7P-08

日常会話翻訳のための日中対訳文の自動推定

中島 浩平* 何 婉瑩‡ 王 振章† 張 文玉† 町田 翔* 延澤 志保*
*東京都市大学知識工学部 ‡東京都市大学知識工学部 / 大連交通大学外国語学院 †大連交通大学外国語学院
*東京都市大学大学院工学研究科

1 はじめに

現在，日中間でビジネスや留学，旅行などでの交流機会が増加しており，日本では中国語学習への興味や関心が高くなっている。しかし，日中辞書だけでは正しい中国語に翻訳することは難しく，翻訳サイトなどで作成される翻訳文は精度が向上しているが直訳された文章や誤った文章が出力されてしまうことがある．近年，Google の翻訳器がフレーズベース翻訳からニューラルネットワークを利用した翻訳に切り替わっている．そのため Google の機械翻訳の精度を再調査した．正解が 33%，意図推定可能が 23%と精度が向上しているものの日常会話の翻訳が完全とは言えなかった．

本研究では日本語日常会話文を複数の翻訳器で翻訳した結果を用いて適切な対訳文を出力する手法を提案する．

2 文の類似度に着目した対訳候補選択

2.1 異言語文間の意味的類似度計算

羅ら [1] は機械翻訳を利用した異言語文間の意味的類似度の計算のために，機械翻訳の影響について実験的に評価を行った．その結果，機械翻訳の精度が類似度計算に強く影響することが示された．意味的類似度とは言語表現の間の意味的な類似の程度を表す指標のことである．従来は単語や文書間の意味的類似度について主に検討されていたが近年では文間の意味的類似度の計算が主要としている．これによって複数の翻訳器で翻訳された訳文が似ていれば似ているほど翻訳の精度が高いものと考えられる．

2.2 文字の共起に着目した対訳候補選択

三浦は英日機械翻訳，下里は日英機械翻訳の対訳候補の評価を目的とし，Web のヒット数を元に共起率を計算し，複数の翻訳文の中からより自然な文を選択する手法を提案している [3, 4] ．共起とは文章中のある単語が出現した際に，特定の別の単語が出現することであり，その頻度を数値化したものが共起率である．三浦らの手法では，Web 検索で複数の単語を AND 検索した際のヒット数の多い単語同士ほど共起率が高いものであるとし，ヒット数に着目して訳文の選択を行っている．

3 提案手法の概要

本研究は，日本語文を 3 つの翻訳器で翻訳し，それらの訳文を比較し，修正することで 3 つの訳文から正しい文を出力する．

提案する手法は 2 つのステップで複数の翻訳文を比較して正しい翻訳文を出力する．システムの手順を図 1 に示す．まず，翻訳したい日本語文を複数の翻訳器にかけ，

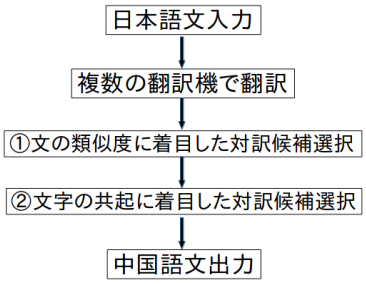

図 1: フローチャート

複数の対訳候補文を得る．

本研究では，それぞれの対訳候補文が正しければお互いの類似度は高くなるものと仮定する．そこで，STEP1 として，複数の対訳候補文に共通して含まれる形態素は正しい翻訳語句の可能性が高いものと仮定し，形態素の出現状況を基にそれぞれの対訳候補文の評価を行う．

STEP1 で複数の対訳候補文が選択された場合には，絞り込むため，STEP2 に進む．STEP2 では，それぞれの対訳候補文について，その文が自然な文と認められるかの推定を行う．具体的には，それぞれの対訳候補文に含まれている語同士の共起率を調べ，より共起率の高い翻訳文を選択する．

3.1 文の類似度に着目した対訳候補選択

山崎らは複数の翻訳文を形態素ごとに分け，出現頻度で点数をつけ (表 1) 他の対訳候補文との類似度が高い文を選択する手法を提案している [2]．表 2 の例では，例えば Google の翻訳文は 7 形態素に切り分けることができ，

表 1: 文の類似度に基づく候補文評価値の基準

形態素の共通出現文数	3 文	2 文	1 文
山崎らの手法 [2]	2 点	1 点	0 点
提案手法	3 点	2 点	1 点

Automatic Selection of Japanese-Chinese Translation of Everyday Conversation.
Kohei Nakajima*, WanYing He‡, ZhenZhang Wang†, WenYu Zhang†, Sho Machida*, and Shiho H.Nobesawa*.
* Faculty of Knowledge Engineering, Tokyo City University
‡ Faculty of Knowledge Engineering, Tokyo City University / School of Foreign Languages, Dalian Jiaotong University
† School of Foreign Languages, Dalian Jiaotong University
* Graduate School of Engineering, Tokyo City University

そのうち，例えば「能」は対訳候補3文すべてに出現しているため，山崎らの手法では形態素「能」の点数は2点となる．同様に7形態素それぞれについて点数を調べて足し合わせることで，Google翻訳文は類似度評価値9点を得る．山崎らの手法では表2の例ではGoogle翻訳文

