

# 日中対訳コーパスを用いた専門用語辞典の自動構築

C0119082 神木 千澄

## 1. はじめに

中国語を母国語する人が専門的な日本語を用いて修学および就労している人数は、いずれも他言語と比較して最も多い。しかし日中間での専門用語辞典は非常に数が少ない。例えばIT辞典の場合、CJKI（日中韓辞典研究所）から出版されとされているものしか確認できなかった。また、収録語が和英で26万件に対し、日中では6.8万件と数が落ちるため、IT分野において確立された日中の対訳辞典は存在しないと考えている。

そこで本研究では日本語と中国語の対訳論文を用いて、対応づいた専門用語を獲得するシステムを構築し、アプリケーションとして公開する。

## 2. 関連研究

### 2.1 専門用語の抽出に関する研究

Shimohata[1]らは、隣接する単語のばらつき度合いによる単語列のユニット性を測る手法を提案。

### 2.2 専門用語の訳語推定に関する研究

パラレルコーパスを用い、2言語間で候補同士の同時出現確率を計算し、その値が高いものを翻訳対として抽出する手法が一般的に行われている。Kupiec[2]の手法では、80~90%の高い精度を達成している。

### 2.3 日中間における専門用語獲得の研究

Longら[3]は、SVMを用いて専門用語対訳対の同義・意義関係の判定を行う手法を提案し、約90%の適合率を達成している。

以上の知見を基に、本研究では日中対訳単語対の獲得に上記の手法を組み合わせることで解決を目指す。

## 3. 日中間対訳単語対の抽出システム

本研究のシステムの概要は右上の図1の通りである。まずコーパスから日本語と中国語をそれぞれ形態素解析し名詞句のみを抽出する。その後2.1の手法を用いて専門用語を抽出する。そして2.2及び2.3の手法を用いることで訳語推定を行い、専門用語対を獲得する。

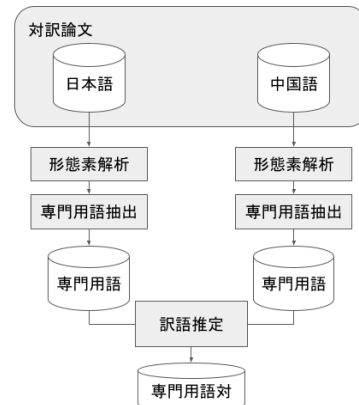


図1. 本研究で構築するシステムの概要図

## 4. 研究計画

表1 研究計画

|           | 8月 | 9月 | 10月 | 11月 | 12月 | 1月 | 2月 |
|-----------|----|----|-----|-----|-----|----|----|
| アルゴリズムの作成 | ■  | ■  |     |     |     |    |    |
| アプリケーション化 |    |    | ■   | ■   |     |    |    |
| 卒業論文作成    |    |    |     |     | ■   |    |    |
| 卒業論文提出    |    |    |     |     |     | ■  |    |
| 卒業論文最終発表  |    |    |     |     |     |    | ■  |

## 5. 進捗状況

現状、日中対訳単語対を獲得するシステムは完成している。プログラムはPython3言語を用い、テキストデータセットは、アジア学術論文抜粋コーパス<sup>1</sup>を用いた。今後はより高精度で高速に実行できる手法がないか検討し改良をける。

## 6. おわりに

日中対訳専門用語辞典の重要性、そのコーパスに基づく自動生成手順の概要等について述べた。

## 参考文献

- [1] 下畑 “共起頻度と語順制約を利用した分野依存性の高い定型表現の自動抽出”, 言語処理学会第3回年次大会, 1997.
- [2] Julian Kupiec “An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora.” In 31<sup>st</sup> Annual Meeting of the Association for Computational Linguistics, pp. 17–22, Association for Computational Linguistics, 1993.
- [3] 龍梓, 董麗娟, 宇津呂武仁, 三橋朋晴, 山本雄, “日中対訳文を用いた同義対訳専門用語の同手法”, 情報処理学会論文誌, 1882-776, 情報処理学会, Vol.56, No.3, pp.960-971, 2015.

<sup>1</sup> <https://jipsti.jst.go.jp/aspect/>