論文内容の要旨

博士論文題目　Extracting Translation Knowledge from Parallel Corpora
(対訳コーパスからの翻訳知識抽出)

氏名　　山本 薫

(論文内容の要旨)

本論文では、対訳コーパスからの翻訳知識抽出について研究を行った。目的は、機械翻訳や翻訳支援などに使える翻訳知識を獲得することにある。従来研究では、単語対応の研究が多く報告されているが、翻訳は単語対応だけでは限界がある。そこで、本研究では、特に、複数から成る対訳表現を抽出することを目的とした。このような翻訳知識は、機械翻訳の辞書作成支援のみならず、翻訳者や言語学習者にとって有用な資源と考えられる。この研究目的を達成するために、以下の3つの研究報告をした。

はじめに、統計的な係り受け関係を利用した対訳表現抽出を報告した。この研究では、語順制約に縛られない係り受け関係を利用することにより、日本語と英語の対訳コーパスから語レベルのみならず句レベルの対応の抽出を目指した。実験を行ない、統計的な係り受け処理から得られた句レベルの係り受け関係は、日本語と英語のように言語制約が異なる対訳でも有効に働くことを確認した。

次に、対訳表現抽出における翻訳単位の比較検討を行なった。自然言語処理ツールから得られる、単語区切り、文節区切り、単語の依存関係という三つの異なる言語的手がかりを元に翻訳単位を生成し、それぞれが対訳表現抽出にどのように寄与するかを実験した。結果、文節区切りは、固有名詞などの連続的に出現する複数語の対訳抽出に有効に働くことが確認された。このことにより、新分野の専門用語辞書の構築に有用であると考えられる。さらに、単語の依存関係を用いた場合、慣用表現や訳し分け知識などを獲得するのに有効に働くことが確認された。

最後に、データマイニング手法による対訳表現抽出を報告した。対訳表現抽出をデータマイニング分野で研究されている系列データマイニング問題とみなし適用する。対訳文から対訳系列を生成し、系列データマイニングを適用することにより、連続的と非連続的に共起する翻訳単位の生成と数えあげを効率的に実現した。系列データマイニングを効率的に解くPrefixSpanアルゴリズムを使って実験を行ない、連続的と非連続的に出現する複数語からなる対訳表現を抽出した。

氏 名	山本 薫

(論文審査結果の要旨)

平成13年12月25日に開催した公聴会の結果を参考に平成14年2月12日に本博士論文の審査を行った。以下のとおり、本博士論文は、提案者が独立した研究者として、研究活動を続けていくための十分な素養を備えていることを示すものと認める。

山本 薫は、本博士論文において、対訳データから機械翻訳に関する知識を自動的を行う方法を提案した。提案内容は次のような項目に渡り、それぞれの観点から抽出手法の性能評価を行い、効果を明らかにした。

1. 語順制約に縛られない係り受け関係を利用することにより、日本語と英語の対訳コーパスから語レベルのみならず句レベルの対応の抽出を試みた。実験を行ない、統計的な係り受け処理から得られた句レベルの係り受け関係は、日本語と英語のように言語制約が異なる対訳でも有効に働くことを確認した。

2. 対訳表現抽出における翻訳単位の比較検討を行なった。自然言語処理ツールから得られる、単語区切り、文節区切り、単語の依存関係という三つの異なる言語的手がかりを元に翻訳単位を生成し、それぞれが対訳表現抽出にどのように寄与するかを実験した。結果、文節区切りは、固有名詞などの連続的に出現する複数語の対訳抽出に有効に働くことが確認された。

3. 対訳表現抽出をデータマイニング分野で研究されている系列データマイニング問題として定式化した。対訳文から対訳系列を生成し、系列データマイニングを適用することにより、連続的と非連続的に共起する翻訳単位の生成と数えあげを効率的に実現した。

このように、対訳データに対して様々な視点と制約によって対訳表現の抽出法を提案し、実験によってその効果を実証した本研究は、独創性高く、しかも実用的であり、自然言語処理の分野において高い貢献があると評価する。

よって、本論文は博士(工学)の学位論文として価値あるものと認める。