Chapter 21
Machine Learning Based Taxonomy and Analysis of English Learners' Translation Errors

Ying Qin
Beijing Foreign Studies University, Beijing, China

ABSTRACT

This study extracts the comments from a large scale of Chinese EFL learners' translation corpus to study the taxonomy of translation errors. Two unsupervised machine learning approaches are used to obtain the computational evidences of translation error taxonomy. After manually revision, ten types of English to Chinese (E2C) and eight types Chinese to English (C2E) translation errors are finally confirmed. There probably exists three categories of top-level errors according to the hierarchical clustering results. In addition, three supervised learning methods are applied to automatically recognize the types of errors, among which the highest performance reaches F1 = 0.85 on E2C and F1 = 0.90 on C2E translation. Further comparison to the intuitive or theoretical studies on translation taxonomy shows some phenomenon accompanied by language skill improvement of Chinese learners. Analysis on translation problems based on machine learning provides the objective insight and understanding on the students' translations.

1. INTRODUCTION

Errors are unavoidable for EFL students during language learning. Error analysis plays an important role in language pedagogy by observing students' performances in real communication situations (Richards, 2015). And translation practice could reflect the primary ability of L2 learners as they process second language (Dodds, 1999, pp. 58-61). Compared to the errors in students' free essay writings, translation errors have some unique features because translation is an activity in a constrained language environment.

DOI: 10.4018/978-1-6684-6291-1.ch021

Copyright © 2022, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

When transforming the source text into the target, students have to consider the semantic equivalence of the two texts as well as the expression in target language. Therefore, translation error could reveal the default of language application when L2 learners try to use the word and grammar to convert the original text into the target language. Additionally, Séguinot (1990) pointed out that translation error analysis could be used to deeply explore the procedure of translation besides being used to judge the quality of translation. Thus, translation error analysis is of great significance to L2 teaching and language study (Yakimovskaya, 2012).

However, the taxonomy of translation error lacks commonly agreed distinctions probably due to that the causes of errors are very complicated. Pym (1992) divided translation error into binary and non-binary error to study the division between translation teaching and language teaching. Binary error refers to the error with no ambiguity and usually could be corrected with the right one. For example, errors of spelling, inflection, word selection and syntax are typical binary errors. On the other hand, inaccurate or unfaithful translation is regarded as non-binary error. Another translation error taxonomy is whether the error is content-related or language-related (Secară, 2005). The former will lead to semantic difference between the source and the target text, such as mistranslation and omission. Language-related error is not as serious as the content-related, generally not leading to misunderstanding. For example, ignorance of case, misuse of possessive and improper collocation is usually viewed as minor errors. More categories of translation errors are summarized by Corder (1974) including addition, selection, omission and ordering according to different translation methods. Even more categories are proposed by American Translators Association (ATA) with as many as 24 kinds of common translation problems.

Some scholars put forward hierarchical categories of translation errors. According to Richards (1971), the top level of error type is a three-kind framework: a) interference errors, generated by L1 transfer; b) intralingual errors, resulted from incorrect (incomplete or over-generalized) application of language rules and c) developmental errors, caused by the construction of faulty hypotheses in L2. Each type of error can be further divided into several sub-types.

Chinese scholars have studied the translation problems made by different learners as well. Sun (2011) described four common Chinese to English (C2E) translation errors made by second-year English majors. Chen (2012) applied the corpus-based approach to study the top 10 C2E errors in the national university entrance exams. Both of them adopt single-layer error taxonomy. Nevertheless, the taxonomy of translation errors is subjective, varying with different researchers or purposes.

In recent years empirical approach in translation error analysis is attracting more and more attention (Campoy et al., 2010). The authors lately build a Chinese learners' translation corpus with comments of teachers on translation problems. Based on the learners' corpus, the authors attempt to apply machine learning methods including clustering and classification to explore the taxonomy of translation errors. The goal and contribution of this study is to generalize translation errors in Chinese EFL learners in a more objective way than the previous studies, which might provide a hint to language learning and translation teaching.

In the following paper the authors first give the introduction to the learners' translation corpus lately built, followed by the description of machine learning approaches including clustering and classification algorithms employed in this study. Section 4 is the experimental results and translation error instances and analyses based on the corpus. The authors also compare the results to the related studies on translation error taxonomy in this section. The last section contains the conclusion and future work.

15 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/machine-learning-based-taxonomy-and-analysis-of-english-learners-translation-errors/307462

Related Content

Using Open-Source Software for Business, Urban, and Other Applications of Deep Neural Networks, Machine Learning, and Data Analytics Tools
Richard S. Segalland Vidhya Sankarasubbu (2022). *International Journal of Artificial Intelligence and Machine Learning (pp. 1-28).*
www.irma-international.org/article/using-open-source-software-for-business-urban-and-other-applications-of-deep-neural-networks-machine-learning-and-data-analytics-tools/307905

Designing a Real-Time Dashboard for Pandemic Management: COVID-19 Using Qlik Sense
Rahul Rai (2021). *Machine Learning and Data Analytics for Predicting, Managing, and Monitoring Disease (pp. 190-203).*
www.irma-international.org/chapter/designing-a-real-time-dashboard-for-pandemic-management/286252

A Method Based on a New Word Embedding Approach for Process Model Matching
Mostefai Abdelkaderand Mekour Mansour (2021). *International Journal of Artificial Intelligence and Machine Learning (pp. 1-14).*
www.irma-international.org/article/a-method-based-on-a-new-word-embedding-approach-for-process-model-matching/266492

An Overview of Biomedical Image Analysis From the Deep Learning Perspective
Shouvik Chakrabortyand Kalyani Mali (2020). *Applications of Advanced Machine Intelligence in Computer Vision and Object Recognition: Emerging Research and Opportunities (pp. 197-218).*
www.irma-international.org/chapter/an-overview-of-biomedical-image-analysis-from-the-deep-learning-perspective/252628

Development of a Charge Estimator for Piezoelectric Actuators: A Radial Basis Function Approach
Morteza Mohammadzaheri, Mohammadreza Emadi, Mojtaba Ghodsi, Issam M. Bahadur, Musaab Zarogand Ashraf Saleem (2020). *International Journal of Artificial Intelligence and Machine Learning (pp. 31-44).*
www.irma-international.org/article/development-of-a-charge-estimator-for-piezoelectric-actuators/249251