6C-01

翻訳品質に基づいた専門用語の半自動抽出手法の提案

園尾　聡　　田中　浩之　　木下　聡
(株)東芝　研究開発センター　知識メディアラボラトリー

1. はじめに

近年、多言語での円滑な情報伝達を実現するため、機械翻訳システムの研究開発が盛んに行われている。辞書や文法に基づく知識ベースの翻訳システムにおいて、対象ドメインに適応した高品質な機械翻訳を実現するためには、そのドメインに応じた専門用語辞書の開発が重要な課題となる。これに対し、対象ドメインコーパスにおける単語の出現頻度及び連接頻度に基づいて、専門用語の自動抽出を行う手法が提案されている[1]。

しかしながら、頻度ベースで抽出した専門用語の中には、構成単語が機械翻訳システムの辞書中に存在し、組み合わせによって適切な翻訳結果が得られるため、辞書登録作業が不要な単語も含まれる。一方で、機械翻訳システムの翻訳品質向上のため辞書登録が必要な単語が優先的に抽出されないという課題があった。

本稿では、機械翻訳システムの翻訳品質向上を目的とした専門用語の半自動抽出手法を提案する。提案手法では、単語の統計量に加えて、機械翻訳処理で得られる解析・変換情報を特徴量とすることで、翻訳品質の視点から辞書登録が必要な単語候補を抽出する。さらに、対象ドメインにおける訳語推定を組み合わせることで、辞書開発作業の効率化を実現する。

2. 提案手法

提案手法の処理フローを図１に示す。まず、対象ドメインコーパスの部分テキストを機械翻訳し、翻訳品質の悪化要因となっている誤訳単語を選定する。これを教師データとし、機械学習によって、辞書登録の優先度を推定する。機械学習には、出現頻度や文字種などの単語に関する特徴量(f_{word})と、その単語を含む文を翻訳した際に得られる構文解析、単語アライメントなどの機械翻訳に関する特徴量(f_{mt})を採用した。表１に今回用いた特徴量を示す。

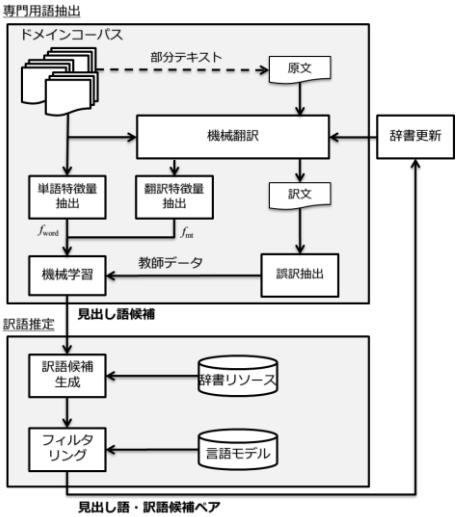

図 1: 提案手法の処理フロー

これらの特徴量を用いて、教師データに含まれる単語の対象ドメインコーパスにおける頻度カバー率を最大化するように、特徴量に対する重み付けパラメータを最適化する。最適化には、多変量のノンパラメトリック法のひとつである Powell 法[2]を用いた。

次に、辞書開発作業を効率化するため、抽出された専門用語に対して訳語候補を推定する。訳語候補推定は、専門用語を構成する各単語について、既存辞書リソースを用いて訳語候補を取得し、それらの訳語候補を組み合わせることで新たな訳語候補を生成する要素合成法[3]を用いた。単純に訳語候補を組み合わせた場合、候補数が膨大となってしまうため、対象ドメインの言語モデルを用いて訳語候補のフィルタリングを行った。最終的に出力された見出し語-訳語候補ペアを元に、専門用語辞書の開発を行った。

3. 評価実験

日-中機械翻訳システム向けの専門用語辞書開発を想定し、評価実験を行った。対象ドメインコーパスとして、日本語特許明細書(IT 分野、32 文書、337,864 字)を用いた。そこから機械翻訳システムに辞書登録が必要な 453 語の専門用語を人手で選定し、教師データおよび評価用の正解データとした。なお、評価用データには全ての教師データ（クローズドセット）を用いた。

A semi-supervised technical words extraction based on translation quality.
Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation.

