5Q-01

日常会話を対象とした中日対訳文の自動選択

山﨑　亘涼†　　孟愛林*　　張文玉*　　原田千聖†　　町田翔†　　延澤志保†

†東京都市大学　知識工学部　情報科学科　*大連交通大学外国語学院／東京都市大学知識工学部

1 はじめに

現在，日本を訪れる外国人の数は年々増加しており，日本政府観光局によると2015年は1600万以上の人が日本を訪れた．訪日外国人の中でも近年特に中国人の数が多く2015年に日本を訪れた1600万人のうち1/4を占める400万人以上が中国人であった[1]．中国人が日本で日本語を扱う際の補助として，スマートフォン等で使用できる機械翻訳サービスが多く提供されている．しかし，翻訳機は便利な反面未だ精度が安定しておらず，時に誤った日本語を出力するという問題点がある[2]．翻訳機の精度を向上させるため，複数の翻訳機によって出力された翻訳文を比較し，正しい文を選択することで精度の高い出力が得られると考える．

従来研究に翻訳機を利用し意味的類似度の評価を行った研究がある[3]．意味的類似度とは，言語表現間の意味的な類似の程度を表す指標であり，この研究では翻訳機の品質が類似度計算の精度に大きく影響されることが示されている．また複数の訳文の中から共起率によって正しい訳文を選択する研究が行われている[4]．この研究の中で web 検索を用いて共起率を求める方法が使われており，2つの単語を AND 検索した時のヒット数の多さを共起率の高さとしている．

本研究では中国語の日常会話文を複数の翻訳機で翻訳した訳文を対象に，その中から正しい文を選択することを目的とする．

2 提案手法

本研究では中国語文を3つの翻訳機で翻訳し，それらの訳文を3つのステップで比較することで，3つの訳文の中から正しい文を選択し出力することを目的とする．

本手法の流れを図1に示す．

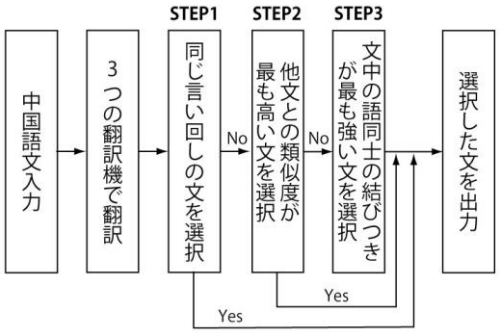

図１：本手法の流れ

2.1 同じ訳文に着目した対訳候補選択

STEP1では，「謝謝(ありがとう)」のような文を機械翻訳すると複数の翻訳機が同じ訳文を出力することが多い点に注目し，2つ以上の翻訳機が同じ対訳文を出力していた場合はその対訳文を正解として選択し出力する．例えば表1の場合は，翻訳機1，3の文を正解とする．

表1:STEP1の例

中国語(正解訳)	不客气(どういたしまして)
翻訳機1	どういたしまして
翻訳機2	遠慮しません
翻訳機3	どういたしまして

2.2 文の類似度に着目した対訳候補選択

STEP2では，従来研究で意味的類似度が翻訳機の品質に影響することが示されていた[3]ことから，それぞれの文中の形態素が他の翻訳機で出力された訳文中の形態素とどれだけ共通しているか形態素ごとに点数を付け，合計点数の最も高い文を出力とする．

点数は表2のように，他1文に存在する形態素には1点，他2文どちらにも存在する形態素にはより高い2点を与える．

表2:形態素ごとの点数

3文に出現する形態素	2点
2文に出現する形態素	1点
1文に出現する形態素	0点

例えば表3の場合「彼」という形態素は全ての

Automatic Sentence Selection of Chinese-Japanese Translation for Daily Conversation

Kosuke Yamazaki†, Ailin Meng*, Wenyu Zhang*, Chisato Harada†, Sho Machida†, Shiho Nobesawa†

