氏　　名	Maike（マイケ） Erdmann（エルドマン）
博士の専攻分野の名称	博　士（情報科学）
学 位 記 番 号	第　２４６６４　号
学位授与年月日	平成 23 年 3 月 25 日
学位授与の要件	学位規則第 4 条第 1 項該当 情報科学研究科マルチメディア工学専攻
学 位 論 文 名	Extraction of Bilingual Terminology from Wikipedia （Wikipediaからの対訳辞書抽出）
論文審査委員	（主査） 教　授　西尾章治郎 （副査） 教　授　藤原　　融　　教　授　細田　　耕　　教　授　鷹田　憲久 准教授　原　　隆浩　　准教授　秋吉　政徳 情報通信研究機構グループリーダー　木俵　　豊

論 文 内 容 の 要 旨

With the demand for bilingual dictionaries covering domain-specific terminology, research in the field of automatic dictionary extraction has become popular. However, accuracy and coverage of dictionaries created from parallel corpora, the standard resource for bilingual dictionary extraction, faces several issues.

Therefore, in this thesis, we propose bilingual dictionary construction from Wikipedia, a large-scale multilingual encyclopedia. The extraction of bilingual terminology from Wikipedia enables us to overcome some of the issues of parallel corpus-based dictionary extraction.

As opposed to other research on bilingual dictionary extraction from Wikipedia, we analyze not only interlanguage links but also other kind of link information. Besides, we propose a method to filter out incorrect translations through supervised learning with features (characteristics) extracted from the link structure of Wikipedia.

We analyze the article texts of Wikipedia as well, in order to further improve bilingual dictionary extraction. We propose a novel algorithm to calculate the similarity of Wikipedia articles in different languages, with the aim of identifying incorrect and missing interlanguage links in Wikipedia. After that, we propose the usage of Web search engine results to find translations for terms not described by a Wikipedia article, using a method of translation prediction more sophisticated than any other research before.

The thesis consists of five chapters. In Chapter 1, we explain the background and motivation of our research. After that, we describe the contribution as well as the structure of the thesis.

In Chapter 2, we propose the extraction of bilingual terminology from Wikipedia link information, and prove that our approach performs significantly better than a bilingual dictionary extracted from a parallel corpus as well as a manually created dictionary.

In Chapter 3, we improve the filtering of incorrect term-translation pairs through supervised learning, allowing us to take into account many different features of a term-translation pair for

evaluating its correctness, and show the effectiveness of that approach, especially when using a large number of features.

In Chapter 4, we propose two different approaches of analyzing Wikipedia article texts to improve the coverage and accuracy of extracted bilingual dictionaries: Wikipedia article similarity calculation and the usage of Web search engine results.

Finally, we draw a conclusion of our research and present plans for future work in Chapter 5.

論文審査の結果の要旨

異なる言語間で単語の翻訳関係を定義した対訳辞書は、機械翻訳や言語横断検索など幅広い分野で、基盤技術として必要とされている。そのため、対訳辞書の自動構築に関する研究が活発に行われてきたが、従来研究では精度と網羅性を両立させた対訳辞書の構築が技術的課題であった。一方、ブロードバンドなどの高速通信インフラの普及に伴い、ユーザが協調してコンテンツを作り上げるソーシャルメディアが爆発的に普及した。インターネット上で構築されたソーシャルメディアを解析し、有用な情報を抽出することで、精度と網羅性の高い対訳辞書を構築できる。

本論文では、大規模なソーシャルメディアの一種であるオンライン百科事典のWikipediaとWebをマイニングすることで、精度と網羅性の高い対訳辞書の構築を目的としている。その主要な成果を要約すると以下の通りである。

(1) Wikipediaに含まれる多様な情報を解析して、対訳辞書を構築する手法を提案している。本手法では、異なる言語で記述された単語の間の翻訳関係を示す「言語リンク」以外に、リンク情報を統合的に利用することで精度と網羅性を向上させている。
(2) 機械翻訳で一般的に利用される編集距離などに加え、特殊リンクの有無、被リンク数、リンク数の比率といった多様な情報を複合的に利用して機械学習手法であるSVMを用いて対訳辞書構築の精度を向上させる手法を提案している。この手法は、特に対訳辞書の自動構築の際に含まれる誤訳などのノイズ情報を除去することを主目的としており、実験によりその有効性を証明している。
(3) 従来研究では言語リンクが存在しない場合に網羅性が低下するという問題があった。そのため、本研究では言語リンクが存在しないページ間の「隠れリンク」を発見する手法を提案している。本手法では、二つの異なる言語で記述されたページ間の類似度を各種の指標によって評価し、隠れリンクに有用な指標を検証している。さらに、Webマイニングの技術と融合することで、精度と網羅性の向上させている。

以上のように、本論文は、Wikipediaを用いた対訳辞書の構築に関する先駆的な研究として、情報科学に寄与するところが大きい。よって本論文は博士（情報科学）の学位論文として価値のあるものと認める。