氏　　名	区　　建			
学位の種類	博士(情報科学)			
学位記番号	博情第 280 号			
学位授与年月日	平成 25 年 9 月 24 日			
論文題目	A study on Translation, information extraction and CLIR based on OOV term processing(OOV 用語処理に基づく翻訳,情報抽出,クロスランゲージ情報検索に関する研究)			
論文審査委員	主査	島津　明	北陸先端科学技術大学院大学	教授
		東条　敏	同	教授
		白井　清昭	同	准教授
		由井薗 隆也	同	准教授
		任　福継	徳島大学	教授

論文の内容の要旨

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.

論文審査の結果の要旨

本論文は、辞書にない語に対する自然言語処理の新技術を提案し、その効果を示している。社会の変化、科学技術の発展などにより新しい語が次々と生み出され、機械翻訳、情報検索などのシステムは、辞書がカバーしない語、out of vocabulary term（OOV 用語）に対処しなければならない。

このような OOV 用語について、本論文は、特に、生物医学用語および人名などの固有名の OOV 用語を対象に、新しい翻訳法や定義情報の抽出法などを示している。具体的には、OOV 用語の英語定義情報の抽出法、OOV 用語の中国語定義情報の抽出法、OOV 用語の文脈情報（クラス、属性など）の抽出法、OOV 用語の翻訳法、OOV 用語の中国語訳の正確さの自動評価法などを示している。

OOV 用語（英語）の定義情報は、Web を検索し、定義が存在する確度の高いスニペットから求められる。このために、スニペットにおける定義表現の特徴、検索結果の順位、Web ドメインの信頼性順位、名詞の曖昧性、Wiki ページの種類の優先度などの情報が利用される。OOV 用語の文脈情報は、定義情報、オントロジーなどに基づいて求められる。OOV 用語の中国語の定義情報は、言語横断検索により得られる。中国語では、語間にスペースがないこと、語の曖昧性などが考慮され、定義動詞、動詞との共起、動詞の位置などの情報が利用される。OOV 用語の定義情報抽出法などに関する論文は、国際会議 NLP-KE'12 において最優秀論文賞を受賞している。

生物医学分野の OOV 用語の翻訳では、従来法が対応しないハイブリッド型 OOV 用語の翻訳法を提案している。ハイブリッド型というのは、Kenny-Caffey syndrome を「Kenny-Caffey 氏症候群」と訳すような用語である。提案法は、Web から規則により訳の候補を求め、候補をフィルターにより絞り、機械学習により訳を選択する。新たな素性 11 種類を提案するとともに、4 種類の機械学習、パラメータ最適化などを網羅的に組合せて試している。固有名の OOV 用語の翻訳法は、訳候補を既存法により求め、フィルターを適用し、ランキング関数により中国語訳を選択する。訳の正確さについては、用語の文脈情報が定義候補に含まれる程度により自動評価する方法を提案している。

OOV 用語の翻訳実験の結果は、本論文の方法が、従来法を大きく上回っていることを示してい

る。

以上、本論文は、OOV 用語について新しい機械翻訳技術を示したものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。