Human vs Machine Translatorship Attribution at the Document Level

Mingyue Jian

1 Introduction

Authorship attribution has been a well-established research task within the field of Natural Language Processing (NLP). The goal of the task is to apply computational modelling methods to attribute anonymous documents to different candidate authors based on the various features measured. Common methods for authorship attribution include TF-IDF, n-gram, machine learning classifiers like support vector machines (SVMs), logistic regression and so on [1]. Translatorship attribution is a similar task that aims to distinguish translators by different features measured in the translated document from the same source. It has been proven by Mona Baker that there will be a distinctive fingerprint for a literary translator, which can be used as a basis for translatorship identification for distinguishment [2]. However, research in this area is limited (only a few papers available) and has primarily focused on translatorship attribution for human translators [3, 4].

With the increasing popularity of machine translation in recent years, the best performance in translation tasks has been achieved using Neural Machine Translation (NMT) [5–7]. In addition to high accuracy in sentence-level translation, previous studies have also shown that NMT demonstrates superior performance in document-level translation [8]. This project aims to conduct translatorship attribution for both human and machine translators at the document level, specifically for translations of the same literature Solaris by Stanisław Lem in different languages. Research questions that will be addressed include: Is there a unique fingerprint between human and machine translators? How does the translation source affect the translation fingerprint for translators (human/machine)?

The potential applications of this project include in plagiarism detection for human translation work, particularly in academic institutes. Additionally, the findings of this project could provide insight into the stylistic differences between human and machine translation for further improvement in the machine translation field.

2 Dataset Construction

Dataset utilised in this project was constructed from drafts, as no pre-existing dataset was suitable for the specific aims of the project.

2.1 The selection of the literary work for the project

The literature work selected for this project is "Solaris" by Polish writer Stanisław Lem. There are several reasons for this selection. Firstly, "Solaris" is a science fiction novel published in 1961, which features modern language. Secondly, the book is divided into chapters, making it suitable for attribution research. Thirdly, the book features long sentences with extensive use of metaphors, which requires translators to introduce creativity in order to preserve the original intent, thereby providing a more thorough test of the machine translators' ability. Fourthly, as the target language for the task is English, two different English translations of the book are available. Furthermore, the second English translation (2011) is by Bill Johnston, who translated the book directly from Polish, while the first English translation (1970) was translated from the French version (1966) of the book by co-translators Joanna Kilmartin and Steve Cox. This allows for a more diverse comparison based on layers of translation.

2.2 The selection of machine translator

Two machine translators were selected for this project in order to facilitate parallel comparison. Large-language models (LLMs), such as GPT-3, are considered to be the best language models for many NLP tasks as they are trained on vast amounts of data. However, due to the memory limitations of the attention mechanism, these models may struggle to process long sentences. Additionally, GPT-3 is not currently open-source for translation tasks. These factors led to the decision to not utilize LLMs in this project. Among the remaining translation engines, Google Cloud Translation [9] and DeepL API [10] were selected for their utilisation of neural machine translation, support for Polish, English, and French translations, and their prevalence in the field.

2.3 The construction of the dataset

The original version of the book (Polish) was obtained from docer . The first English version was downloaded from Library Genesis, while the second English version and the French translation were obtained from Scribd. After acquiring the source files, each version was manually separated by chapter, and extraneous elements such as footers, front-pages, and non-textual content were removed. Table1 displays the original version and the human translations of the dataset after manual cleaning and preprocessing.

The original version(Polish) of the book was subsequently translated into English and French by chapter using the APIs provided by Google and DeepL respectively.

Language	Translator	Word Count	Published Year	Translation Source
Polish	-	56,623	1961	-
French	Jean-Michel Jasienko	71,620	1966	Polish version
English	Joanna Kilmartin and Steve Cox	67,456	1970	French version
English	Bill Johnston	75,544	2011	Polish version

Table 1: The original source and translations for Solaris.

2.4 Dataset processing

The following steps were taken to process the dataset:

1. The format of punctuation marks in the HTML code was converted to the standard format.
2. Special characters, excluding alphanumeric characters and whitespace, were filtered out. Punctuations were removed as they were not relevant to the scope of the project. Special characters were replaced with their corresponding letters, as determined by examining parallel chapters for confirmation. Remaining special characters were replaced with an empty string.
3. The entire corpus was converted to lowercase.
4. Proper nouns, including character and place names, were filtered out. These were identified manually in the source text.
5. The cleaned and relevant data was written to a new directory for future use.

