

観光客に向けた情報提供のための技術的要素の検討

水嶋 雄里[†] 北村 森夫[‡] 黒瀬 英生^{††} 且 政宏^{††} 小坂 宏美^{††} 葛原 充洋^{††} 寺元 貴幸[†]
 津山工業高等専門学校[†] 株式会社ワードシステム[‡] 津山市役所^{††}

1 はじめに

近年、訪日外国人を含めた観光客数は増加傾向にある [1]。よって、観光客への情報提供を適切に実施することが、観光地にとってのサービス向上につながるといえる。

情報提供の質を高めるためにイベントやニュースなどの速報性のある情報を提供することは観光客にとって有用である。しかし、これには時間・人員・金銭等の面からコストがかかるという問題が想定される。

そこで、本研究ではオープンデータおよび webAPI 等を利用し、提供する情報の質を向上させるシステムのための技術的要素の検討を実施する。具体的には、速報性のある情報にさらに付加情報を与えた観光案内文を既存のデータから自動的に生成する手法を提案する。また、施設名などについて一意に翻訳される機械翻訳の手法を提案する。これらにより、観光案内文を複数の言語で生成することにより観光地にとっての低コストでのサービス向上を検討するものである。

2 観光案内文の生成手法の検討

2.1 観光案内文の生成手法

オープンデータから得られるデータと他の API を使用して得られるデータを連携して観光案内文を自動的に生成する手法を設計した。具体的には、定型文のフォーマットを作成してデータを埋め込むことで生成する。

2.2 観光案内文の翻訳手法

観光案内文では、案内に用いる施設名に公式の施設名との表記ゆれがないことが望ましい。翻訳文についても表記ゆれをなくすことを目的として、観光案内文の日英での機械翻訳の手法を設計した。図 1 に機械翻訳について設計したフローを示す。これは、機械翻訳において未知語が翻訳されることなく出力される処理を利用して翻訳済みの単語を埋め込むことができるという仮定に基づいている。

フローに基づいて、対訳コーパスから固有表現の対訳

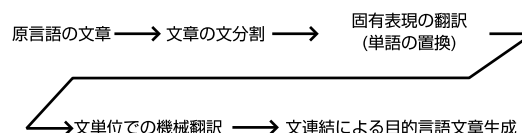


図 1 機械翻訳 システムフロー

単語対を抽出する手法を設計した。ヒューリスティックに基づく手法として、Dice 係数を用いた対訳単語対の抽出を用いることとした。式 (1) に Dice 係数の計算式を示す [2]。

$$Dice(X, Y) = \frac{2 \cdot f_{XY}}{f_X + f_Y} \quad (1)$$

$$(0 \leq Dice(X, Y) \leq 1)$$

f_X, f_Y はそれぞれ単語 X, Y が独立に出現する頻度である。また、 f_{XY} は単語 X, Y が対訳文間で同時に出現する頻度である。Dice 係数の値に閾値を設定し、これを超えるものについて単語アライメントすることで対訳単語対を抽出することができる。このとき、対訳抽出すべき単語は施設名等の固有表現であるため該当しない単語は閾値を超えた場合でも除去する。

ただし、小規模の対訳コーパスから施設名などの出現頻度の低い単語の対訳関係を推定することは難しい。よって、単語の頻度を計算する処理を実行する際に使用する対訳コーパスと Dice 係数を計算して対訳単語を抽出する対訳コーパスのそれぞれに分割する。単語の頻度を計算する処理を実行する際に、使用する対訳コーパスは対訳単語を抽出する対訳コーパスを基本としてデータを増やすことで作成する。具体的には、まず類語辞書を用いて辞書に既存の単語の対訳関係からこれらの単語を類語に置換して対訳コーパスをのデータを増やす。次に、対訳を抽出する対象の対訳コーパスに類似するドメインの対訳コーパスを用意して対訳コーパスを結合する。

3 観光案内文の生成実験

3.1 観光案内文の生成手法

観光案内文の生成については、イベント会場の住所と開催日の天気予報の付加情報を与えた観光案内文を生成する。観光地が保有するデータには、岡山県津山市が公

Discussion of the way to serve tourism information

[†]Yuri Mizushima, Takayuki Teramoto · National Institute of Technology, Tsuyama College

[‡]Morio Kitamura · WordSystem Co., Ltd.

