社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93 (2)
2009－ＮＬ－189 (2)
2009／1／22

検索エンジンを用いた日中翻訳支援システムの提案

－日中機械翻訳における語順誤訳の訂正－

王 キョ嘉† 柳 クン† 村岡 洋一‡ 秋岡 さやか‡

† 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

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．この状況に対し，本研究は日中翻訳サイトによる翻訳された誤訳文の一部を自然な中国語に訂正するシステムを提案する．語順誤訳文を対象に，１６種類の文型パターンに分類する．膨大な情報量を持つウェブ検索を利用し，検索結果から分析を行い，自然な中国語の候補文を提示する．

Proposal of Japanese-Chinese Translation Support System using Search Engine

－Correction of word order mistranslation on Japanese-Chinese machine translation－

Jujia Wang† Jun Liu† Yoichi Muraoka‡ Sayaka Akioka‡

† WASEDA University Graduate School of Fundamental Science and Engineering

‡ WASEDA University Faculty of Science and Engineering

Abstract

Recently, the number of people who try to study Chinese has increased rapidly by advancement of Japanese enterprise to China and International Cultural Exchange. So the chance to write sentences in Chinese increases, too. However, making correct Chinese sentences is difficult only in a Japanese-Chinese dictionary for the Japanese. And lots of the translation made with the translation site and the translation software is called literal translation. Therefore, it is difficult to make natural Chinese sentence. To this situation, the research proposes the system that corrects a part of the translated mistranslation sentence by the Japanese-Chinese translation site to natural Chinese. The word order mistranslation of the translation is classified in 16 kinds of phrase pattern. We use the web searching with huge volume of information, analyze the retrieval result, and present the candidate sentences of natural Chinese.

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加や，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．それに伴い，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．

この状況に対して，本研究は自然な中国語を提示するために，既存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する．既存の日中翻訳サイトの語順誤訳文を対象に，１６種類の文型パターンに分け，フレーズ検索を用いて，検索結果の分析から自然な中国語の候補文を提示する．

２．関連研究

2.1　中国語作文支援のアプローチ

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，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．

2.2　検索エンジンを使った英作文支援システムの構築

早稲田大学山名研究室は検索エンジンを使い，英作文支援システムを構築した[5]．この研究は日英翻訳の問題点を解決するための検索エンジンを使った翻訳手法を検討し，人手での作業においては，それぞれのフレーズを検索エンジンで検索し，検索結果を見て比較するという手間がかかってしまったり，ワイルドカードを使用して検索を行ったときにそれぞれの検索結果を見ていったりするのが大変であるという問題を注目した．これらの問題点を解決するために，GoogleAPI を用いて，検索エンジンを使った英作文支援システムを構築し，各フレーズにおいて検索式を入力して調べるという手間が省くことができた．

３．日中翻訳の問題点

3.1　翻訳市場の現状

日本翻訳連盟 JTF 翻訳白書 2005 年度より，中国語翻訳の市場は 2005 年度日本の翻訳市場規模は 2,000 億円と推定され，中国語の売上は全言語売上高の 7 %を占めた．また，翻訳ソフトの翻訳精度が低く，専門家翻訳のみ採用されているのは現状である．この状況に対し，対訳機能無料化，専門分野の辞書と支援ツール有料化の傾向がある．

3.2　翻訳サイトによる翻訳結果の分析

ここで，既存の翻訳サイトの翻訳現状及び問題点を調査する．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．今回は「誤用から学ぶ中国語」[9]と「中国語表現法マニュアル」[10]の二冊本から，10 文字～20 文字の日本語の３００文を抽出した．表１は翻訳文を単語数（フレーズ）が４以下（少単語数クラス），単語数が５～９（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．誤訳は語順誤訳，語順＋語彙誤訳，語順＋文型誤訳，語彙誤訳，文型誤訳，

その他誤訳の６種類に分けた．図１から語順を含む誤訳（語順誤訳，語順＋語彙誤訳，語順＋文型誤訳）は誤訳のうち５０％近くことが判明した．単純な語順誤訳は２９％である．本研究は語順誤訳を対象に，検索エンジンを利用し，語順誤訳訂正の可能性を検討する．

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	4 以下	8.35%
中単語数クラス	5～9	76.66%
多単語数クラス	10 以上	14.99%

