英語科学技術論文執筆支援のための日英対訳例文データベース自動構築手法

吉岡直輝† 松本和幸† 任福継†

†徳島大学

1. はじめに

用例ベースの英作文支援システムとして，例文データベースに登録されている例文から目的に合致する文を検索し，文中の名詞等の単語を入れ替えることにより，英文を作成するシステムがある．しかし，このようなシステムで用いる例文データベースを構築する際，入れ替え対象の名詞やフレーズ部分の日英対応付けなどを，人手によって行う必要がある．

我々は，これまでに機械翻訳で応用されているSuper-Functionの理論を応用して，日英対訳コーパスからの日英対訳例文データベースの自動構築を行ってきた．しかし，文が属するカテゴリに依存して入れ替え対象の単語の種類が変わることもあるため，従来手法では，様々なカテゴリへの対応が困難である．

そこで，本研究では，英語科学技術論文執筆支援のための日英対訳例文データベース自動構築手法を提案する．

2. 英作文支援システム

我々の研究グループで構築するシステムは，ユーザが書きたい文にマッチする例文を例文データベースから探し出し，抽出した文の名詞の部分を入れ替えて任意の文を作成することができるシステムである．下記に例を示す．上部が例文データベースから参照した文で下部が抽出した文の名詞の部分を入れ替えて作成した文である．ユーザは②の「学校」を「図書館」に入れ替える．そうすると，英語の対応箇所「school」の部分が，「library」に自動で置き換わる．

日本：私は①学校②にいく。
英：I① go to school②.

日本：私は図書館②に行く。
英：I go to library②.

A method to construct English-Japanese builingual sentence database to support writing papers of science and technology in English

† Naoki Yoshioka
† Kazuyuki Matsumoto
† Fuji Ren
Tokushima University (†)

3. Super-Function

Super-Functionはある事象を別の事象に変換する関数である．Super-Functionに基づいて機械翻訳を行う場合，Super-Functionを原言語と目標言語（ここでは日本語と英語）の対応を示す関数と定義する．また，双方の言語は定数か変数に属すると定義する．変数を名詞とその修飾語，それ以外を定数とした場合，図2のような関係が得られる．

図の円は定数，矢印は変数を表す．図2の定数と，変数の関係を示したのが表1である．変数には位置と条件が記述されている．条件はそれぞれ主格，冠詞，指定なしを示す．図および表の φ は空文字を示す．文の最初に定数が存在しない場合に用いられる．

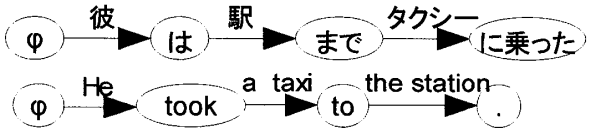

図1．例文「彼は駅までタクシーに乗った．」の関係

表1．変数と定数の関係

J	E	L_j	L_e	Condition
φ	φ	1	1	lp
は	took	2	3	a
まで	to	3	2	the
に乗った.	.	–	–	–

4. 提案手法

本手法は科学技術論文の対訳アブストラクト例文データベースを作成するために提案した。実験は以下の手順で行う．形態素解析された日英入力文における単語列を篠山ら[1]が提案した名詞分類規則に基づいて変数/定数に分類する．両言語には変数/定数を示すタグを付け，原言語の文における変数となる語について辞書を参照し，対応する日英単語の照合を行っていく．この際，英語については冠詞の除去とステミングを行ってから照合する．複合語の場合，単語ごとに日本語の訳語と，英語を比較し，完全一致した場合のみ対応付けを行う．対応付けできなかった単語は品

詞の並びでスコア付けして照合を行う．ここでも照合できなかった単語は単語N-gramのアブストラクト単位のDF値を使って頻出する単語を変数から定数に書き換えていく．本手法の流れを図3に示す．

