politeness, etc., as summarised by Lai et al (2019, p. 3579), while maintaining its meaning, other features, and contents is known as text style transfer. This task can be used for a variety of tasks, including paraphrasing, summarising information, changing poetry or song lyrics, concealing the author's identity, and scenario-adaptive text analysis.

Although 'informal' data is widely available across the internet, existing NLP tasks and models are unable to effectively use it and function well for such data because of informal speech as well as grammatical, spelling, and semantic mistakes. Because of this, formality style transfer, a particular style transfer job which seeks to maintain the substance of an informal statement while making it semantically and grammatically proper, has recently drawn more and more attention (Chawla and Yang, 2020, p. 1). See fig.1 for an example of a formality style transfer.

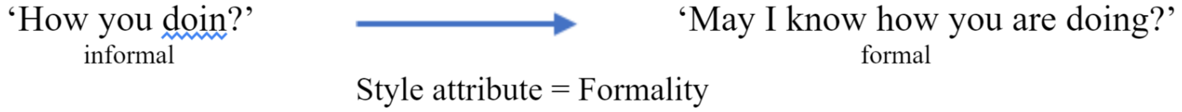


Figure 1: An example of formality style transfer

2.2 History of Style Transfer

Text style transfer (TST) already has an established place in the field of NLP, and has recently gained substantial interest following promising results offered by deep neural models used in TST studies.

Due to the increasing demand for TST, ongoing research in this sector has evolved, ranging from conventional language techniques to much more modern neural network-based techniques. Conventional methods focus on term substitution and templates. For instance, early research in NLG for weather forecasting resulted in the creation of domain-specific templates to characterise distinct weather types with differing degrees of uncertainty for various operators (Jin et al, 2022, p. 157).

To create text with pragmatic restrictions such as formality under a small-scale, well-defined schema, language-based systems and schema-based NLG systems are the starting points for research that more explicitly focuses on TST (Jin et al, 2022, p. 157). For the most part, this previous effort needed domain-specific templates, hand-featured phrase sets that communicate a certain characteristic (for example, being talkative), and occasionally a look-up table of phrases with the same meaning but several distinct attributes.

Gatys et al (via Jin et al, 2022, p. 184) initially investigated a Convolutional Neural Network (CNN) to collect information and stylistic aspects from photos individually. Their results showed that CNNs successfully retrieved content information from a random image and style data from a famous art piece. However, due to a lack of parallel corpora and trustworthy assessment measures, language style transfer has lagged behind some other fields like computer vision in terms of advancement (Fu et al, 2018, p. 663).

2.3 Low-resource languages

2.3.1 Overview

Historically, the use of TST has been focused in NLP for languages with a large corpus of training data to work from, with English being the focal language in most research (Garcia et al, 2021, p. 1). However, there has been a growing number of research papers in the past few years, particularly papers focused on languages in South and South-East Asia, that have explored how to use style transfer for low-resource languages. This research has explored the use of style transfer not only within the languages themselves, but also in the context of multilingual machine translation.

2.3.2 Challenges

Currently, models are prone to errors when attempting to operate with a limited corpus (Garcia et al, 2021, pp. 1-2; Krishna et al, 2022, p. 3). Considering this, Garcia et al (2021, p. 2) developed the first stages of a multilingual model, dubbed Universal Rewriter, by building on a pre-existing model. Their model focused on sentiment transfer (of which style transfer forms a part), among other aspects of MT. They used English style-transfer data to rewrite sentences in other languages, with the intent to compensate for limited data in their target languages by using English equivalents. This was generally successful, though native speakers identified some of the outputs as sounding unnatural. Krishna et al later tested the model's function, and found that there was also an issue with sentences being copied from the training data, rather than genuinely transferring style, for languages with little-to-no translation data (2022, pp. 2 - 3). These complications remain unsolved.

2.3.3 Style transfer within a low-resource language

Focusing on style transfer within a low-resource language, Wibowo et al (2020) worked on a model to transfer between Bahasa Indonesia (standard Indonesian) and informal Indonesian.

NLP and MT models solely use Bahasa Indonesia, not the significantly variant standard Indonesian. Further complicating the problem, informal Indonesian employs many loan words from other languages, and some words have different meanings to their use in Bahasa Indonesia, which can confuse NLP models.