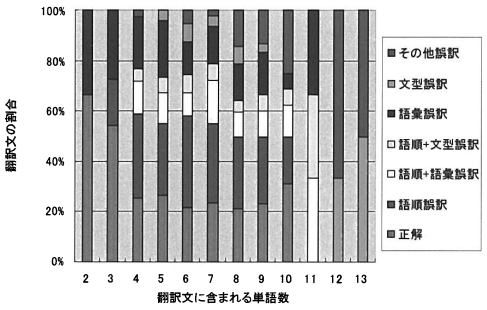

図１：単語数毎の正解訳と誤訳の割合

４．語順誤訳パターン

4.1　中国語の特徴

中国語は，シナ・チベット語族に属し，中国語祖語を元に，タイ諸語などの南方諸語やモンゴル語，満州語など北のアルタイ諸語の発音，語彙，文法など特徴を取り込みながら文化したと考えられている．特徴として，声調を持ち，孤立語で，単音節言語であることが挙げられる[7]．

日本語と中国語は同じ漢字を用いているものの，基本的に違う言語構造を持つ言語である．両言語にはそれぞれ独自の特質，例えば，独特の文法構造，表現形式などがあり，それらがそれぞれの言語の個性ある特徴を成しているので，翻訳を行う場合は文法構造，表現形式が平行して単純にそのまま対応することはない．一方，両言語における発想の相違が存在することにより，同じ意味内容の表現はいつも同じ表現形式がなされるとは限らない．むしろそれぞれ別の表現形式をとることにより，独特な様相を呈することが多い[8]．更に，中国語には，日本語の格助詞に相当するものがなく，語尾変化もないので，単語間の文法的関係は，主として，単語の並べ方（語順）によって表されている．従って，中国語では，語順が極めて重要である．

4.2　中国語の文型パターン

中国語の文は，述語の性質によって，名詞述語文・形容詞述語文・動詞述語文の三つの種類に区分できる．このうち，名詞述語文と形容詞述語文は簡単であり，動詞述語文は中国語文の大部分を占めている．文の構造はその文の動詞の統語論的特長によって，著しく限定される．本研究は中国語の特性によって分類し，10 種類の文型パターンを分けた．

1. 主語＋名詞述語（フレーズ）
   例：他的生日幾月幾号.
   （彼の誕生日は何月何日ですか．）
2. 主語＋（状）＋形容詞＋（補語）
   例：這間屋子太小了.
   （この部屋はとても小さいですね．）
3. 主語＋（状）＋自動詞＋名詞・文
   例：昨天我出去了.
   （昨日私は出かけました．）
4. 主語＋（状）＋他動詞＋目的語
   例：我特別喜歓運動.
   （私はスポーツが大好きです．）
5. 二重主語文
   例：中国人口多.
   （中国は人口が多い．）
6. 主語＋是＋（述語＋的）
   例：我是昨天晩上到的.
   （私は昨晩着いたのです．）
7. 存在文
   例：他在食堂.
   （彼は食堂にいます．）
8. 把字文
   例：我把相机帯来了.
   （私はカメラを持ってきました．）
9. 被字文
   例：那個録相帯被人借走了.
   （あのビデオは人に貸し出された．）
10. 連動文
   例：媽媽去商店買東西.
   （母は買い物をしにお店に行っています．）

4.3　語順誤訳の種類

前節の文型パターンをもとに，前章で判明した語順誤訳文を１６種類のパターンに分けた（表２）．1. 主語＋名詞述語（フレーズ）と 2. 主語＋（状）＋形容詞＋（補語）の表現方法は限られているため，高い精度で翻訳された．それ以外の語順誤訳文型パターンを中国語形態素解析ツール ICTCLAS で解析した．

表２：語順誤訳の文型パターン

1. 主語＋（状）＋自動詞＋名詞・文	
日本語	彼は二回結婚したことがある.
訳文	他二次结婚过.
正解訳	他结婚过二次.
パターン	m/q/vi/u → vi/u/m/q
日本語	彼は映画を見に行った.
訳文	他去了看电影.
正解訳	他去看电影了.
パターン	vf/y/v/n → vf/v/n/y
日本語	私は一時間待ちました.
訳文	我一小时等了.
正解訳	我等了一小时.
パターン	m/n/v/y → v/u/m/n

2. 主語＋（状）＋他動詞＋目的語	
日本語	私はもう宿題を書き終えた.
訳文	我已经写结束了作业.
正解訳	我已经结束了写作业.
パターン	d/v/v/u/n → d/v/u/v/n
日本語	僕は毎日八時間寝る.
訳文	我睡觉每天八小时.
正解訳	我每天睡觉八小时.
パターン	vi/r/m/n → r/vi/m/n
日本語	昨日一時間テレビを見た.
訳文	昨天一小时看了电视.
正解訳	昨天看了一小时电视.
パターン	m/n/v/u/n → v/u/m/n/n
日本語	彼はあなたを二年間だました.
訳文	他二年欺骗了你.
正解訳	他欺骗了你二年.
パターン	m/q/v/u/r(n) → v/u/r(n)/m/q
日本語	彼はねずみが一番いやだ.
訳文	他鼠最讨厌.
正解訳	他最讨厌鼠.
パターン	n/d/v → d/v/n

