4R-3

語のグループ化を用いた特許文動詞の自動訳し分け

高野　雄一†　　横山　晶一†

山形大学大学院 理工学研究科‡

1. はじめに

近年,特許のような知的財産が,社会における貴重な存在として認識されており,これに伴う特許申請数の増加が著しい.また,国際的な特許の共有化に伴い国際特許も増加中にあり,正確で迅速な機械翻訳が求められている.

日英機械翻訳における訳文品質の分析[1]において,訳文品質低下の原因は訳し分けの不適切さであると報告されている.訳し分けとは,ある文中の単語を翻訳するときに訳の候補が複数ある場合,その文に最も適した訳を選択するということである.例えば,「含む」という動詞は,「全体の一部として含む」意味合いの文では"include",「要素・成分として含む」という意味合いでの文では"contain"と訳される.この訳し分けの精度を向上させるためには,文中で使用された単語の意味（語義）を解析する必要がある.

本研究では単語の意味解釈をした上での訳し分けのために,文章を意味のつながりで示すことの可能な語のグループ化を行う[2].語のグループ化とは“「男」「少女」を＜人＞と分類,「荷物」「鞄」を＜具体物＞と分類にする”などと,語を分類付ける方法のことをここでは言う.語のグループ化が訳し分けに役立つかどうかを調べ,従来よりも精度の高い訳し分けが可能なシステムを作成する.

2. 提案手法

適切な訳し分けを行うためには,文の意味を考慮する必要がある.しかし,詳細説明や要件が長大で難解であるという特許文の特徴から,精度の高い機械翻訳が困難であるという現状がある.

本研究では動詞の訳し分けを改善することによる翻訳精度の向上を目標とする.入力テキストから冗長な部分を排除し,動詞の前後のテキストから適切な訳し分けとなる対訳動詞を抽出し,修正する.方法を図 1 に示す.入力テキストから,対象の動詞と動詞前後のテキストを抜き出す.抜き出したテキストをグループ辞書により置換する.置換したテキストを訳し分け辞書のスコアに従って,訳し分け候補動詞のスコアを計算し,スコアが最も高くなった動詞を対訳として出力する.

Translation Disambiguatioin of Verbs in Patent Sentences using Word Grouping

† Yuichi Takano, Shoichi Yokoyama

‡ Graduate School of Science and Engineering(Informatics), Yamagata University

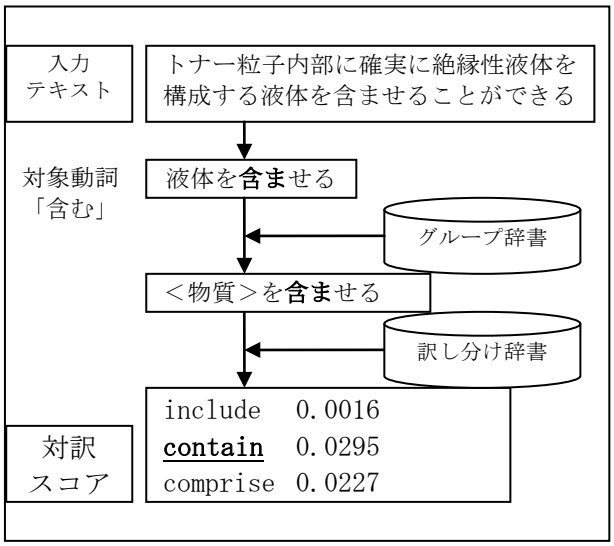

図 1:対訳動詞算出

表 1:動詞訳し分け辞書

動詞	出現数	スコア	テキスト
含む	1071	0.00305	＜道具＞に含まれる
(include)	621	0.00177	＜機械＞に含まれる
	602	0.00171	＜情報＞に含まれる
含む	542	0.00299	＜道具＞に含まれる
(contain)	363	0.00201	＜動物＞に含まれる
	332	0.00183	＜食料＞に含まれる

2.1 グループ辞書

グループ辞書とは同一の意味・概念となる語を一つにまとめた辞書である.例えば,「男性」,「女性」,「子供」などの語を＜人間＞というグループにまとめる.語を一つにすることで,意味的に同様である文を同様のものとして扱うことが可能となる.

2.2 動詞訳し分け辞書

訳し分け辞書の作成には対訳付き特許テキストデータとグループ辞書を用いる.

テキストデータから対象の動詞を含めた係り受けや N-gram をとる部分を抜き出す.抜き出した文を形態素解析し,グループ辞書を用いて各語を置換する.置換した文と,対象の動詞,対訳となる動詞を 1 つにまとめ,訳し分け辞書に追加する.訳し分け辞書では出現数を数えておき,出現数に応じてそのテキスト形でのスコアを決定する.現

在,スコアの算出方法として,テキスト出現数と,対訳動詞においてのテキスト総数の商をスコアとして扱っている.訳し分け辞書の作成例を表 1 に示す.