Wibowo et al sought to begin addressing these issues in their research. They created a corpus of informal Indonesian, and manually annotated it with the Bahasa Indonesia equivalents. They then explored the efficacy of four different translation models. Their most successful model was phrase-based MT, or PBSMT. Processing language at the phrase level seems to be particularly effective in style transfer, as this is the level at which much nuance and pragmatic information is held in many languages.

The team then trained their PBSMT model with their corpus. They tried creating an additional artificial corpus of Bahasa Indonesia, which improved performance in early iterations before actively damaging the model's performance.

2.3.4 Multilingual style transfer

Within the style transfer of low-resource languages, it is common for the translation models to be semi-supervised in their approach. Fully supervised models are not viable within low-resource languages, and unsupervised models are still error-prone in this field (Garcia et al, 2021, p. 5; Krishna et al, 2022, p. 3). The researchers supplement the models with actions such as manually annotating data to create a corpus (Wibowo et al, 2020, p. 2), or acknowledge that they require large-scale feedback from native speakers to verify the success of and further improve their models (Garcia et al, 2021, p. 8).

There has yet to be devised an alternative to these semi-supervised methods, though there is a desire to be able to develop models that are completely unsupervised, so that low-resource languages can be fully accommodated in style transfer.

2.3.5 The future for low-resource style transfer

Multilingual and low-resource style transfer is still relatively undeveloped, and so solutions to the issues within the field are still in the early stages of exploration. Hu et al have proposed that future research in text style transfer should focus on designing models that are trained on the styles of languages other than English (2022, p. 24). Given that current models attempt to apply English stylistic properties to the style transfer of other languages, leading to unnatural language (Garcia et al, 2021, p. 8), it is clear that this needs to be addressed in order for the field of style transfer in low-resource languages to progress.

2.4 Types of Text Style Transfer methods

2.4.1 Overview

As laid out by Hu et al (2022, p. 4), based on the data sets utilised for model training, TST study is divided into three groups (see Fig.2).

- **Parallel Supervised** - Parallel corpora are collections of text pairings that each convey the same idea. In this specific data setting, the TST models are trained using known pairs of text written in different styles. NMT approaches like sequence to sequence (Seq2Seq) models and data augmentation are widely utilised to transfer the textual style.
- **Supervised Non-Parallel** - In the nonparallel supervised scenario, TST models seek to transfer the style of text without being aware of the text pairings that they are meant to match. The majority of recent TST studies fall into this category since parallel datasets are scarce in real-world TST applications.
- **Completely unmonitored** - The style labels are accessible to facilitate supervised training of the TST models in both parallel and nonparallel supervised data scenarios. Solely having an unlabelled text corpus leads to complications where in order to accomplish TST without prior knowledge of style labels, the TST models must be trained in an unsupervised manner.

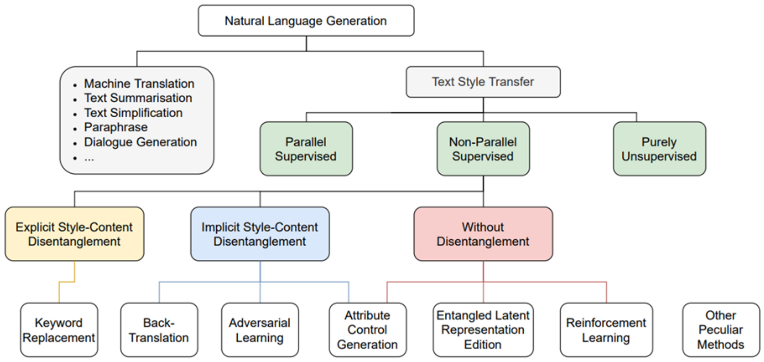


Figure 2 Taxonomy of style transfer methods (Hu et al, 2022, p. 5)

2.4.2 Methods used on Parallel Data

As Jin et al (2022, pp. 169-170) identified, the below methods are used on parallel data:

- **Seq-to-seq**: The seq2seq encoder-decoder model converts an input into an encoded representation of the model before sending it to the decoder, which generates the desired output. In this case, the size of the input and output vectors need not be constant. The purpose of this model's design is to allow it to handle data with no length restrictions. A single recurrent neural network will serve as both a decoder and an encoder.
- **Inference techniques**: To prevent the model from duplicating too many parts of the input sentence and failing to make enough edits to flip the attribute, first the words in the source sentence that need to be replaced are detected. Then, the words are altered by negative lexically constrained decoding, which prevents naïve copying.
- **Data Augmentation**: Through the use of a formality classifier, the back-translated English text is guaranteed to be formal by extracting informal content from internet forums and generating back-translations, which are translations from informal English to formal English using a pivot language like French.