2.5 Dataset overview

Following the data processing, six folders were created, each containing 14 chapters in separate documents for analysis. Table2 provides details of the cleaned documents.

3 Methodology

The methodology employed in this project is n-gram analysis. To accomplish this, term frequency-inverse document frequency (TF-IDF) vectors were constructed to describe 1-gram, 2-gram, and 3-gram in each chapter across the entire corpus.

Folder Name	Word Count	Description
fr-h	66,245	the first English translation version by human
po-h	74,536	the second English translation version by human
fr-dl	67,447	the English translation from French by DeepL
fr-goo	67,042	the English translation from French by Google
po-dl	70,236	the English translation from Polish by DeepL
po-goo	65,434	the English translation from Polish by Google

Table 2: cleaned data

3.1 The feature set

Frequently employed for authorship attribution, features such as function words and most frequent-used words are also considered to be important fingerprints of style for translators.

- For 1-gram, two features are gathered for analysis:
  1. The distribution all the unique word, by chapters and by documents. This will give an overview of word distribution for different translations.
  2. The distribution of function words for each chapter. There are several reasons for the importance of the function words for translatorship attribution [4,11]. (1) Function words are one of the discriminations for translators as they are not greatly affected by the source context as lexical words. (2) They are reliable, irreplaceable and always occupy a high proportion for every English document because of the structure of English language, so it's always easily comparable for different document.
  3. The distribution of top 200 most frequent words for each chapter.
- For 2-gram and 3-gram, the only features gathered are the n-grams by document as well as the most frequent n-gram by documents. As function words are all unigram, and also because distinctive results are already shown by the features mentioned.

3.2 Visualisation

The constructed TF-IDF vectors are of very high dimensions based on the unique n-gram, it's hard to visualise it by features for analysis. Thus, principal component analysis (PCA) is applied for better visualisation. PCA transforms vectors into a lower dimensional space in a way that preserves the maximum amount of variance. Via PCA, features from TF-IDF can be plotted into a 2D graph by matplotlib for better visualisation and analysis.

3.3 Classification

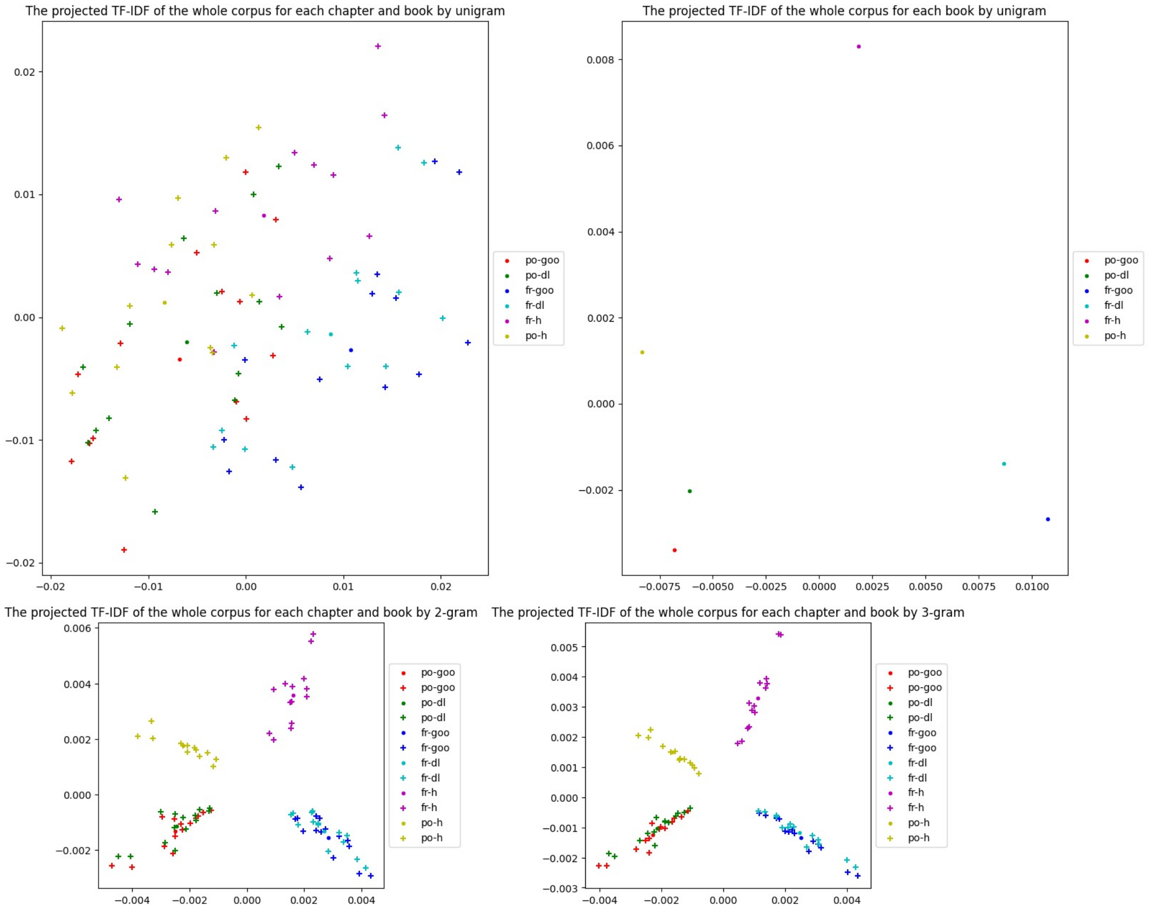
Classification in machine learning is applied for feature classification for document pair. Compared to logistic regression, the project applied support vector machines (SVMs) for classification because of the better accuracy for classification tasks for pairs. Accuracy for classification results will be calculated to display the confidence of difference between the features of two documents.