表 2: 文の類似度に基づく候補文評価の例

翻訳機	翻訳文	類似度評価値	
	ちょっと手伝ってくれませんか	先行研究	提案手法
Google	你/能/帮/我/一/下/吗	9	2.29
Weblio	能/我/微/帮/助/吗	8	2.33
Excite	能/帮/助/一/下/吗	9	2.50

とExcite翻訳文が同点最高点を得てともに選択される．この手法は各形態素の点数の合計を類似度評価値としているため，形態素数が多い文が有利となること，類似度評価値が整数値のみとなり複数文が同じ点数を得る可能性が高いことの2つの問題がある．

そこで本研究では，山崎らの類似度評価値に対して2つの改良を提案する．まず，表1のように他の対訳候補文に出現しない形態素にも点数を与えることで，出現可能性の低い形態素の存在を無視していた問題を解決する．さらに，それぞれの形態素の点数の合計ではなく，それぞれの形態素の相加平均を類似度評価値とする．これに従って類似度評価値を計算すると，表2のように値は実数値となり，出現可能性の低い形態素も含めた文全体を考慮した類似度評価値を算出することができる．これにより，表2の例では，STEP1での出力をExcite翻訳文に絞ることができた．

3.2 文字の共起に着目した対訳候補選択

STEP2では，対訳候補文中の文字の共起率が高い文を選択することを目的とする．それぞれの対訳候補文をWeb検索し，そのヒット数に基づいて3種類の検索エンジンで検索し，より多くのヒット数を得た候補文をSTEP2の出力とする．STEP2における選択の例を図3に示す．この例の場合，3つの検索エンジンにおいてヒット数が

表 3: 文字の共起に着目した対訳候補選択の例

	この言葉を日本語に翻訳してください	Google検索	Sogou検索	百度検索
Google翻訳	请把这个词翻译成日文	612,000	34,654	950,000
Excite翻訳	请把这个言词翻译成日语	26,300	25,327	50,200

多かったGoogle翻訳の文を出力する．このように，Webでのヒット数が高かった候補文をより自然な文として選択する．

4 実験結果と考察

対訳候補文は正解，意図推定可能，意図推定不可能，不正解の4つに分類する．意図推定可能な文とは，翻訳語句や文法に誤りがあり正解とはいえないが原文の意図を推定することが可能なものを指す．語句の翻訳に失敗して日本語語句のローマ字表記など中国語以外の語句が含まれている文は不正解として扱う．それ以外の，翻訳語句や文法などの誤りのため原文の意図が推定できないものを意図推定不可能とする．

本研究の実験結果を図2に示す．図2にあるように最

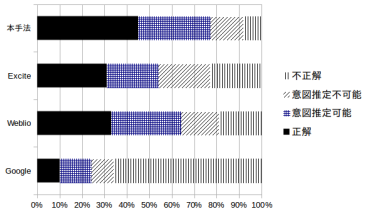

図 2: 本手法の結果

も正解率の高い翻訳器単体では64%に止まったが本手法を用いることで73%まで向上させることができた．

STEP1だけを用いた結果を図3に示す．

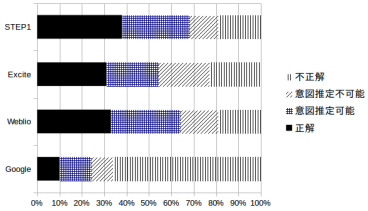

図 3: STEP1 のみの結果

STEP1で68%まで向上させることができた．STEP2だけで見ると，適用された文が13文であったが，そのうち5文でしか有効でなかった．

5 まとめ

本稿では，日本語の日常会話文を中国語へ翻訳するときの精度を向上させる手法を提案した．
本稿の手法では複数の翻訳器を用いて得られた対訳文を，文の類似度と文字の共起に着目して比較することで，より正しい文を選択した．実験の結果，最も正解率の高い翻訳器の64%を上回り，73%まで向上させることができた．既存の翻訳器単体を組み合わせることによって，その正解率よりも高い正解率を出すことに成功した．

参考文献

[1] 羅 文涛, 林 良彦, “機械翻訳を利用した異言語文間の意味的類似度計算の評価,” 言語処理学会 第22回年次大会発表論文集, pp.833–836, 2016.

[2] 山崎 亘涼, 孟 愛林, 張 文玉, 原田 千聖, 町田 翔, 延澤 志保, “日常会話を対象とした中日対訳文の自動選択,” 情報処理学会 第79回全国大会, vol.2, pp.559–560, 2017.

[3] 三浦 大, “複数の訳文候補に基づく英日機械翻訳,” 東京都市大学卒業論文, 2012.

[4] 下里 昌輝, 延澤 志保, “オノマトペを対象とした日英対訳語句の自動推定,” 電子情報通信学会 2016 年総合大会, vol.D-2, p.41, 2016.

[5] 張 文玉, 町田 翔, 孟 愛林, 延澤 志保, “構文に着目した日中機械翻訳候補文の自動修正,” 情報処理学会研究報告, vol.2017-NL-232, No.7, 2017.