2.4.3 Methods used on Non-Parallel Data

For non-parallel data, Jin et al (2022, pp. 170-171) identified three primary categories of unsupervised approaches:

- **Disentanglement**: In the embedding latent space, it disentangles text into its content and attribute, and then uses generative modelling.
- **Prototype editing**: ensures that the final text is still fluid by eliminating just the sentence fragments with the incorrect characteristics (for example, formal instead of informal) and substituting them with the words having right attributes.
- **Pseudo-parallel Corpus Construction**: trains the model on pseudo-parallel data as though it were supervised. Extraction of aligned phrase pairings from two mono-style datasets during retrieval is one method for creating pseudo-parallel data.

2.4.4 An example of parallel supervised text style transfer and multilingual analysis on conversational texts

2.4.4.1 Data Collection and Cleaning

The source scripts or datasets of conversations are retrieved in one language and then translated into another language using translators such as Google Translate. A set of grammatical rules are used to remove the majority of grammatical problems from the raw transcripts. Several words and phrases that appear repeatedly, notably in comedic or sarcastic contexts are condensed into a single repetition (Tikhonova et al, 2021, p. 2).

2.4.4.2 Data Primary Analysis

Language and style is analysed prior to training the models and statistics such as the total number of words, the average number of words in each sentence, the proportion of complicated words etc. is recorded. The language is then examined using a speech complexity metric. A word is regarded as complicated if it has more than four syllables. The variety of complex words in the conversation is used to comprehend the way of life, field of work, and other distinct characteristics of the participants in the conversation. At the end, sentiment analysis of one language is performed by determining the proportion of positive and negative words spoken in the conversation (Tikhonova et al, 2021, pp. 2-4).

2.4.4.3 Final Dataset Structure

Separate versions of the dataset are produced using the collected data. Based on the context of each project, a dataset or a combination of datasets is chosen to train the final model (Tikhonova et al, 2021, p. 4).

2.4.4.4 Training the models

Models are trained in two phases, with the second part including fine-tuning. Fine tuning entails making minor adjustments to the previously trained models and applying them to a comparable task. Using this fine-tuning strategy, large versions of models are trained in both languages (Tikhonova et al, 2021, pp. 4-5).

2.5 Challenges in TST

TST has restrictions much like other NLG tasks. In particular, there are issues with the methodologies currently in use for evaluating the results produced by TST and natural language production models. For instance, for particular NLG tasks, there are presently no accepted automatic assessment measures to score the contextual quality or informativeness of created texts. Additionally, there is limited agreement on the best way to perform human evaluations (Hu et al, 2022, p. 2). Chawla and Yang provide an overview of the key NLP challenges in mastering language complexity; namely, addressing informal phrases and abbreviations, decoding the contexts in which sentences are written, understanding sarcasm and rhetorical questions, the prevalence of grammar and spelling mistakes, and struggling with slang and idiomatic expressions (2020, p.9). Brands provides a higher-level view of why NLP is challenging, with a focus on the problem of low-resource languages, the need for large amounts of training data in order to create accurate TST models, and the time and resources needed to develop good TST models (2022).

Section 3 - Conclusion

The review provided a thorough analysis of Machine Translation and Text Style Transfer inculcating various languages, and its recent research progression. We can conclude that style transfer works best on high resource languages, and that low-resource languages face a plethora of issues which are still not a primary focus in much of the current research. We also summarise the issues that low resource languages face to perform TST tasks. It is also clear that most of the existing TST studies fall under the non-parallel supervised category. In the case of Machine Translation, we can conclude that Neural Machine Translation is the next step up from Statistical Machine Translation which requires larger sets of bilingual corpora. NMT requires a lot of processing power and current research is mainly focused on making it more efficient. With continued research, Machine Translation and Text Style Transfer have the potential to revolutionise the way in which text is written and be understood in the digital age. The potential applications of TST and MT are vast. We hope that this review will help readers gain a thorough understanding of the important facets of this area, delineate the key categories of TST techniques, and cast some light on future research.

