A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models

Iustina Ilisei
Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom
iustina.ilisei(a)gmail.com

ABSTRACT
In the field of Descriptive Translation Studies, *translationese* refers to the specific traits that characterize the language used in translations. While translationese has been often investigated to illustrate that translational language is different from non-translational language, scholars have also proposed a set of hypotheses which may characterize such differences. In the quest for the validation of these hypotheses, embracing corpus-based techniques had a well-known impact in the domain, leading to several advances in the past twenty years. Despite extensive research, however, there are no universally recognized characteristics of translational language, nor universally recognized patterns likely to occur within translational language. This thesis addresses these issues, with a less used approach in the field of Descriptive Translation Studies, by investigating the nature of translational language from a machine learning perspective.

While the main focus is on analyzing translationese, this thesis also investigates two related sub-hypotheses: simplification and explicitation. To this end, a multilingual learning framework is designed and implemented for the identification of translational language. The framework is modeled as a categorization task, the learning techniques having the major goal to automatically learn to distinguish between translated and non-translated texts. The second and third major goals of this research are the retrieval of the recurring patterns that are revealed in the process of solving the task of categorization, as well as the ranking of the most influential characteristics used to accomplish the learning task. These aims are fulfilled by implementing a system that adopts the machine learning methodology proposed in this research.

The learning framework proves to be an adaptable multilingual framework for the investigation of the nature of translational language, as this dissertation illustrates by applying it to the investigation of two languages: Spanish and Romanian. This thesis experiments with different research scenarios and learning models in order to assess to what extent translated texts can be differentiated from non-translated texts in specific contexts. The findings show that machine learning algorithms, aggregating a large set of potentially discriminative characteristics for translational language, are able to differentiate translated texts from non-translated ones with high scores. The evaluation experiments report performance values such as accuracy, precision, recall and F-measure on two datasets.

The present research is situated at the confluence of the three areas of Descriptive Translation Studies, Machine Learning and Natural Language Processing, justifying the need to combine these fields for the investigation of translationese and translational hypotheses.

KEYWORDS: comparable corpora, explicitation, Romanian, simplification, Spanish, translationese.

Completion of Thesis
Place: University of Wolverhampton, UK
Year: 2013
Supervisors: Prof. R. Mitkov, Prof. G. Corpas, Prof. D. Inkpen.