E-017

ウェブ検索を用いて中国語作文支援システムの構築

王　キョ嘉† Jujia Wang　柳　クン† Jun Liu　村岡　洋一‡ Yoichi Muraoka　秋岡　さやか‡ Sayaka Akioka

概要

近年，日本企業の中国への進出や国際文化交流などにより，中国語を学ぼうとする人々の数が急激に増えている．それとともに，中国語で文章を書くチャンスも増加している．しかしながら，日本人にとって，日中辞書だけで正しい中国語文章の作成が難しく，翻訳サイトや翻訳ソフトなどで作成されるのは直訳と呼ばれる訳文が多く，自然な中国語文章作成は困難である．このため，中国語作文を作成するための支援システムが必要だと思われる．本研究は，特定したコーパスを用いる手法と異なり，膨大な情報量を持つウェブ検索を利用し，検索結果の分析から自然な中国語の候補文を提示し，ユーザが候補文から一番適切なフレーズを選択する手法である．

１．はじめに

近年，中国の WTO 加盟による日中間のビジネス機会増大や，留学・旅行・インターネットなどによる交流機会増大を背景として，日本では中国語学習への興味や関心が高くなり，中国語を習得したいというニーズが増えている．そして，中国語作文を作成するチャンスも増加している．中国語の文法が日本語と明らかに異なる点以外，特別な文型や動詞なども日本人にとって学習の難点だと思われる．従って，日中辞書だけで文の構造に関する情報が得られず，正しい中国語文章の作成はできなく，翻訳サイトや翻訳ソフトなどで作成されるのは直訳と呼ばれる訳文が多く，自然な中国語文章作成は困難である．このため，中国語作文を作成するための支援システムが求められている．

中国語を作成するための支援には，中国単言語コーパスを用いる方法と，日中対訳コーパスを用いる方法がある．前者は，大規模なテキストを確保することで，汎用性が確保することができるため，様々な分野の文章に対して適用することができるが，単言語コーパスは，あらゆる分野の文章を含むので，訳文の精度は低くなる．後者は，新聞記事や小説など既存の翻訳資源を利用した分野を特定したコーパスを用いるので，当該分野においての訳文は，高い精度を保つ．しかしながら，中国語は非常に複雑なので，人手でより大規模の対訳コーパスの構築をしなければならず，時間やコストがかかる欠点がある．

本研究はこれらの特定したコーパスを用いる手法と異なり，膨大な情報量を持つウェブを利用し，フレーズ検索とワイルドカード検索を用いて，検索結果の分析から自然な中国語の候補文を提示し，ユーザが候補文から一番適切なフレーズを選択する手法である．

† 早稲田大学大学院基幹理工学研究科情報理工学専攻

‡ 早稲田大学理工学術院

２．関連研究

中国語作文支援のアプローチとして，大きく分けて２つのアプローチがある．

1. 単言語コーパスを用いる方法
2. 文対応が付けられた対訳コーパスを用いる方法

単言語コーパスを用いた研究例として，CCL[1]，国家語委現代漢語語料庫[2]，Kiwi[3] などがあり，いずれも大量のデータを入手できるため，実際の利用例を参照したり，一般に使用されている表現かどうか調べるといった目的に利用できる．しかし，中国語の習熟度が高い利用者でなければ，中国語だけの例文を見てもすぐに意味を理解することができないという欠点がある．

対訳コーパスを用いた研究例として，日中対訳コーパス[4] がある．日本語と中国語の例文を同時に参照することにより，例文中にわからない単語やフレーズがある場合もスムーズに意味を理解し，利用方法を習得できるという利点がある．しかし，中国語単言語コーパスに比較すると，利用できるデータ量が圧倒的に少なく，目的とする例文を発見できない場合があるという問題がある．

３．WEB を利用した中国語作文の検討

3.1　翻訳サイトによる翻訳結果の分析

WEB 検索を用いて，必要なフレーズを抽出することが実現できるかどうかを検討してみた．まず，翻訳サイト EXCITE を利用し，翻訳結果の精度および問題点を分析した．ここで，10 文字～20 文字の日本語の３００文を例文として実験を行った．表１は翻訳文を単語数が５以下（少単語数クラス），単語数が６～９（中単語数クラス），単語数が１０以上（多単語数クラス）に分類する．図１は翻訳文に含まれる単語数毎の正解訳と誤訳に分類される割合を示す．誤訳は語順ミス，言葉違い，全然違う三種類に分けた．図１から誤訳のうち５０％近くが語順ミスであることが判明した．本研究は語順ミスを対象に，検索エンジンを利用することによって，自然な中国語に直せるかどうかについて検討する．

表１：翻訳文の単語数に応じたクラス分類

クラス	単語数	占める割合
少単語数クラス	5 以下	4.35%
中単語数クラス	6～9	71.74%
多単語数クラス	10 以上	23.91%

3.2　WEB 検索によって語順ミスの訂正可能性

前節で調べた結果，語順ミスは一つの正しく翻訳されない原因だと判明した．この節で，WEB 検索を用いて，自然な中国語に直せるかどうかについて検討する．

ここで一つの例を挙げる．

日本語：彼は映画を見に行った

中国語の翻訳結果：他去了看电影

この翻訳結果には，”了”の位置が間違っている．中国語

の文法には確か過去に行われたことを示す場合，”了”を使うのは間違っていない．しかし「去」の動詞は”去”の目的を表す動作目的語（目的語が動作である）なので，” 去”の後に，”了”の動態助詞を使えない．過去のことを表すには，文末に”了”を付ける．この文は WEB で検索してみると，同じ表現が見つからなかった．そして，この文をばらばらにキーワードで検索してみる．”他 + 去 + 看电影”の検索結果から，”去看电影”という表現が出てきた．次に”他 + 去看电影”というキーワードを入れて検索した結果，”去看电影了”という表現が出てきた．これによって”他去看电影了”という自然な中国語は WEB 検索により検索可能だということが分かった．