References

Babhulgaonkar, A., & Sonavane, S. (2020). Language Identification for Multilingual Machine Translation. *2020 International Conference on Communication and Signal Processing (ICCSP)*, 401–405. https://doi.org/10.1109/ICCSP48568.2020.9182184

Brands, I. (2022, April 6). *Why Is NLP Challenging?* Biostrand. https://blog.biostrand.be/en/why-is-nlp-challenging

Brownlee, J. (2017, September 21). *What Is Natural Language Processing?* MachineLearningMastery.Com. https://machinelearningmastery.com/natural-language-processing/

Chawla, K., & Yang, D. (2020). *Semi-supervised Formality Style Transfer using Language Model Discriminator and Mutual Information Maximization*. arXiv. https://doi.org/10.48550/arXiv.2010.05090

Ding, Y., Liu, Y., Luan, H., & Sun, M. (2017). Visualizing and Understanding Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1 (pp. 1150-1159). https://doi.org/10.18653.v1.P17-1106

Fu, Z., Tan, X., Peng, N., Zhao, D., & Yan, R. (2018). Style Transfer in Text: Exploration and Evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11330

Garcia, X., Constant, N., Guo, M., & Firat, O. (2021). *Towards Universality in Multilingual Text Rewriting*. arXiv. https://doi.org/10.48550/arXiv.2107.14749

Hu, Z., Lee, R.K.W., Aggarwal, C.C., & Zhang, A. (2022). *Text Style Transfer: A Review and Experimental Evaluation*. arXiv. https://doi.org/10.48550/arXiv.2010.12742

Hutchins, J. (2010). Machine translation: a concise history. *Journal of Translation Studies 13 (1 & 2), pp. 29-70.*

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., & Mihalcea, R. (2022). Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, *48*(1), 155–205. https://doi.org/10.1162/coli_a_00426

Kim, Y., Graça, M., & Ney, H. (2020). When and Why is Unsupervised Neural Machine Translation Useless? In *Proceedings of the 22nd Annual conference of the European Association for Machine Translation* (pp 34-44), Lisboa, Portugal. European Association for Machine Translation . https://doi.org/10.48550/arXiv.2004.10581

Krishna, K., Nathani, D., Garcia, X., Samanta, B., & Talukdar, P. (2022). *Few-shot Controllable Style Transfer for Low-Resource Multilingual Settings*. arXiv. https://doi.org/10.48550/arXiv.2110.07385

Lai, C.-T., Hong, Y.-T., Chen, H.-Y., Lu, C.-J., & Lin, S.-D. (2019). Multiple Text Style Transfer by using Word-level Conditional Generative Adversarial Network with Two-Phase Training. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3579–3584. https://doi.org/10.18653/v1/D19-1366

Lake, B. M., & Murphy, G. L. (2021). *Word meaning in minds and machines*. arXiv. https://doi.org/10.48550/arXiv.2008.01766

Phadke, M. M., & Devane, S. R. (2017). Multilingual machine translation : An analytical study. *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 881–884. https://doi.org/10.1109/ICCONS.2017.8250590

Stahlberg, F. (2020). Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research*, *69*, 343–418. https://doi.org/10.1613/jair.1.12007

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020) Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1 (pp. 5-21). https://doi.org/10.1016/j.aiopen.2020.11.001

Tikhonova, M., Telesheva, E., Mirzoev, S., Tarantsova, P., Petrov, S., & Fenogenova, A. (2021). Style transfer in NLP: a framework and multilingual analysis with Friends TV series. *2021 International Conference Engineering and Telecommunication (En&T)*, 1–6. https://doi.org/10.1109/EnT50460.2021.9681722

Wen, T.-H., & Young, S. (2020). Recurrent neural network language generation for spoken dialogue systems. *Computer Speech & Language*, *63*, 101017. https://doi.org/10.1016/j.csl.2019.06.008