4 Results

Result analysis is derived from the implementation results and discussed based on the proposed research questions.
Q1. Is there a unique fingerprint between human and machine translators?
Q2. Will the translation source affect the translation fingerprint for translators (human/machine)?

4.1 Projected TF-IDF features by n-gram of the whole corpus

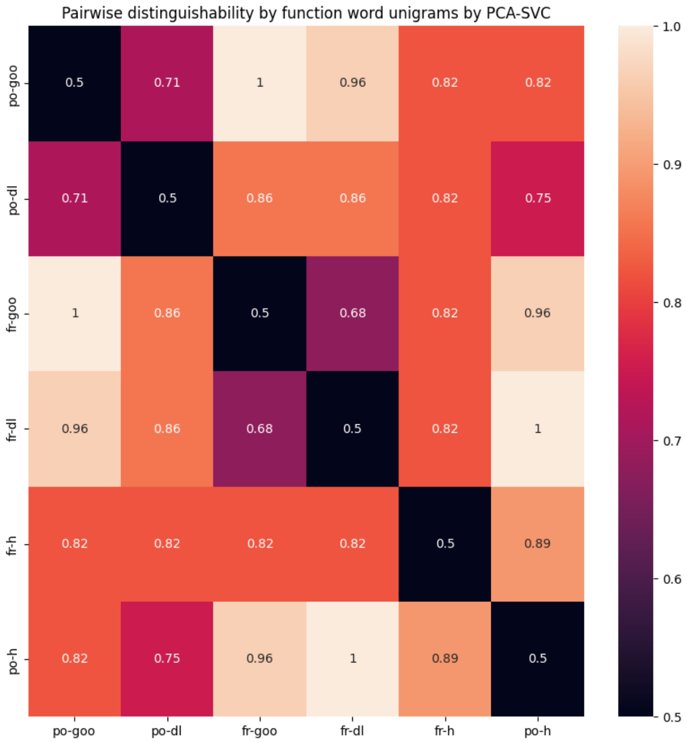
When plotting the projected TF-IDF of the entire corpus for each chapter and book untilising n-gram, distinguishment between each translation work is clearer as the number n increases. Based on Graph 1, a significant difference can be observed from 2-gram and 3-gram between human and machine translation, and with translations from different language sources. For human and machine translation, the dots for human-translated documents are on the top, whereas the machine translated documents are at the bottom. Same with translations from different language sources, with language source being French on the right, and Polish on the left. These findings suggest that translatorship attribution could be based on the language source or even the layers of translations. Additionally, it can be inferred that machine translations from the same source language are highly similar to one another, despite being generated by different engines.



