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 $F_1 = 0.85$ on E2C and $F_1 = 0.90$ 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.

KEYWORDS

Classification, Error Analysis, Flat Clustering, Hierarchical Clustering, Learners’ Translation Corpus, Machine Learning, Translation Error Taxonomy

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. 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

DOI: 10.4018/IJCALLT.2019070105

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
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.

2. LEARNERS’ TRANSLATION CORPUS WITH NATURAL COMMENTS

The translation texts the authors collected are written translation assignments of English major postgraduate students. Besides the sources and students’ translations, the authors collect the comments of the teachers. Due to the fact that translation errors may be attributed to numerous causes and located on numerous levels, the authors do not require the teachers to use pre-defined error types to markup the problems in the students’ translation assignments. Instead, they can make comments on the translation problems in their own words. Most teachers prefer to use natural sentences to describe translation problems. Therefore, the comments can fully reflect the opinion of the commenter on the translation problems. The length of comments may vary from a single word to several paragraphs.

The corpus contains original texts of 23 topics, with 15 C2E and 8 E2C topics respectively. It is also a balanced corpus including six fields such as technology, politics, literature, and law. Table 1 shows the field distribution of translations in the corpus. Totally the corpus consists of 1091 students’ translation texts over one million words. The number of commenters for C2E and E2C translations is 123 and 29 respectively. Most of them are teachers and some are their assistants, all with high translation competence. Therefore, the study of these comments may reflect the common viewpoints
Learning Objects: Projects, Potentials, and Pitfalls
[www.igi-global.com/chapter/learning-objects-projects-potentials-pitfalls/20024?camid=4v1a](www.igi-global.com/chapter/learning-objects-projects-potentials-pitfalls/20024?camid=4v1a)

Using Social Network-Mediated Bridging Activities to Develop Socio-Pragmatic Awareness in Elementary Korean