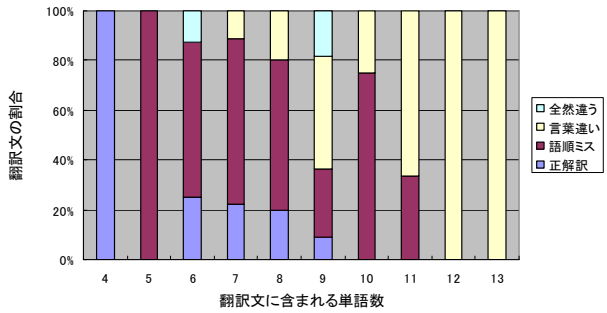

図１：単語数毎の正解訳と誤訳の割合

４．提案手法

4.1　システムの構築

本システムは，検索エンジンを使って，全文のヒット数だけで翻訳文のよさを判断することは難しく，検索フレーズの加工と検索結果の分析を行う．図２に基本となる支援システムを示す．まず，WEB 翻訳サービス API を利用し，翻訳文を得る．その文から，形態素解析により，キーワードを抽出する．結合パターン対により生成したキーワードを検索エンジンに入力する．次に，検索エンジンから得られた検索結果から木構造の生成と解析という分析を行う．最終的に翻訳文に対して，訂正及び共起表現などを提示する．

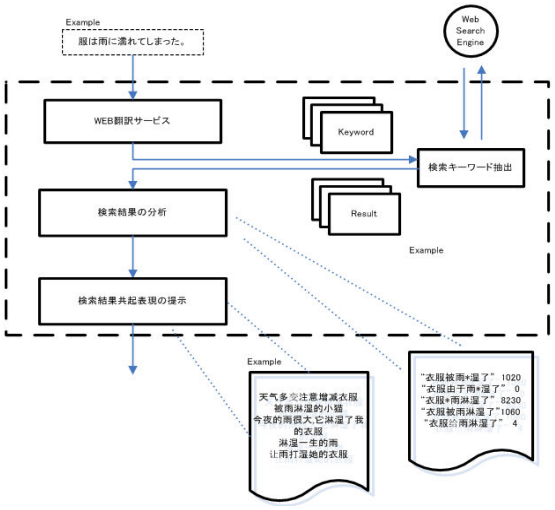

図２：ウェブ検索を用いて中国語作文支援システム

4.2　評価実験

今回は語順ミスの１４２文をサンプルとして，本システムで調べた．単語数と平均キーワード数および検索回数の関係を表２に示す．図３は単語数毎の語順ミス訂正率を示す．図３から，本システムを使って，５５％が自然な中国語に直せ，２７％が意味の似ている文が検索されることが分かった．これによって，ユーザの書きたい文章が８０%以上表示されることを言える．

表２：単語数と平均キーワード数及び検索回数の関係

単語数	平均キーワード数	平均検索回数
5	3	2
6	3	2.3
7	3	2.6
8	3	2.9
9	4	3
10	5	3.3
11	5	4.1

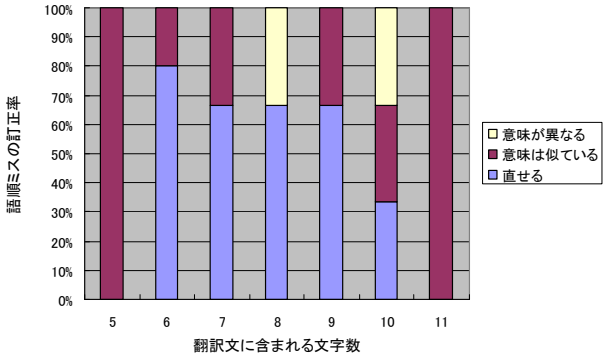

図３：単語数と語順ミスの訂正率

５．考察および今後の課題

本研究は中国語作文作成支援システムとして，WEB 検索を用いた．今回は語順ミスを対象として評価実験を行った．実験結果は本システムを利用し，ユーザの書きたい文章が８０%以上表示できることが分かり，WEB 検索で自然な中国語を作成することが可能と言える．

今回は語順ミスのみに対応したが，今後言葉違いの訂正などにも対応していく予定である．それから，検索精度が更に上がるために，システムを改善していく予定である．

参考文献

[1] 北京大学漢語語言学研究中心，“ CCL 語料庫検索系統（網絡版）”，
http://ccl.pku.edu.cn:8080/ccl corpus/jsearch/
[2] 中国国家語言文字応用委員会，“ 国家語委現代漢語語料庫”,http://www.clr.org.cn/retrieval/
[3] 東京大学の中川研究室，“ Kiwi System,
http://kiwi.r.dl.itc.u-tokyo.ac.jp/kiwi-0.1/
[4] 曹大峰(北京日本学研究中心，国立国語研究所)，千葉庄寿(麗澤大学) 北京日本学研究センター，“ 日中対訳コーパス（中日対訳語料庫）第一版の利用方法”, 2007.1.25