Graph 1

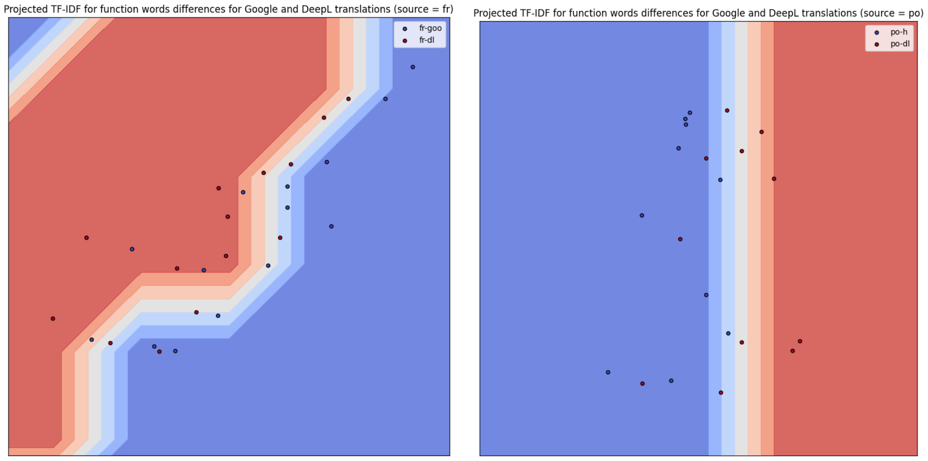
4.2 Projected TF-IDF features by function words of the whole corpus

Graph 2 illustrates the pairwise distinguishability by function word by PCA-SVC. The distinguishability is the accuracy result for SVC. The darker colour (the higher the number) in each cell, the harder it is to distinguish the style of the two documents, and vice versa. As can be seen in Graph 2, when measuring function words, it's highly likely to distinguish between human and machine translations, particularly when the source language is different, as indicated by the two leftmost columns – with the exception of the black diagonal cells, the remaining cells have values above 0.75. It is noteworthy that, when compared to machine translation from French, the second human English translation, which has a source language of Polish, shows more similarity to machine translation from Polish (0.82 and 0.75) than from French (0.96 and 1). On the other hand, the first human English translation, which has a source language of French, displays the same classification accuracy score for both languages by machine translations (0.82). It is not sufficient to convey the distinguishability solely through classification accuracy numbers, and it is necessary to establish benchmarks in the future to provide a clearer understanding.



Graph 2

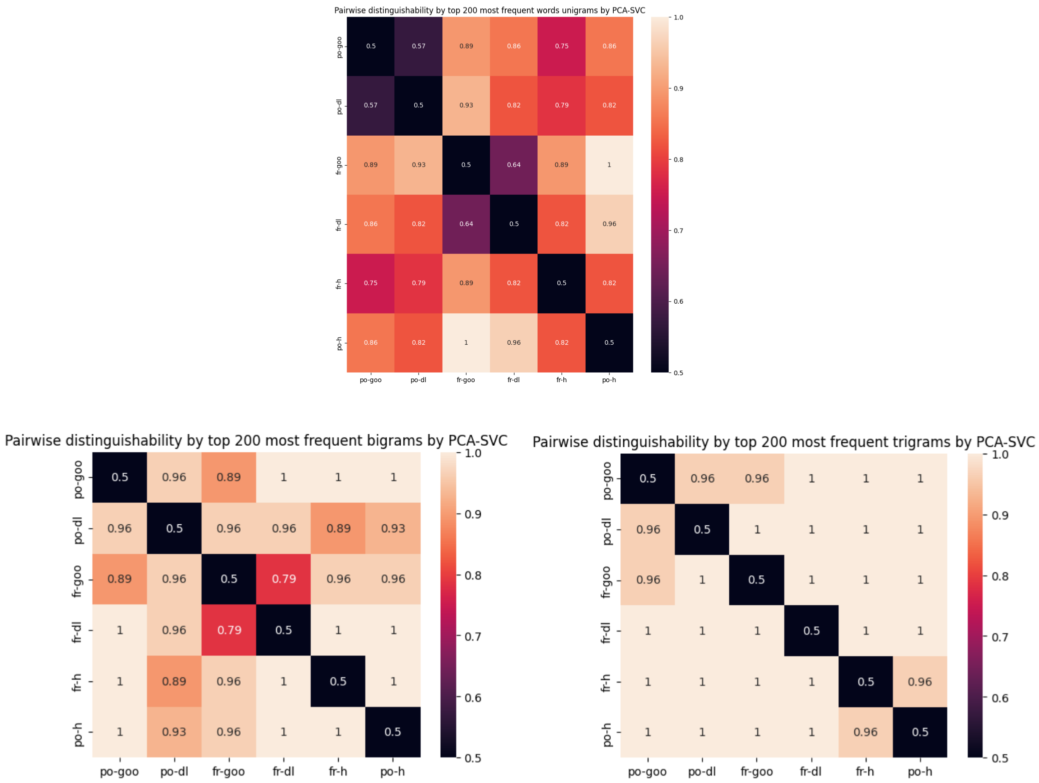
Upon further examination, it can be inferred that it is difficult to differentiate machine translations from one another when they are translated from the same source. Graph 3 depicts the projected TF-IDF graphs for machine translations of function words based on different source languages. The distinctiveness of the chapters on the graph is limited, as evidenced by the classification accuracy scores of 0.68 (source = fr) and 0.71 (source = po).



