5Q-02

オノマトペを含む日中対訳を対象とした自動選択

原田　千聖 †　山崎　亘涼†　孟　愛林*　張　文玉*　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

近年,中国人留学生の数が増加傾向にあるJASSOの調査では平成16年には約7.7万人だった中国人留学生が,平成28年には約9.4万人に増加した[1].彼らは生活の中で意味の分からない日本語を,機械翻訳機で翻訳を行い,意味を調べる場合がある.中国人留学生と会話している最中に伝え辛い表現があった際には,機械翻訳機を使用してお互いの言葉を翻訳する.しかしながら,翻訳する表現のなかにオノマトペが含まれていた場合,説明的な意味を加えることが難しいため,翻訳結果によっては内容が理解できない場合がある.

2 関連研究

複数の日英機械翻訳機を対象にオノマトペを含む句を翻訳し,共起を用いて複数の対訳語句から最適な翻訳結果を取得する研究がある[2].しかし,下里らの手法は,評価値が翻訳結果の形態素の数に依存するため,正しい翻訳結果であっても,他の機械翻訳機よりも形態素の数が多かった場合,正しい翻訳結果が推定されない問題がある.

そこで本稿では,オノマトペを含む句を対象に,複数の日中機械翻訳機が出力した対訳語句から,共起を用いて最適な翻訳結果を選択する手法を提案する.また,評価値が形態素の数に依存することを防ぐため,文法に則ったパターンごとに共起率の重み付けを行う.

3 研究手法

4種類の機械翻訳機(Excite 翻訳,Yahoo 翻訳 Google 翻訳,Weblio 翻訳)が出力した対訳語句を入力する.次に,文法に着目して定義した2つの条件に沿って共起率を求める.共起率とは,ある単語とある単語が同時に出現する確率のことをいう.次に,共起率と品詞の組み合わせごとに定義した重みを使用して対訳語句の評価値を求める.評価値が最も高い対訳語句を最適対訳語句として出力する.

本稿で扱う教師データとテストデータはオノマトペと被修飾語からなる句のみを扱い,被修飾語は名詞と動詞のみに限定する.オノマトペを含む句は「擬音語・擬態語4500」からランダムに選択し,65句を教師データに使用しており,57句をテストデータに使用している.教師データとテストデータの一部を表1に記す.

表 1: 各データで使用したオノマトペを含む句の例

教師データ	テストデータ
ほかほかなジャガイモ	とろけるチーズ
あっさりした味	ふんわりケーキ
ほんのり香る	こんがり焼いた
そっと落とす	どんどん作る
さっとゆでる	グツグツ煮立てる

3.1 文法による条件の設定

中国語の文法で修飾語を扱う場合,結构助词を使用する.结构助词とは,修飾語と被修飾語の間に使用する助詞のことである.被修飾語が名詞の場合は「的」,動詞の場合は「地」を使用する.この特徴に着目して,2つの条件を設定し共起率を求めることで文法と語句の翻訳が正しい句の選択可能となる.

条件1:対訳語句が3つの形態素で構成されており,结构助词を含む場合

结构助词以外の形態素の共起を求める.

例えば対訳語句がもつ形態素が単語Aと结构助词と単語Bの場合,AとBの共起を求め,その値を対訳語句の共起率として定める.

条件2:対訳語句が4つ以上の形態素で構成されており,结构助词を含む場合

条件2となる場合の例を図1に示す.结构助词を区切りに前半部分をA群,後半部分をB群として分ける.そして,パターン1やパターン2のようにA群とB群それぞれから一つずつ形態素を選択し,共起率を求める.求めた共起率の中で最も大きい値を句の共起率として定める.

渐渐(名词) 沥沥(名词) 的(结构助词) 炒饭(名词)
A　B

パターン1 : 渐渐(名词)　炒饭(名词)

パターン2 : 沥沥(名词)　炒饭(名词)

図1:条件2の例

条件1と条件2に当てはまらない語句は,その対訳語句が持つすべての形態素の共起を求める.

3.2 共起率の算出

本稿では Web ページ内にある単語とある単語が同時に出現する確率を共起率として扱う[3].例えば,「塩」と「胡椒」は同時に使用される頻度が多いため

Automatic Selection of Japanese-Chinese Translation for Onomatopoeia Phrases
Chisato Harada†, Kosuke Yamazaki†, Ailin Meng*, Wenyu Zhang*, Shiho Nobesawa†
†Faculty of Knowledge Engineering, Tokyo City University
*Dalian Jiangtong University / Tokyo City University

