Proposal: Taz Scott-Talib, Alberto Lopez Garcia

Authorship attribution and analysis is a hot topic in modern natural language processing, but is largely centred around the English language, and works written exclusively in English. While a variety of recent methods have successfully distinguished writings from different authors using lexical, syntactic and semantic information (Wu et al. 2021), this issue become much more complex and challenging when taken to the world of translated works (Caballero et al. 2021). Due to major differences in syntax between various languages, the task of translation involves much more than finding like-for-like word replacements, resulting in fundamentally different text structures.

Our team intends to apply a known authorship attribution NLP model (e.g., in the style of Bagnall 2016 or Wu et al. 2021) to translated works of major authors. Using multiple works from the same authors and from the same translators for consistency, we aim to identify both the original author of a translated text, as well as its translator. Based on our results, we believe we might be able to answer certain questions on the effect of literary translation on both text style and authorship analysis.

For instance, we aim to observe how a successful authorship analysis model applies to translated text and determine what hyperparameters or preprocessing steps might need to be modified to maintain strong results. In addition, we may be able to determine whether a translator is able to accurately convey the writing style of the original author, or whether their work has syntactic qualities of its own, creating a writing style consistent across works that can be associated with the translator. This will of course be gauged based on our model's ability to identify the translator and/or the original author of a translated text. Finally, our study should identify the aspects of authorship lost in translation: what qualities (vocabulary, sentence structure, syntax, etc.) are most/least valuable when it comes to attributing translated works to their original authors? Characteristics that have little to no effect in this task may have not been accurately captured in translation.

While a few papers have studied authorship attribution in the context of translated works (see Mikros 2020 and Caballero et al. 2021), the topic is relatively untouched in the world of NLP. These authors also focus more on whether the task is possible, rather than identifying the characteristics of translated text that allow for analysis. We believe our questions go beyond this, and our motivation would allow us to make statements about the effects of translating literary works on writing style and semantics.

Listed after the references are two datasets which we found that may suit our task.

References

D. Bagnall, "Author Identification using Multi-headed Recurrent Neural Networks", *CLEF Author Identification challenge*, 2016. [Abstract]. Available: ArXiv, https://doi.org/10.48550/arXiv.1506.04891.

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. 93195-93208, 2021. [Abstract]. Available: IEEE Xplore, https://ieeexplore.ieee.org/abstract/document/9467290.

G. K. Mikros, "Finding the author of a translation. An experiment in authorship attribution using machine learning methods in original texts and translations of the same author." *Words and Numbers. In Memory of Peter Grzybek (1957-2019)*, 2020. Available: University of Athens, https://eclass.uoa.gr/modules/document/file.php/ISLL125/Test%20paper%20(Finding%20the%20author%20of%20a%20translation).doc.

H. Wu, Z. Zhang and Q. Wu, "Exploring syntactic and semantic features for authorship attribution", *Applied Soft Computing*, vol. 111, 2021. [Abstract]. Available: ScienceDirect, https://doi.org/10.1016/j.asoc.2021.107815.

Potential challenges

One problem we expect to encounter while carrying out this experiment involves the interpretation of predictions by the model. Namely, what does an inaccurate prediction mean—that the model is not powerful enough and was not able to recognize the author's and translator's styles, or, on the other hand, that the essential style of the author was "lost in translation"? Or even: if the model predicts author *B* in a translated work by author *A,* can we conclude from this that the translator was somehow 'influenced' by the works of author *B*? (however, in reality, this is just a special case of the original style getting lost in translation).

Another challenge is to find a significant volume of works by the same author and the same translator.

Datasets

https://www.unesco.org/xtrans/bsform.aspx

https://www.rochester.edu/College/translation/threepercent/translation-database/