Wibowo, H. A., Prawiro, T. A., Ihsan, M., Aji, A. F., Prasojo, R. E., Mahendra, R., & Fitriany, S. (2020). Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation. *2020 International Conference on Asian Language Processing (IALP)*, 310–315. https://doi.org/10.1109/IALP51396.2020.9310459

Xu, R., Ge, T., & Wei, F. (2019). *Formality Style Transfer with Hybrid Textual Annotations*. arXiv. https://doi.org/10.48550/arXiv.1903.06353

Yadav, D., Desai, J., & Yadav, A. K. (2022). *Automatic Text Summarization Methods: A Comprehensive Review.*

Project Report

Group 42

1 Introduction

Over the course of 11 weeks, our group worked towards the completion of a literature review on the subject of Natural Language Processing (NLP) in Machine Translation (MT) with a focus on Style Transfer. This report will detail various aspects of the process which we went through to achieve this, as well as some interesting interpretations of our findings.

Section 2 describes how the group worked together to complete the tasks involved in completing the literature review, as well as both preparing for and then performing the presentation. **Section 3** presents the challenges faced by the group and the solutions we implemented in order to mitigate them. **Section 4** covers some interesting findings and limitations from the literature that, while not irrelevant to the overall research, did not belong in the review itself. Finally, **Section 5** concludes what the group have learned in the overall process as well as what can be carried forward for use in future projects.

2 Teamwork and Project Management

2.1 Overview

Our group was comprised of 7 students all from the MSc Data Science and Artificial Intelligence course, each with a different background and set of experiences. Three of the group members had a background in foreign language learning, linguistics and translation, two members had experience looking at NLP, and the other two had experience with Artificial Intelligence, Machine Learning and sentiment analysis. Nevertheless, our broad shared interest in language, linguistics, multilingual NLP, and MT brought us together.

We started by reading around the topics that we were considering for the project, which included:

- NLP in the context of chatbots
- Applications of NLP and machine translation in the medical field
- Multilingual semantic analysis
- Applications of multilingual NLP on the web
- Translation of style in subtitles and dubbing

We then decided to focus on style transfer after coming across an interesting paper which discussed how the personality of each character of a popular sitcom could be preserved after translating subtitles or dubbing into another language (see Tikhonova et al., 2021).

2.2 Project Planning and Management

Once we had decided on a suitable topic for our Literature Review, we began our project planning by dividing the work between us. Being a larger group of 7 people with diverse backgrounds, we wanted to make sure that every member could give a worthwhile contribution. The four main areas we looked at when deciding group roles were:

- Prior technical knowledge and experience in the field

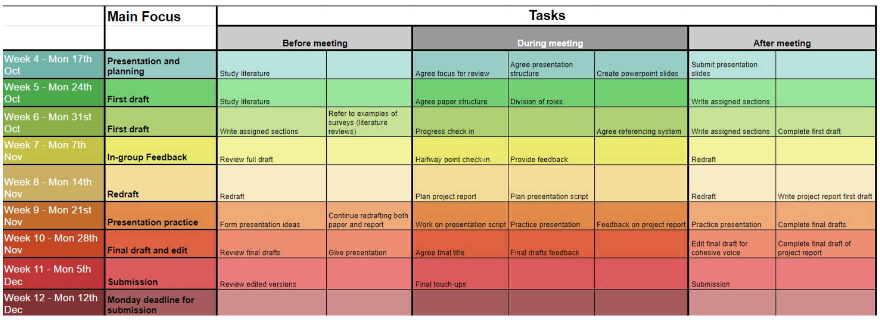
- Organisational skills
- Public speaking skills
- Proofreading and formatting

At this point, we assigned the tasks as shown in Table 1. Naturally, more tasks were added over time and roles were changed to better suit natural abilities.

Everyone	Read literature, provide references, scripting presentation
Rahul	Prepare for and answer questions after the presentation, write the NLP section
Aiswarya	Give presentation, write conclusion and scope for future work section
Artemis	Give presentation, write section on low-resource languages
Alex	Write part of project report, write section about machine translation
Kevin	Give presentation, write introduction
Alice	Edit structure, write abstract, write part of project report
Digvijay	Proofreading, add key words, create references, focus on format and presentation

Table 1: Distribution of tasks

We made sure to be very clear about what everyone was expected to do before each weekly meeting (see Figure 1). We used a Gantt chart (see Figure 2) to track our progress throughout the project.