Graph 3

4.3 Projected TF-IDF features by top 200 most frequent n-grams of the whole corpus

Graph 4 illustrates the pairwise distinguishability of n-grams using PCA-SVC, separately for n=1,2,3. It's evident that, as the number n utilised as features increases, the likelihood of a higher accuracy score for SVC classification also increases. Human and machine translations can be clearly distinguished using trigrams, as evidenced by the accuracy scores of 1.0.



Graph 4: Heatmaps of SVC accuracy on chapter distinguishing between each pair of books, for the top 200 most popular 1-grams, 2-grams, and 3-grams.

Machine translations from Polish, while only being distinguishable with very low accuracy (0.57) using unigrams, are distinguishable with much higher accuracy using higher-order-grams (0.96). However, translations from French are much more difficult to distinguish by bigrams, with an accuracy of only 0.79 (and 0.64 for unigrams), but again are easily separable in trigrams (accuracy 1).

Another interesting result is that the result of human translation from the French text is more stylistically alike to machine translations of the original Polish text (accuracy of 0.75 and 0.79 with unigrams) than when the source languages match (accuracy 0.82-0.89) and especially compared to the results when languages are swapped (accuracy 0.96-1.0). This is an interesting result because it demonstrates that source language is far from being the only factor in the writing style of a machine-translated document.

4.4 Model evaluation

The model developed in the project has provided some insights into translatorship attribution between human and machine translators at the document level. However, these conclusions may not be generalisable to other scenarios. The reason lies in that, firstly, the corpus of the project is limited to a single piece of literature, which is both small and specific in nature. Additional research with larger and more diverse corpora is needed to further validate the findings of this project. Secondly, only two human translators were considered – it may be the case that other human translators, maybe especially novice ones, have writing styles more similar to machines. Apart from that, the result of this study demonstrates that attributing translatorship between human and machines is possible.

5 Conclusion

In conclusion, this project applies several computational linguistics methods to the task of translatorship attribution for human and machine translation at the document level. It's demonstrated that translatorship can be distinguished by n-grams and function words, although performance between the two differs. These findings show that the style of translation will be greatly affected by the source text for both humans and machines. It also shows that machine translators, especially Google Translate, each have their own writing style which is somewhat consistent between source languages. Although these styles can be hard to distinguish from each other, they are much more easily distinguishable from human writing styles.

References

[1] S. Swain, G. Mishra, and C. Sindhu, "Recent approaches on authorship attribution techniques—an overview," in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 1. IEEE, 2017, pp. 557–566.

[2] M. Baker, "Towards a methodology for investigating the style of a literary translator," *Target. International Journal of Translation Studies*, vol. 12, no. 2, pp. 241–266, 2000.

[3] C. Caballero, H. Calvo, and I. Batyrshin, "On explainable features for translatorship attribution: Unveiling the translator's style with causality," *IEEE Access*, vol. 9, pp. 93 195–93 208, 2021.

[4] E. Mohamed, R. Sarwar, and S. Mostafa, "Translator attribution for arabic using machine learning," *Digital Scholarship in the Humanities*, 2022.

[5] F. Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020.

[6] S. Wang, Z. Tu, Z. Tan, W. Wang, M. Sun, and Y. Liu, "Language models are good translators," *arXiv preprint arXiv:2106.13627*, 2021.

[7] I. Rivera-Trigueros, "Machine translation systems and quality assessment: a systematic review," *Language Resources and Evaluation*, vol. 56, no. 2, pp. 593–619, 2022.

[8] E. Matusov, "The challenges of using neural machine translation for literature," in *Proceedings of the qualities of literary machine translation*, 2019, pp. 10–19.

[9] Google, "Cloud translation documentation — google cloud." [Online]. Available: https://cloud.google.com/translate/docs

[10] DeepL, "Deepl translate api: Machine translation technology." [Online]. Available: https://www.deepl.com/pro-api?cta=header-pro-api/

[11] J. N. G. Binongo, "Who wrote the 15th book of oz? an application of multivariate analysis to authorship attribution," *Chance*, vol. 16, no. 2, pp. 9–17, 2003.