3. 二重主語文	
日本語	彼は大学を卒業しました.
訳文	他毕业了大学.
正解訳	他大学毕业了.
パターン	r/y/n → n/r/y
日本語	この子は大変マナーが悪い.
訳文	这个孩子非常礼仪坏.
正解訳	这个孩子礼仪非常坏.
パターン	n/d/n/a → n/n/d/a

4. 主語＋是＋（述語＋的）	
日本語	今日は何曜日ですか.
訳文	今天是几星期？
正解訳	今天是星期几？
パターン	vshi/m/n → vshi/n/m

5. 存在文	
日本語	日本はたくさんの島が存在する.
訳文	日本很多的岛存在.
正解訳	日本存在很多的岛.
パターン	m/ude/n/v → v/m/ude/n
日本語	机の上に写真が並べてある.
訳文	在桌子上面照片排列着.
正解訳	在桌子上面排列着照片.
パターン	p/n/f/n/v/uzhe → p/n/f/v/uzhe/n

6. 把字文	
日本語	私は鍵を部屋に忘れてしまいました.
訳文	我把钥匙在房间里(上)遗忘了.
正解訳	我把钥匙遗忘在房间里(上)了.
パターン	pba/n/p/n/f/v/y → pba/n/v/p/n/f/y

7. 被字文	
日本語	資料は風に吹かれてばらばらになった.
訳文	资料风被吹七零八落了.
正解訳	资料被风吹七零八落了.
パターン	n/pbei/v/v/y → pbei/n/v/v/y

8. 連動文	
日本語	彼らは遊びに行きました.
訳文	他们去了游玩.
正解訳	他们去游玩了
パターン	vf/y/v → vf/v/y

記号意味：r 代名詞，v 動詞（vshi 是），n 名詞，u 助詞，m 数詞，q 助数詞，p 前置詞（pbei 被，pba 把），d 副詞，y 助詞(語気)，f 方向詞

５．提案手法

5.1 システムの概要

本研究は，前章の１６種類のパターンをベースに，システムを構築した．まず，初心者が日本語単文を入力する．システム側は自動的に WEB 翻訳サービス（EXCITE）を利用し，翻訳文を得る．長文が入力された場合，システム側で句単位に分ける．そしてチェックしたい文を選択し，検索する．元々正しい文章は検索した後違う文章になってしまうことを防ぐために，最初は文章をそのまま検索させる．検索結果がある値以下であれば，システム側で形態素解析を行い，フレーズを抽出してから再度検索させる．同様にヒットされなかったら，不自然な文章と判断される．次に，不自然な文章と判断された文を既存の文型パターンと照合する．一致パターンが存在する

場合，パターンの交換ルールで単語位置を変換し，検索させる．最後にヒット数が多いかつ原文との差が小さいパターンが訂正文として表示し，その他のヒットされた表現が関連情報として提示する．

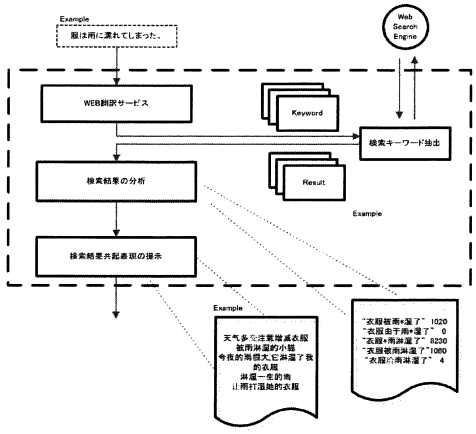

図２：システムの概要

5.2　例文訂正の説明

日本語文：彼は大学を卒業しました．
訳文：他毕业了大学.

まず，訳文は正解かどうかを確認する．部分的なフレーズをピックアップし，検索エンジンに投げる．ヒット数が０だったため，不自然な中国語の可能性が高いと判断される．そして，中国語形態素解析を行う．

他/rr 毕业/v 了/ule 大学/n

このパターンを既存の語順誤訳パターンと照合し，一致するパターンがあるため，それの変換ルールに従い，フレーズ位置を変換し，「他大学毕业了」の候補が提示される．二つのフレーズを同時に検索エンジンに投げ，「他毕业了大学」のヒット数が 0 であり，「他大学毕业了」のヒット数が 1,850 であったため，「他大学毕业了」の表現が一般的であることが分かる．