	Main Focus	Tasks						
		Before meeting		During meeting			After meeting	
Week 4 - Mon 17th Oct	Presentation and planning	Study literature		Agree focus for review	Agree presentation structure	Create powerpoint slides	Submit presentation slides	
Week 5 - Mon 24th Oct	First draft	Study literature		Agree paper structure	Division of roles		Write assigned sections	
Week 6 - Mon 31st Oct	First draft	Write assigned sections	Refer to examples of surveys (literature reviews)	Progress check in		Agree referencing system	Write assigned sections	Complete first draft
Week 7 - Mon 7th Nov	In-group Feedback	Review full draft		Halfway point check-in	Provide feedback		Redraft	
Week 8 - Mon 14th Nov	Redraft	Redraft		Plan project report	Plan presentation script		Redraft	Write project report first draft
Week 9 - Mon 21st Nov	Presentation practice	Form presentation ideas	Continue redrafting both paper and report	Work on presentation script	Practice presentation	Feedback on project report	Practice presentation	Complete final drafts
Week 10 - Mon 28th Nov	Final draft and edit	Review final drafts	Give presentation	Agree final title	Final drafts feedback		Edit final draft for cohesive voice	Complete final draft of project report
Week 11 - Mon 5th Dec	Submission	Review edited versions		Final touch-ups			Submission	
Week 12 - Mon 12th Dec	Monday deadline for submission							

Figure 1: Weekly project schedule with detail

3 Challenges and Solutions

3.1 Illness

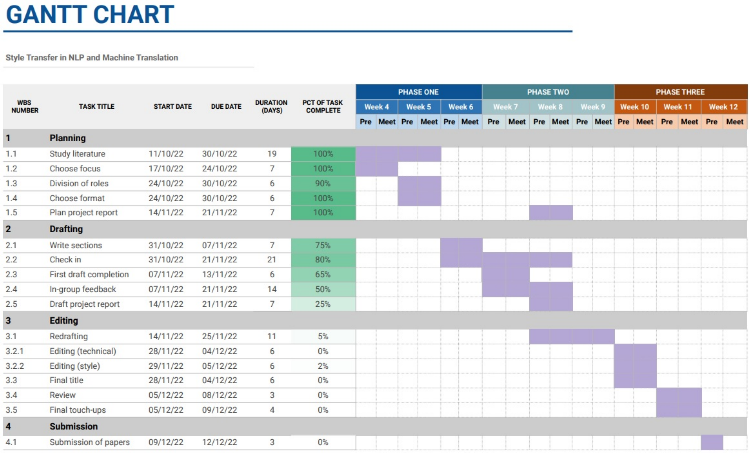
Unfortunately, two of the students in our group became very unwell and were hospitalised during the semester; one on the 1st of December and another on the 9th of December. The first student was responsible for proofreading, formatting, writing the abstract and the section about low-resource languages for the Literature Review, and writing part of the project report. The second student was responsible for writing the section covering machine translation in the Literature Review and part of the project report. Both students informed the group on the same day and the group discussed how it could redistribute some of the tasks. By employing good communication, we ensured that both students could properly recover and that all tasks were covered.

3.2 Logistics

One of the main challenges we faced was finding a suitable day and time to meet each week and review the progress of the project. We tried to use planning tools such as Doodle (www.doodle.com) to organise a convenient slot, however we were unable to find one where everyone was available. We resolved this issue by meeting during timetabled lecture hours for the module after Week 4, when the lectures had ended. Whilst scheduling conflicts sometimes arose meaning that not every group member could attend each meeting, overall we found our sessions to be productive.) to enable us to have oversight of the progress of the project

3.3 Delays against Project Schedule

At the end of October, we suffered a setback due to clashes with assignments on other modules, and we weren't able to start the next section on the planned date. However, wanting to keep the schedule on course, we were able to shorten the duration of the working stage - essentially working faster - and we were able in this way to catch up on the planned start date for the next stage in week 10. By constantly checking in with our progress on our Gantt chart and using it as a flexible reference point, we gained a degree of oversight with relation to our strict deadline.



