4F3-2

専門辞書拡張型機械翻訳システムにおける中間コード選択

Intermediate code selection in machine translation system extensible for domain-specific dictionaries

関西学院大学理工学部情報科学科 ○ 増田 雄介 ， 北村 泰彦

Yusuke Masuda and Yasuhiko Kitamura

School of Science and Technology, Kwansei Gakuin University

Abstract A number of machine translation systems are available on the web, but translating technical terms is not easy. We developed a machine translation system extensible for domain-specific dictionaries. It substitutes an intermediate code for each technical term in an input sentence, translates the substituted sentence, and substitutes again the parallel translation in domain-specific dictionaries for the intermediate code. However, the quality of translation is low when we use intermediate codes that do not reflect the meaning of the corresponding technical terms. We therefore developed a method to select intermediate codes that reflect the meaning of technical terms and improved the quality of translation.

1 はじめに

現在，Web 上の機械翻訳ツールにより誰でも手軽に翻訳することが可能になっている．しかし，既存の機械翻訳ツールには専門用語の対訳が難しいという問題がある [1]．例えば,「ワンセグというサービスが人気です.」を翻訳すると,「Service named Wanseg is popular.」となり,「ワンセグ」という単語が適切に翻訳されない．

そこで，専門辞書を機械翻訳の辞書とは別に作成し，機械翻訳ツールと連携させることで，専門用語の翻訳を可能にする専門辞書拡張型機械翻訳システムを開発した．このシステムでは，文中の専門用語を一旦中間コードに置換して翻訳し，その中間コードを専門辞書に載っている対訳で置換する．例えば,「ワンセグというサービスが人気です.」という文を入力すると，文中の専門用語「ワンセグ」に中間コード「@1」を置換し,「@1 というサービスが人気です」という文になる．これを英語に翻訳すると「Service named @1 is popular.」となる．最後に，中間コードに,「ワンセグ」の対訳である「1seg」を置換して「Service named 1seg is popular.」となる．

しかし，中間コードに無意味な記号を用いることで単語の地名，事象といった属性の情報が失われ，周辺の前置詞や冠詞が適切に翻訳されない場合がある．例えば，入力文「ワンセグで見たアルファードが欲しい.」を翻訳すると，出力文は「I want ALPHARD seen with 1seg.」となる．しかし，専門用語である「1seg」の前置詞が「on」になるべきところ「with」となっており，適切に翻訳できない．

そこで，本研究では中間コードとして，置換する対象の専門用語と同じ意味のカテゴリに属し，機械翻訳が扱える一般的な単語を用いる手法を提案し，翻訳の質の向上を図る．

2 中間コードを用いた機械翻訳

専門用語の属性の情報を持った単語を中間コードにするため，まず専門辞書内の専門用語をカテゴリ別に分類する．中間コードとして，それぞれのカテゴリに一致し，かつ機械翻訳ツールが扱える単語を用意する．

表 1: 中間コード例

人		製品		施設	
日	英	日	英	日	英
太郎	Taro	テレビ	television	ビル	building
花子	Hanako	ラジオ	radio	家	house
次郎	Jiro	カメラ	camera	学校	school

表 1 に示すのは，カテゴリごとに分けられた中間コード例である．一番上の行はカテゴリ名であり，その下に中間コードの対訳を示す．カテゴリの分類は，代表的なシソーラスのひとつである日本語語彙大系の単語意味体系に基づいている [2]．中間コードが複数存在する理由は，文中に同じカテゴリの専門用語が複数個出てきた場合，中間コードが重ならないようにするためである．