5.3　評価実験

今回は中国語検定三級の過去問題集から，新たな代表的な２５０文を用意し評価実験を行った．表３はシステム判断結果を示す．表３から，元々正解だった訳文は正しく判断されるのは 36.8%であり，純語順誤訳の訂正率は 62.8%であることが分かる．表４，表５は単語数毎の語順誤訳訂正率及び文型パターン毎の語順誤訳訂正率を示す．

表３：システム判断結果

	人間判断	システム判断
正解	68	25
純語順誤訳	43	27
その他	139	198

表４：単語数毎の語順誤訳訂正率

単語数	語順誤訳数	訂正した数
4	3	2
5	11	7
6	14	8
7	7	5
8	4	3
9	2	1
10	2	1

表５：文型パターン毎の語順誤訳訂正率

文型パターン	語順誤訳数	訂正した数
主語＋（状）＋自動詞＋名詞・文	10	8
主語＋（状）＋他動詞＋目的語	20	13
二重主語文	1	1
主語＋是＋（述語＋的）	2	1
存在文	4	2
把字文	1	0
被字文	1	0
連動文	4	2

６．考察

前章の評価実験結果から，単純な語順誤訳の訂正率は 62.8%となった．人間が正解だと判断した文をシステムは判断できなかった原因は完全一致でフレーズを検索させたため，ヒット数が０のケースが多かったからである．

語順がパターン照合できなかったのは，以下の理由と考えられる．

１．たとえ似た文型の日本語を入力しても，翻訳サイトに翻訳された訳文は異なる．

例：原文①：私は二年間中国語を勉強する．
　　中国語訳①：我二年学习中文.
　　原文②：私は 2 時間中国語を勉強する．
　　中国語訳②：我学习２小时中文

中国語訳①はシステム側に存在する文型パターンであり，状況語が変わっただけで，翻訳結果の文型が変わってしまい，訂正ができなかった．

２．たとえ人間が同じ文型構成だと見えても，中国語形態素解析ツールによる解析された品詞が異なる場合がある．

例：原文①：私は車を買う計画がない.
中国語訳①：没有买车的计划我.
没有/d 买/v 车/n 的/ude 计划/n 我/rr
原文②：私は家を買う計画がない.
中国語訳②：没有买家的计划我.
没有/v 买家/n 的/ude 计划/n 我/rr
同じ構成に見えるが，形態素解析は“买车”は動詞＋名詞と分析したことに対し，“买家”は名詞と判断されてしまった.

3．登録されていないパターンがある.
例：原文：あなたはどうしてちゃんと考えないのですか
中国語訳：你不考虑怎么

7．まとめ

近年の急速な経済発展を背景に，日本企業にとって中国の重要性が非常に高まっている.中国進出企業のビジネス活動支援を主な目的として，日中機械翻訳が注目されている．しかしながら，機械翻訳に翻訳されたのは，自然な中国語ではないため，中国語作成の支援ツールが必要である．この状況に対して，本研究はウェブ検索及び事前登録した文型パターンを利用し，既存の日中翻訳サイトの語順誤訳文を対象に，自然な中国語を訂正する手法を提案した.

評価実験は既存のパターン照合が一致した文章に対して，完全に自然な中国語に対応できるが，それの割合は 63.8%であった．パターン照合できなかった原因は翻訳サイト及び中国語形態素解析ツールの精度に関係がある一方，パターンの充実も必要だ思われる.

今後の課題として，精度向上のため，パターンを増やしていく予定である．また，今回は語順誤訳のみに対応したが，今後語彙誤訳や文型誤訳の訂正などにも対応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
[5] 大鹿　広憲，「検索エンジンを使った英作文支援システムの構築」，山名研究室 2004 年度修士論文
[6] Tara Calishain, Rael Dornfest 著　山名　早人　監訳　石川　隼輔，堀井　洋，村上　明子，鹿島　久嗣，小柳　光生，「GOOGLE HACKS」第 2 版　オーム社　2005 年 8 月 24 日
[7] 中国語－フリー百科事典『ウィキペディア（Wikipedia）』，
http://ja.wikipedia.org/wiki/%E4%B8%AD%E5%9B%BD%E8%AA%9E
[8] 日中翻訳における表現構造に関する考察
国立国語研究所外国人研究員　　張　建華
1998 年
[9] 郭　春貴著，「誤用から学ぶ中国語－基礎から応用まで」　白帝社　2001 年 11 月 5 日
[10] 蘇　氷，山内　智恵美著，「中国語表現法マニュアル」　駿河台出版社　2005 年 5 月 20 日
[11]楊鵬,村上仁一,徳久雅人,池原悟，「結合価パターンを用いた日中機械翻訳システムの構築」，研究報告-情報学基礎（FI）,2008 年 1 月 Vol.2008 No.4
[12]張華平，ICTCLAS3.0(20080903)漢語分詞システム, http://ictclas.org/