

論文の内容の要旨

論文題目 **Formalization and Realization of Example-based Machine Translation**
(用例ベース翻訳の定式化と実現)

氏 名 荒牧 英治

機械翻訳は、自然言語処理の最大の応用分野であるとともにキラーアプリケーションであり、その研究は50年も前から開始され、現在に至るまで続けられてきた。

当初はルールベースの翻訳システムが研究されていた。しかし、ルールを手で作成するには、大きなコストと時間がかかる上、ルールが増すとともにルールを管理するのが困難となる。

近年では、インターネットの発展とともに電子化テキストの量が増加しつづけ、それら大量のコーパスを用いたコーパスベース翻訳に注目が集まっている。コーパスベース翻訳は統計ベース翻訳[Nagao1984]と用例ベース翻訳[Brown1993]の2つのアプローチがある。

統計ベース翻訳は、単語にばらして翻訳モデルを学習する。この方式は、英語＝フランス語など類似した言語間ではうまく働くが、日英のように大きな構造変換が行われる言語間の翻訳をとらえるのが困難である。

本論文は用例ベース翻訳に焦点をあてる。用例ベース翻訳は、入力文の各部分に対して類似した用例を選択し、それらを組み合わせて翻訳を行う。経験豊かな人間が翻訳を行う場合でも用例を利用して翻訳を行っており、この方式は他の手法よりも自然な翻訳文の生成が可能だと考えられる。また、用例の追加により容易にシステムを改善可能である。

以上のような利点を持つものの、一般的な用例ベース翻訳システムは実現されていない。

その理由の一つとして、用例ベース翻訳は入力文や用例を解析するための高精度のパarserを必要とするが、これまでは十分な精度を持つparserが存在しなかったことが挙げられる。しかし、近年、parserの精度は急速に向上し、日本語や英語などではすでに90%以上の精度を持つ。すなわち、最近になってはじめて、実証的な研究が行えるベースが整ったと言える。

本稿は実証的な用例ベース翻訳を実現するために(1)頑健なアライメント、(2)訳語選択に関して提案する。さらに、従来まで経験則に頼っていた用例ベース翻訳を定式化する

二章ではアライメントについて述べる。言語間の対応関係を推定する処理であるアライメントは機械翻訳のキーテクノロジーである。特に、アライメント結果を直接用いる用例ベース翻訳にとって、アライメントは翻訳精度に大きな影響を及ぼす重要な処理である。提案手法は、頑健なアライメントのために、parserで解析された構造情報を利用し、翻訳辞書など既存の言語資源をフルに用いる。翻訳辞書を用いたアライメントでは、(a)辞書対

応の曖昧性, (b)辞書に載っていない語の扱いが問題となる.

この問題を解決するために, (a)に対しては, 対応の整合性を用いる対応内の整合性, 対応同士 (対応間) の整合性を提案する. (b)に対しては, 構造の情報を利用する.

三章では用例の選択を中心に翻訳システム全体について述べる. アライメント結果にはしばしばエラーが含まれる. このため, 大量の用例の中から適切なものを選ぶ技術が必要となる.

提案手法は, 従来から用いられている用例の大きさ, 類似度に加え日英対応の確信度を用いこの問題を解決する.

四章では, 用例ベース翻訳を定式化を行う. 大きな用例ほど大きなコンテキストを扱えるため, 用例ベース翻訳は基本的に大きな用例を用いて翻訳を行う. すなわち, 用例ベース翻訳は大きさを重視, 統計ベース翻訳は頻度を重視するアプローチだと言える. しかし, 統計ベース翻訳とちがい用例ベース翻訳は定式化されておらず, 統計ベース翻訳と比べてアドホックである.

提案手法は, 大きな用例を好む翻訳確率を提案し, それを用いて用例ベース翻訳を定式化する.

五章では, 提案した翻訳システムの評価を行う. 評価は複数の自動評価手法と人手による主観評価の両方からなり, 提案した翻訳システムおよびその手法の妥当性を証明する.

六章では結論を述べ, 今度の機械翻訳研究の方向性を議論する.