専門用語の属性の情報を持った中間コードを用いる専門辞書拡張型機械翻訳システムの翻訳手順を説明する．

1. 形態素解析

   翻訳入力文を形態素解析器を用いて分かち書きする．

2. 専門辞書検索

形態素解析器で抽出された名詞・未知語を，対訳とカテゴリ名から成る専門辞書と照らし合わせ，専門用語を見つける．

3. 中間コード置換

   専門用語を同じカテゴリの中間コードで置換する．

4. 機械翻訳ツールを用いた翻訳

   中間コード交じりの文を機械翻訳ツールで翻訳する．翻訳には機械翻訳ツールのひとつである Excite 翻訳を利用する．

5. 専門用語置換

   翻訳して得られた翻訳文中の中間コードを専門辞書に載っている専門用語の対訳に置換する．

例えば,「ワンセグで見たアルファードが欲しい.」という文では，文中の専門用語「ワンセグ」,「アルファード」は，カテゴリ「製品」の中間コードとして登録されている「テレビ」,「ラジオ」に置換され,「テレビで見たラジオが欲しい.」という文になる．この文を英語に翻訳すると「I want the radio seen on the television.」となる．ここで，中間コードである「television」,「radio」を専門用語「ワンセグ」,「アルファード」の対訳である「1seg」,「ALPHARD」にそれぞれ置換して「I want the ALPHARD seen on the 1seg.」となる．

表 2: 各システムの生成した翻訳文比較

Excite 翻訳	I want Alfard seen with Wanseg.
意味なし中間コード	I want ALPHARD seen with 1seg.
意味あり中間コード	I want the ALPHARD seen on the 1seg.

表 2 に示すのは「ワンセグで見たアルファードが欲しい.」という文を Excite 翻訳，意味なし中間コードを用いたシステム，意味あり中間コードを用いたシステムで生成した翻訳文の比較である．意味あり中間コードを用いたシステムでは，専門用語だけでなく前置詞「on」や冠詞「the」といった専門用語周辺の情報が翻訳文に反映され，翻訳の質が向上している．

3 評価

「Web2.0:次世代ソフトウェアのデザインパターンとビジネスモデル（前編）」(http://japan.cnet.com/column/web20/story/0,2000055933,20090039,00.htm) に記載されている日本語で，専門用語を含んだ文章を用いて翻訳品質の評価を行った．これらの文章を Excite 翻訳，意味なし中間コードを用いたシステム，意味あり中間コードを用いたシステムの 3 種類のシステムで翻訳した．各システムが生成した翻訳文を入力文の原文である「O Reilly - What Is Web2.0」(http://oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html) に記載されている英文と照合し，以下の条件を一つ満たす度に得点を 1 点加え，得点の合計でシステムの評価を行う．

α . 専門用語が適切に翻訳されている．

β . 条件 α を満たし，かつ前置詞，冠詞が適切に翻訳されている．

以下の表 3 は 3 種類のシステムが生成した翻訳文を用いた評価結果である．翻訳文は 85 文で，得点の最大値は文章数の 2 倍の 170 点である．システムの得点を最大値で割ったものを条件達成率とする．

表 3: 翻訳文の得点によるシステムの評価

システム	Excite 翻訳	意味なし中間コード	意味あり中間コード
条件 α	41	82	84
条件 β	33	43	80
計 (条件達成率)	74(43.5%)	125(73.5%)	164(96.5%)

結果，意味あり中間コードを用いたシステムでは，Excite 翻訳と比較して約 53%，意味なし中間コードを用いたシステムと比較して約 23%，専門用語と前置詞，冠詞が適切に翻訳された文を生成する割合が増加しており，翻訳品質の向上を確認した．

4 まとめ

本研究では，専門辞書拡張型機械翻訳システムの中間コードに，置換する専門用語と同じカテゴリの単語を用いることで，専門用語とその前置詞，冠詞が適切に翻訳された文を生成し，翻訳の質を向上させた．

しかし，例えば「ブックマークする」のように「名詞＋自立動詞」の組合わせの場合，名詞の部分に中間コードを置換するだけでは適切に翻訳出来ない場合があった．このような問題を含めた翻訳の質の向上を今後の課題とする．

参考文献

[1] Toru Ishida. Language Grid: An Infrastructure for Intercultural Collaboration. In IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06), pp.96-100, 2006.

[2] 池原他（編）：日本語語彙大系，岩波書店，1997.