2.3 訳し分け評価

作成した訳し分け辞書を用いて,入力されたテキスト中にある動詞の訳し分け判定を行う.判定ではテキストの置換まで,訳し分け辞書作成と同様の処理を行う.置換を行った後,訳し分け辞書でのスコアを用いて,各対訳英語動詞でのスコアを算出する.スコアの合計が最も高くなった動詞を正しい対訳動詞として出力する.

3. 実験

本研究のシステムを利用し,日本語テキストを Google 機械翻訳(http://translate.google.co.jp/)に通し,翻訳結果の修正をする実験を行った.

3.1 実験設定

実験の流れを図２に示す。今回訳し分け辞書の作成に用いる学習データとしては,特許明細書文アラインメント[3]の日英対訳テキストからランダムに 500 万文を抽出し用いた.この特許文は日本から米国へ出願された対応特許の明細書の文を NICT の Align で対応付けたものであり,日本語文とそれを人手で英訳した文が収録されている.

グループ辞書には日本語語彙大系[4]を使用する.置換の際には語の上位語に変換する. 基本的には第６層目に変換し,置換元の語がそれ以下の層に該当する語の場合はそのままの形で用いる.

訳し分けの対象として扱う動詞には,出現数が多く複数の訳し分けがある動詞として「含む」(include, contain, comprise)を扱った. 訳し分け辞書に登録するテキストには連続する名詞の場合最後の名詞のみを残し,動詞を中心とした単語 5gram を抜き出す前処理を行う.学習用の特許文から各日本語動詞をテキスト中に含み,動詞の対訳が上記の語となる文を収集し,それら用いて訳し分け辞書を作成した.作成した訳し分け辞書から（出現数 / 出現総数）でスコアを算出する.学習数は include:89542, contain:35492, comprise:3472 であった.

実験の評価に用いるテキストは,特許明細書文学習データとして,用いていないテキストからランダムに抜き出して用いた.Google 機械翻訳の翻訳結果に対し,本システムによる修正の結果,訳し分けがどの程度できているのか精度を調べた.

3.2 実験結果

実験の結果を表２に示す.「正」は正しい訳,「誤」は誤った訳を表し,「正→誤」は Google 翻訳の出力は正しいが訳し分けの修正の結果判定が誤った,ということを表す.

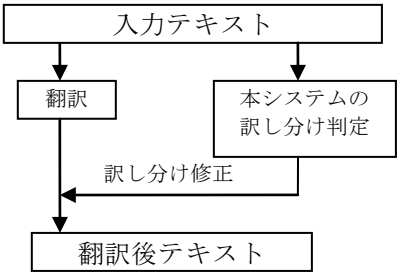

図 2:実験の流れ

表 2:実験結果

	include	contain	comprise
正→正	14	23	11
正→誤	9	5	1
誤→正	13	11	12
誤→誤	14	11	10

誤り文の訂正率 50%,正答の修正を含めた全体での正答率は 62%という結果になった.Google 翻訳の訳し分けの正答率が 47%であったことを考慮するとある程度訳し分けを改善できたと考える.

今回は動詞を中心とした 5gram で訳し分け辞書を作成したが,係り受けを考慮することや,スコアの計算方法を見直すこと,特許文に特化したグループ辞書の導入により,訳し分け精度の改善が期待できると考える.

4. 終わりに

本論文では語のグループ化を用いて特許文動詞の訳し分けをするシステムを作成した.訳し分けの精度を調べるために,Google 機械翻訳の結果を修正する実験を行った.実験の結果,ある程度訳し分けの改善を確認出来た.訳し分け精度の向上のためには特許文用に特化したグループ辞書の作成や,スコア計算方法の改善が必要である.

謝辞

本研究に際し,Japio から,資料の提供を賜りました.ここに感謝の意を表します.

参考文献

[1]麻野間直樹,中岩浩巳：目的言語の単語共起情報を利用した訳語選択と未知語の訳出,言語処理学会第５回年次大会論文集,pp.442-448,(1999)

[2]S.Yokoyama, Y.Takano: Investigation for Translation Disambiguatioin of Verbs in Patent Sentences using Word Grouping, *Proceedings of the 4th Workshop on Patent Translation*(2011)

[3] (財) 日本特許情報機構(Japio)：AAMT / Japio 特許翻訳研究特許情報データベース(2008)

[4] 池原他:日本語語彙大系(CD-ROM 版),岩波書店(1999)