GANTT CHART

Style Transfer in NLP and Machine Translation

WBS NUMBER	TASK TITLE	START DATE	DUE DATE	DURATION (DAYS)	PCT OF TASK COMPLETE	PHASE ONE						PHASE TWO						PHASE THREE					
						Week 4		Week 5		Week 6		Week 7		Week 8		Week 9		Week 10		Week 11		Week 12	
						Pre	Meet	Pre	Meet	Pre	Meet	Pre	Meet	Pre	Meet	Pre	Meet	Pre	Meet	Pre	Meet	Pre	Meet
1	**Planning**																						
1.1	Study literature	11/10/22	30/10/22	19	100%	■	■	■	■														
1.2	Choose focus	17/10/22	24/10/22	7	100%	■	■																
1.3	Division of roles	24/10/22	30/10/22	6	90%			■	■														
1.4	Choose format	24/10/22	30/10/22	6	100%			■	■														
1.5	Plan project report	14/11/22	21/11/22	7	100%									■	■								
2	**Drafting**																						
2.1	Write sections	31/10/22	07/11/22	7	75%					■	■												
2.2	Check in	31/10/22	21/11/22	21	80%					■	■	■	■	■	■								
2.3	First draft completion	07/11/22	13/11/22	6	65%							■	■										
2.4	In-group feedback	07/11/22	21/11/22	14	50%							■	■	■	■								
2.5	Draft project report	14/11/22	21/11/22	7	25%									■	■								
3	**Editing**																						
3.1	Redrafting	14/11/22	25/11/22	11	5%									■	■	■	■						
3.2.1	Editing (technical)	28/11/22	04/12/22	6	0%													■	■				
3.2.2	Editing (style)	29/11/22	05/12/22	6	2%													■	■				
3.3	Final title	28/11/22	04/12/22	6	0%													■	■				
3.4	Review	05/12/22	08/12/22	3	0%															■	■		
3.5	Final touch-ups	05/12/22	09/12/22	4	0%															■	■		
4	**Submission**																						
4.1	Submission of papers	09/12/22	12/12/22	3	0%																	■	

Figure 2: Gantt chart used for project management

4 Further Findings, Interesting Remarks and Limitations

Throughout the project we came across various subjects of research that were related to our area of interest, but not strictly relevant to the style transfer focus we had decided. However, the importance of the insights this gave us means that we think it is still worthwhile to make note of them here.

4.1 Limits of Human Translation

One part in our limits of machine translation section touched upon the more philosophical question of whether we can ever expect a machine to truly capture the feeling of a given text entirely, separately from its functional content.

Karpińska explores this through the lens of the long standing *Theory of Equivalence* in relation to MT capabilities. The theory represents the as yet unresolved division in focus with relation to the translation of a piece of text. The first focus, known as *Formal Equivalence*, places the most importance on strict adherence to maintaining as much as possible the layout, form, and syntactic structure of the base text when translating into the target text. As Nida says via Karpińska, "the target text should appear as foreign to the reader, on both formal and cultural level" (2017, p.6). Here, the otherness of the base text becomes a feature; a part of its appeal. Rather than downplaying the foreign nature of the base text by attempting to sound more natural, the process of translation with formal equivalence can instead be used to highlight the differences between the two languages. This makes the reader aware that they are reading about something that was originally devised in terms and possibly ways of thinking different to their own.

In opposition to this, is the idea of *Dynamic Equivalence*, which explores the idea that the most important part of the text is the effect it has on the reader, and that a translator should be seeking to replicate that effect as closely as possible when changing the text into the target language. Here the goal is on much more subjective, dealing instead with feeling and intuition rather than concrete things. Using *Dynamic Equivalence*, the content of the translated text could easily differ greatly from the source text in terms of structure and vocabulary, often employing less commonly used words as translations where the translator felt that word could achieve the same nuance that was present in the original text. To translate successfully within the sphere of *Dynamic Equivalence* would require a high level of cultural and linguistic knowledge of both source and target languages on the part of the translator . Kielar notes (via Karpińska) that "a translator is expected to have empathy as well as cultural and linguistic sensibility. The text should sound naturally[sic] in the target language and its readers should stay within their comfort zone and the sphere of their culture" (2017, p.6).