共起率が高くなるが,「塩」と「眼鏡」は同時に出現する頻度が少ないため共起率が低くなる.
もし,中国語訳として適切な場合は,その語句を使用しているWebページが多いため共起率が高くなり,適切でない場合はその語句を使用しているWebページが少ないため共起率が低くなると考える.そこで,本稿では共起率を用いて評価値を定める.共起率は式(1)で求められる[3].オノマトペと被修飾語を $a_1 \cdots a_n$ とし, $a_1 \cdots a_n$ のANDとORを検索キーワードとした時の,検索ヒット件数を $hit(X)$ と表す.検索エンジンは中国の検索エンジンの百度を使用する.

$$S(a_1,\cdots,a_n)=\frac{hit(a_1\cap\cdots\cap a_n)}{hit(a_1\cup\cdots\cup a_n)}\cdots(1)$$

3.3 評価値の算出

表2の品詞の組み合わせは,あらかじめ中国人留学生が評価を行った教師データから定義した.正しい対訳語句であると判断された教師データの句に含まれる5種類の品詞の組み合わせと,その他を加えた6種類を品詞の組み合わせとする.

表2: 品詞の組み合わせ一覧

記号	重み付けを行うパターン	例句(品詞の組み合わせ)
w_0	形谓词,结构助词,名词	热乎/的/马铃薯 (形谓词/结构助词/名词)
	形谓词,名词	
w_1	形容词,结构助词,名词	乱七八糟/的/理由 (形容词/结构助词/名词)
	形容词,名词	
w_2	副词,结构助词,动词	悄悄/丢掉 (副词/动词)
	副词,动词	
w_3	名词,结构助词,名词	咕嘟/咕嘟流 (名词/名词)
	名词,名词	
w_4	动词,结构助词,名词	打寒战/的/寒冷 (动词/结构助词/名词)
	动词,名词	

表2の品詞の組み合わせに基づく重み付けを行い,共起率にかけることで,最適対訳語句を取得しやすくする.
重み付けを行うには,適当な重みを設定し評価を行い,試行錯誤的に重みをチューニングする作業を,何度も繰り替えさなければならない.そのため,適切な重みを求めるには,多大な計算時間がかかる.そこで,遺伝的アルゴリズムを使用する.遺伝的アルゴリズムは,探索過程において,常に解の候補を保持しているため,実用的な時間にあわせた探索が可能となる.石川らの研究[5]にて提案されているアルゴリズムと重みパラメータの推定方法を参考に,品詞パラメータにかける重みを定義した.実験から得た数値の最大値で正規化した値を重みとして定義し,表3に記す.

表 3: 品詞　パターンの重み

記号	w_0	w_1	w_2	w_3	w_4	w_5
重みの値	0.79	0.71	0.02	0.22	0.17	1

重みの値は,1に近い値ほど取得したい品詞の組み合わせであり,0に近い値ほど取得しにくくしたい品詞の組み合わせまたは,使用頻度が非常に少ないとパターンであることが推測できる.
式(2)で評価値 P を定義した.共起率を S とし,共起を求めた品詞の組み合わせに基づく重みを W と表す.

$$P=S*W\cdots(2)$$

評価値Pが最も高い対訳語句を最適対訳語句として出力を行う.

4 実験結果

テストデータを用いて,下里らの手法と提案手法の比較実験を行った.下里らの手法と,本稿で提案した提案手法,本稿で使用した各機械翻訳機の正解率を表4に示す.正解率とは,中国人留学生が選択した最適対訳語句を,各機械翻訳機と各手法が出力した対訳語句に含む確率のことである.表4より,従来手法に比べて提案手法の方が正解率が 高いことが確認できる.そして,提案手法が各機械翻訳機における正解率を上回っていることも確認できる.

表4: 機械翻訳機の正解率との比較

	Excie	Yahoo	Google	Weblio	下里らの手法	提案手法
正解率	46%	61%	37%	37%	54%	65%

5 おわりに

本稿では,オノマトペを含む句を対象に,複数の日中機械翻訳機が出力した対訳語句から,共起を用いて最適な翻訳結果を選択する手法を提案した.実際の実験結果から,提案手法を用いることでそれぞれの機械翻訳機の正解率と従来手法の正解率を上回ることができた.

参考文献

[1] 独立行政法人日本学生支援機構, http://www.jasso.go.jp/,2017年
[2]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
[3]森 純一郎,松尾 豊,石塚 満,"Webからの人物に関するキーワード抽出",人工知能学会論文誌,20巻5号 C,pp.337-345,2005年
[4]小野 正弘,"擬音語・擬態語4500 日本語オノマトペ辞典", 小学館,2007年
[5]石川 英太郎,石田 崇,後藤 正幸,"評価関数の重みパラメータを推定する対話型遺伝的アルゴリズム",電子情報通信学会論文誌,Vol.J 94-D No.11,pp.1888-1898,2011年11月