^{††}Hideo Kurose, Masahiro Dan, Hiromi Osaka, Mitsuhiro Kuzuhara · Tsuyama City Office

表1 観光案内文のフォーマットと生成結果

フォーマット	生成結果
{ イベント名 } は, { 開始日時 } { 開始時間 } に { 開催場所 }(住所: { 開催場所住所 }) で開催されます. この日の天気は, { 開催場所開催日時の天気予報 } です.	津山まなびの鉄道館レンタサイクル使用開始は, 2018年4月1日9時0分に津山まなびの鉄道館(大谷)(住所: 日本、〒708-0882 岡山県津山市大谷)で開催されます. この日の天気は, 昼過ぎまで曇りです.

表2 Dice 係数の閾値を 0.6 としたときの翻訳文と参照訳の例

種類	文
本システム	Thank you very much for visiting <u>Tsuyama Railroad Educational Museum</u> today.
他システム	Thank you very much for visiting <u>Tsuyama Manabi's railway hall</u> today.
参照訳	Thank you for visiting the <u>Tsuyama Railroad Educational Museum</u> today.

開するオープンデータ^{*1}を使用する. 表1にフォーマットと生成した観光案内文を示す.

3.2 観光案内文の翻訳手法

対訳コーパスを増やすために用いる対訳コーパスには, Wikipedia 日英京都関連文書対訳コーパス^{*2}のうち鉄道(交通関連)のカテゴリから 14,002 文対を用いる. 対訳単語を抽出する対訳コーパスには, 岡山県津山市が保有する観光施設の情報等を多言語に翻訳した文章・音声を表示する多言語音声ガイドシステム「つやま声ナビ」^{*3}に収録されている対訳データのうち津山まなびの鉄道館に関する 72 文対を用いる. 翻訳を実験するための原言語文・参照訳は, 「つやま声ナビ」に収録されている対訳データのうち津山まなびの鉄道館に関する URL のみの文対を除いた 70 文対を用いる. ただし, 文単位での機械翻訳は既存の翻訳システム^{*4}を用いる. また, 他システムによる翻訳として既存の翻訳システムに原言語の文を入力として与えた場合の結果を用意する. 表2に例として Dice 係数閾値を 0.6 として設定した場合の翻訳結果を示す. このとき, 下線部は施設名である.

翻訳の質を評価する手法として, 翻訳編集率 (Translation Edit Rate; TER) を評価する [3][4]. 表3に翻訳編集率を示す. このとき, Dice 係数の閾値を変数として測定した.

翻訳の例文から, 参照訳のうち施設名であるは本システムでは考慮されて一意に翻訳されていることがわかる. また, 本システムと他システムの翻訳編集率を比較すると本システムがわずかに低い値を示していることがわかる. よって, 観光案内文において本システムは施設

表3 それぞれの Dice 係数閾値における翻訳編集率

Dice 係数閾値	本システム	他システム
0.2	0.7722	0.7788
0.4	0.7746	0.7788
0.6	0.7752	0.7788
0.8	0.7728	0.7788

名を考慮しており翻訳編集率をわずかに下げているといえる.

4 おわりに

観光案内文の生成では, オープンデータを用いて定型文を生成できることを確認した. 観光案内文の翻訳では, 抽出する単語について置換された目的言語の翻訳文が生成されることを確認した. 既存の翻訳システムに単語置換した文を入力する手法は, 翻訳システム側から見ると未知語の処理がおこなわれた結果といえる. よって, 言語としての自然さに難点がある. また, 生成文・翻訳文ともに言語としての自然さを評価することで向上するための指標とすることが必要である.

参考文献

- [1] 日本政府観光局: 訪日外客統計, 日本政府観光局 (オンライン), 入手先<https://www.jnto.go.jp/jpn/statistics/data_info_listing/index.html> (参照 2018-12-31).
- [2] 北村美穂子, 松本裕治: 対訳コーパスを利用した対訳表現の自動抽出, 情報処理学会論文誌, Vol.38, No.4, pp.727-736 (1997).
- [3] 渡辺太郎, 今村賢治, 賀沢秀人, Neubig, Graham, 中澤敏明: 機械翻訳, pp.37-60, コロナ社 (2014).
- [4] Matthew Snover, Bonnie Dorr, Richard Schwartz, et al.: A Study of Translation Edit Rate with Targeted Human Annotation, Proceedings of Association for Machine Translation in the Americas, 2006.

^{*1}<https://www.city.tsuyama.lg.jp/life/index2.php?id=5530>

^{*2}<https://alaginrc.nict.go.jp/WikiCorpus/index.html>

^{*3}<https://www.city.tsuyama.lg.jp/schedule/detail.php?id=15939>

^{*4}<https://translate.google.com/intl/ja/about/>