It is generally accepted that the best translation, rather than being wholly formal or wholly dynamic in its equivalence, will instead find itself somewhere in between the two, at times in the text relying more heavily on direct translation and at times focusing on the emotions conveyed to the reader through word choice. Between the two ideas of equivalence, it is clear that MT is better suited to the formal side, splitting sentences down into abstract small units and then performing direct translations on them. As Karpińska points out, this makes them best suited to translating technical texts, where the goal is to convey as clearly as possible the functional information contained in the source text, and the source texts themselves are usually written in such a way as to leave little or no room for interpretation (2017, p.7). However, most texts fall outside of this specification, and so in most cases there is at least some necessity for application of dynamic equivalence in the translation of a text. A human translator can choose where to employ more formal translation and when to use their own cultural knowledge to embellish or alter the text, but a piece of AI translation software does not have this ability, and is unlikely to in the near future. As such it's likely that the last area that MT will make progress in is texts where one is required to capture the "feeling" of the sentence, rather than the surface level meaning, examples being works of literature or poetry. Until a machine can be 'taught' the historical and cultural background of a text that it's translating, it won't be able to convey that imagery into the target text language.

Of course, one could argue that even amongst humans, there exists no "perfect translation", as the feelings evoked by a piece of text naturally differ between individuals, and so true dynamic translation of a source text is technically impossible. However, that is outside the realm of discussion on MT so we it didn't fall neatly under the purview of our literature review.

4.2 Lack of Consistency and Self-Regulation within the Field of in Machine Translation

While reading through the various papers on the subject of Machine Translation, one common through line was their use of the same testing software to assess the quality of the translations, this being the BLEU (BiLingual Evaluation Understudy) score. However, midway through the project we came across one paper that called into question the validity of this scoring system (see Post 2018).

In Post's paper, he attacks the credibility of the results of many papers that use the BLEU score by pointing out the inconsistency with which the scoring metric is employed. BLEU is not a monolithic metric for which the only variable is the information fed into it, but rather "a parameterized metric whose values can vary wildly with changes to these parameters" (Post, 2018, p.1). For the results of a study to be accepted as true, they must be reproducible, but with the variance in the parameters fed to BLEU on a case by case basis it is likely that the same machine learning algorithm carried out on the same dataset could have significantly different results, with a variance of as high as 1.8 (Post, 2018, p.2).

The reason that this is a problem is that currently the authors of a paper publishing results of a machine learning model are not required to include the configurations of BLEU they used. Configurations such as number of references used or maximum n-gram length are rarely mentioned, despite this drawing into doubt the reliability of their results. In some cases fairly accurate assumptions can be made - maximum n- gram length is usually 4 and only one reference is necessary - but the uncertainty and lack of information is still inherently unscientific and the subsequent results could be said to be likewise. As a solution, Post suggests that, going forward, research groups should always include details of parameterization within their results, and proposes a Python script SacreBleu which would standardize and more explicitly provide details of the configuration used to obtain the results (Post, 2018, p.4). However, at the moment these practices have not become standard and remain rare, and as such many machine translation papers could be said to be unreliable, which is counterproductive to research. Where this topic pertains to human error within the field, this was not included in the literature review.

5 Conclusion

One of the main goals of this project was to become not only more knowledgeable in our chosen field of research, but also to develop our collaborative skills, and gain the ability to accurately gauge the expected length to completion of various tasks, and improve our time management in order to meet deadlines based upon that expected length. By working together as a group and with each individual member making an effort to contribute, we were able to achieve this with confidence, and we believe that the skills gained here will be of great use for future projects either solo or once again as a group.

6 Bibliography

Karpińska, P. (2017). Computer Aided Translation - possibilities, limitations and changes in the field of professional translation. Journal of Education Culture and Society, 8(2), 133-142.

Post, M. (2018). A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186 -191, Brussels, Belgium. Association for Computational Linguistics.

Tikhonova, M., Telesheva, E., Mirzoev, S., Tarantsova, P., Petrov, S., & Fenogenova, A. (2021, November). Style transfer in NLP: a framework and multilingual analysis with Friends TV series. In 2021 International Conference Engineering and Telecommunication (EnT) (pp.1-6). IEEE.