表 1: 単語および翻訳に関する特徴量

単語に関する特徴量	出現頻度（ドメインコーパス）
	出現頻度（一般コーパス）
	n-gramスコア(一般コーパス)
	形態素数（複合語の場合）
	未知語を含む単語かどうか
	英数字を含む単語かどうか
	英数字のみの単語かどうか
	TF(Term Factor)
	IDF(Inverted Document Factor)
	TF-IDF
	連接頻度([1])
翻訳に関する特徴量	構文解析に失敗した文に含まれる単語かどうか
	翻訳辞書中の単語に対する訳語候補の異なり数
	翻訳辞書中の単語に対する訳語候補の分散
	原文-訳文間で単語アライメントがとれている単語かどうか
	原文-訳文間の単語アライメントで交差している単語かどうか
	原文-訳文間の単語アライメントで交差していない単語がどうか

表 2: 抽出された見出し語候補（抜粋）

見出し語	訳語候補	提案手法による優先度(順位)	頻度による優先度(順位)
光ケーブル	光缆,光纤电缆,光电缆,光纤缆,感光缆	15	107
手段	模块,手段,办法,用层,用排	16	13
アドレス帳	地址簿	17	98
注文	要求,订购,定购,订货,定货	18	82
ストレージ	存储,储存,存储介质,储藏,保管	19	43
アプリケーション	应用,申请,应用程序,应用软件	20	26

優先度上位 1000 語の抽出精度を評価したところ、88 語の辞書登録が必要な見出し語が抽出され、F 値は 12.2% (Prec.=8.8%, Rec.=20.1%)であった。一方で、抽出された単語の文書全体の頻度カバー率は 82.2%であり、高頻度な単語を優先的に抽出できていることが確認できた。

表 2 に提案手法によって抽出された見出し語および訳語候補の抜粋を示す。今回の評価実験では、高頻度な単語が評価用データに含まれていたため、最終的な抽出精度に関しては頻度ベースの手法と同等であった。しかしながら、提案手法による優先度上位を見ると、低頻度な単語であっても優先的に抽出できていることが分かる。

次に、以下の条件下で辞書開発作業を行い、提案手法における作業時間および翻訳システムの翻訳品質改善効果を測定した。

1. ランダムに選択した翻訳結果を提示し、辞書登録が必要な語を抽出し、専門用語辞書として登録（ベースライン）
2. 提案手法によって提示された見出し語及び訳語候補を確認し、専門用語辞書として登録

両作業について、日中バイリンガルが同じ時間で作業した。図 2 に、両作業において単位時間あたりに確認された文数および辞書登録された単語数を示す。ベースラインと比較して、提案手法では、確認文数が 2.8 倍、登録単語数が 1.7 倍となり、作業効率が大きく向上していることが確認できた。また、同一ドメインのオープンテスト文(5,000 文)を用いて、各専門用語辞書を用いた場合の翻訳品質の評価結果を図 3 に示す。提案手法では、翻訳品質の改善効果が高い見出し語が優先的に抽出され、誤訳の改善箇所が 1.5 倍となり、辞書開発の作業効率が向上した。

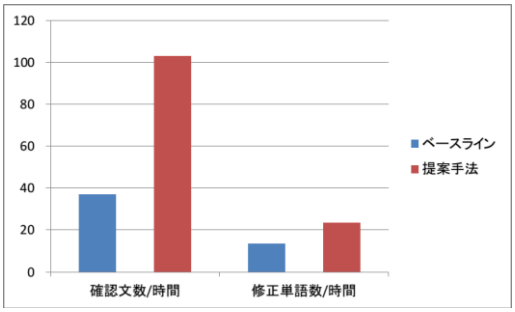

図 2: 辞書登録作業効率の改善効果

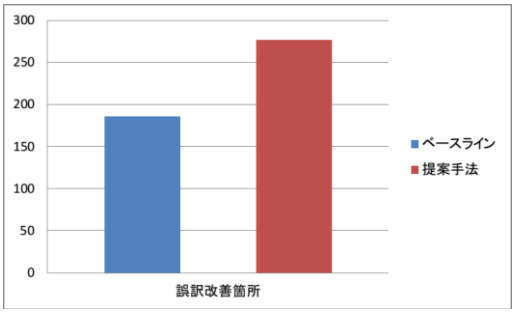

図 3: 翻訳品質の改善効果

4. おわりに

本研究では、機械翻訳システムの翻訳品質に基づいた専門用語の半自動抽出手法を提案した。特許明細書を用いた評価実験において、専門用語辞書の開発効率が 1.5 倍に改善する効果を確認した。

参考文献

[1] 中川裕志、森辰則、湯本紘彰: ″出現頻度と連接頻度に基づく専門用語抽出″, 自然言語処理, Vol.10, No.1, pp.27-45, 2003.

[2] Powell M.J.D. “An efficient method for finding the minimum of a function of several variables without calculating derivatives”, Computer Journal, Vol.7, No.2, pp.155-162, 1964.

[3] 外池昌嗣、宇津呂武仁、佐藤理史: “ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定”, 自然言語処理, Vol.14, No.2, pp.33-68, 2007.