† Faculty of Knowledge Engineering, Tokyo City University

*Dalian Jiangtong University / Tokyo City University

文に含まれているため2点，「の」は翻訳機1，3の2文に含まれているため1点，「再び」は翻訳機1にしか含まれないため0点となる．

表3:STEP2の例

中国語 (正解訳)	他叫什么名字来着？ (彼の名前は何でしたっけ?)	点数
翻訳機1	再び/彼/の/名前/は/なん/です/か	12点
翻訳機2	彼/は/なん/と/いって/名前/は/来/て/い/ます/か	11点
翻訳機3	彼/の/名前/は/何/と/言う/です/か	13点

また点数を決定する上で以下の3つのルールを定めた．

①数字には点数を付けない

②名詞以外の品詞で読みが同じものは同じ語として扱う

③2回以上登場する形態素は1回目に登場した形態素同士を同じ形態素として扱い，2回目に登場した形態素は2回目同士で比較し点数を付ける

①は数字を形態素解析した場合に区切れる場所に違いがあるため除外する．②は表3の「なん」と「何」の様に表記が異なるが意味が同じ語を同一の語として扱うためである．しかし，名詞に関しては表記が異なるだけで意味が大きく異なるものがあるため除外する．③は表3の翻訳機2の「は」のように複数出現する形態素により高い点数がついてしまうことを防ぐためである．

ルールに従って計算すると，表3の例では翻訳機3が最高点となり，翻訳機3を選択する．

最高点の文が複数ある場合はSTEP3に進む．

2.3 語の結びつきに着目した対訳候補選択

STEP3では，文中の語同士の結びつきの強い文を選択することを目的とする．従来研究で web 検索を用いて共起率を求める方法が使われていた[4]ことから，文を検索した結果のヒット数がより多い文は語同士の結びつきが強いと考える．STEP2での候補の文をそれぞれ3種類の検索機で検索し、2つ以上の検索機でよりヒット数の多い文を選択し，出力する．

3 実験結果

本実験には中日対訳文が掲載された web サイト「ゴガクル」[5]において日常会話というタグが付けられた中国語の文100文を使用する．この100文は Google，Yahoo，Excite の3つの翻訳機で翻訳し，いずれかが翻訳成功している文を用いる．また今回は機械翻訳された3文の他に，評価のためゴガクルに掲載された対応する日本語訳を正解として加えた計4つの訳文を対象に行う．

評価は人手で行い，以下の4段階で評価した．

[評価 A]ゴガクルと同じ意味

[評価 B]文法がおかしいがゴガクルと同じ意味

[評価 C]ゴガクルと意味が違う，意味不明

[評価 D]翻訳失敗

評価 A，B を正解，評価 C，D を不正解とする．

本手法を用いた実験結果を表4にまとめる．STEP の列はそれぞれどの段階で出力を決定したかを表している．

表4:ステップごとの実験結果

	STEP1	STEP2	STEP3	合計
	23	58	19	100
評価 A	21	33	10	64
評価 B	2	7	6	15
評価 C	0	18	3	21
評価 D	0	0	0	0
正解率	100%	69%	84%	79%

この実験結果から本手法を用いることでの正解率は79%となった．STEP1での正解率は100%となり複数の翻訳機が同一の訳文を出力した場合，意味の通じる文であることが分かった．また，STEP3での正解率は8割以上でありこちらも多くの場合で意味の通じる日本語を選択することができた．しかし STEP2での正解率は7割以下と他の段階に比べ低い結果となった．

4 おわりに

本手法によって8割近い正解率で意味が通じる日本語を選択することに成功した．半分以上の出力が STEP2で決定していることから，STEP2の精度を上げることが今後の課題と考えられる．

参考文献

[1]日本政府観光局(JNTO)
:http://www.jnto.go.jp/jpn/, 2016.

[2] 赤部晃一, Graham Neubig, 工藤拓, John Richardson, 中澤敏明, 星野翔, 〝ProjectNext における機械翻訳の誤り分析，〞言語処理学会第21回年次大会ワークショップ『自然言語処理におけるエラー分析』, 2015.

[3] 羅文涛, 林良彦編, 〝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.

[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，〞電子情報通信学会総合大会講演論文集，２０１６年

[5] ゴガクル:http://gogakuru.com/, 2017.