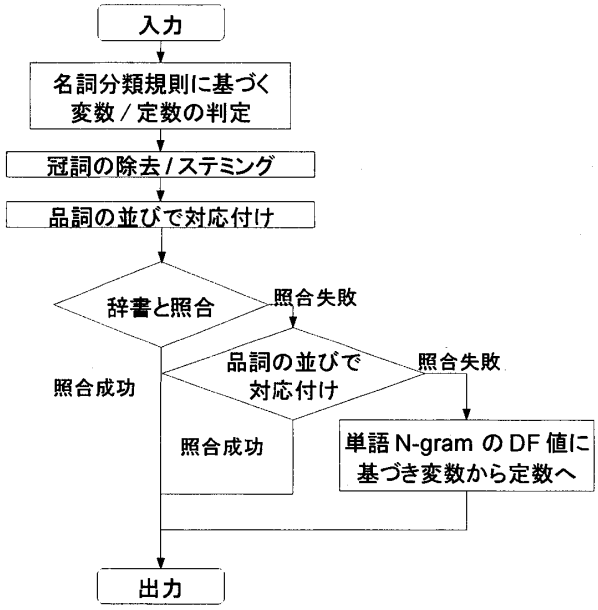

図3．日英対訳例文データベース生成手法の流れ

一般的な表現では，変数と定数の分類がほぼ可能である．しかし，以下に示すような論文表現になると分類が困難になる．

・対訳辞書に登録されていない単語を含む表現
・論文特有の言い回しを含む表現

つまり，日英が対応する箇所を固定できないため自動的にデータベースを作るのが困難になる．

そこで，品詞の並びおよび単語N-gramのDF値を用いて日英文の単語を自動で対応付けを行う．

5. 実験

5. 1. 実験概要

自然言語の分野の日英文が対訳となっている論文のアブストラクトを対象とする．

論文検索サイトCiNii(http://ci.nii.ac.jp/)のキーワード検索により論文のアブストラクトの抽出を行った．155件の論文から抽出し，本実験では50件の論文のアブストラクトを対象とした．辞書と照合し対応付けを行い，その後，品詞の並びで対応付けを行った．

実験結果の評価は，適合率と再現率で行う．すべての文に対して適合率と再現率を算出し，文の数で割った平均値を出す．

5. 2. 実験結果と考察

適合率は42.81%，再現率は11.81%だった．この結果を見ると適合率，再現率ともに悪いことがわかる．理由として次のことが挙げられる．単語毎の翻訳を用いる場合には完全一致ということにしているが，品詞並びによる照合では，学習サンプルから計算した相対頻度スコアをとして，これを元に照合するので，このスコアの閾値を適切に設定するべきである．現在は，スコアの閾値を0.5と決めているがこのパラメータを検討する必要がある．

また，訳語選択については，要素合成法[5]等でも提案されているような，信頼度スコアを与えるべきである．

6. おわりに

本稿では，Super-Functionに基づいた英語科学技術論文の日英文の対応付け手法について述べた。本手法により，英作文支援システムにおける作文の柔軟な対応ができ，対応つけの幅が広がった．しかし，実験を行ったデータ数が少ないことが問題である．

今後はデータ数をふやして，品詞の対応付けのルールの構築と単語N-gramのDF値を用いた変数/定数分類方法について研究する予定である．

7. 参考文献

[1]Manabu SASAYAMA, Fuji Ren and Shingo Kuroiwa : Automatic Super-Function Extraction for Translation of Spoken Dialogue, IEEE NLP-KE2007, pp.141-148, Beijing, Aug.2007.
[2]F. Ren:
``Super-function based machine translation'', Language Engineering,Proceedings of JSCL and TsingHua University Press,pp.305–312.(1997)
[3]E. Brill:
``Some advances in transformation-based part of speech tagging'',Proceedings of AAAI,(1994)
[4]ChaSen version 1.0 is officially released on 19 February 1997 by Computational Linguistics Laboratory,Graduate School of Information Science,Nara Institute of Science and Technology
[5]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 Apr. 2007