Abstract

OOV term translation plays an important role in natural language processing. Although many researchers in the past have endeavored to solve the OOV term translation problems, but existing approaches are not able to handle different types of OOV terms, especially hybrid translations, such as “Kenny-Caffey syndrome (Kenny-Caffey 氏症候群)”. We proposed a novel English definition ranking approach to consider the types of OOV terms before translating them. Thus, different types of OOV terms could be translated differently. Furthermore, the translations mined in other languages are also OOV terms, none of existing approaches offer the context information or definitions of the OOV terms. Users without special knowledge cannot easily understand meanings of the OOV terms. Our English definition ranking method also extracts multilingual context information and monolingual definitions of OOV terms. Moreover, non-existing methods focus on cross language definition retrieval for OOV terms. We propose a novel CLIR for Chinese definition retrieval method for extracting Chinese definitions of OOV terms. Never the less, it has always been so difficult to evaluate the correctness of an OOV term translation and definition without domain specific knowledge and correct references. We propose a novel auto re-evaluation method to evaluate the correctness of OOV translations and definitions.

We tested our methods with both name type and biomedical type OOV terms. We retrieved and processed a total of 743,914 documents (snippets). Our method achieved accuracies of 84.15% for multilingual context information extraction and 75.46% for English definition extraction respectively. Our method also achieved precision of 79.76% and high recall of 99.86% for name type OOV term prediction, and we achieved high precision of 99.93% and recall of 89.21% for biomedical type OOV term prediction. For name type OOV term translation, our method gained little improvements over existing methods with high accuracies of 98.39% and 98.39% in candidate generation and candidate selection respectively. For biomedical type OOV term translation, our method gained much improvements over existing methods, our method of SF+F+W+S+B+P with the base machine learning algorithm Lib-supported vector machine surpasses the existing methods with a recall of 83.05% and precision of 79.72% for OOV translation. Furthermore, our method achieved accuracies of 67.49% for Chinese definition extraction, 85.12% for name type OOV term Chinese definition extraction and 60.00% for biomedical type OOV term Chinese definition extraction. We achieved a precision of 46.99% and a recall of 99.37% for translation auto re-evaluation. We also achieved a precision of 67.90% and a recall of 99.41% for definition auto re-evaluation.