オープンデータおよびwebAPI 等を利活用した観光案内サービスの技術的要素の検討

寺元　貴幸*　水嶋　雄里**

Discussion of technical elements related to new services utilizing Tsuyama City Open Data and webAPI

Takayuki　TERAMOTO, Yuri MIZUSHIMA

These days, foreign tourist to Japan is increasing. According to Japan National Tourism Organization, their year-over-year growth rate is 19.3%. In other words, tourist including them is increasing. Hence, to serve tourism information is needed to them and its necessity is increasing. In Tsuyama City, tourist is increasing and increasing them is considered to keep continuously. Thus, considering to grow up service to them is necessary. Therefore, to serve it correctly can reach to grow up service for tourist spot.

As stated above, considering to grow up service quality. For instance, to serve events or news information is useful for it. Furthermore, to serve tourism information as some languages is useful. In contrast, to serve it is considered costly. Currently, Tsuyama City serves “Tsuyama Koe Navi” where is served tourism information as text and voice in some languages. However, adding many tourism information is difficult due to a cost of development and operation.

We investigate the way to grow up service quality for tourist spot using open data and webAPI where is served in Tsuyama City. In particular, first we investigate the way to generate tourism information with some data and to add useful information. Second we investigate the way to translate it correctly. Finally we investigate the way to generate tourism information as some languages automatically by these two techniques.

Key Words:　Open Data,　Web API,　Tourism Information,　Multilingual

1．緒　　言

近年、訪日外国人数は増加傾向にある。日本政府観光局によると、2017 年の訪日外国人数の前年比伸び率は 19.3%であり、訪日外国人を含めた観光客数は増加傾向にあるといえる[1)]。よって、訪日外国人に向けた情報提供の必要性は高まっているといえる。津山市でも、観光客数は増加傾向にあり[2)]、今後も引き続き増加していくことが予想される。このため、観光客へのサービス向上を考えることが不可欠である。よって、訪日外国人を含めた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。

上記のように、提供する情報の質を向上させることを考える。例えば、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有用である。他にも、訪日外国人を含めた多言語による情報提供も同様である。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。津山市は、多言語音声ガイドシステム「つやま声ナビ」により多言語観光案内をおこなっている。しかし、開発・運用にかかる費用を考慮すると大量の観光案内文の作成は難しいのが現状である[3,4)]。

多言語翻訳としてコストの低い方法は機械翻訳システムを利用することである。近年の研究により、目的言語への翻訳の質は向上している。また、多言語に向けた翻訳が可能であるシステムも多く存在する。しかし、施設名等の固有表現の翻訳については多くの問題がある。具体的には、一般的な固有表現については機械翻訳システムで考慮されることも多いが、新語や地域的な固有表現についてはシステムが公式に定める対訳を考慮することは難しい。これにより、固有表現が一意に翻訳されない可能性

原稿受付　令和 1 年 9 月 26 日
＊総合理工学科　情報システム系
＊＊情報工学科　平成 31 年 3 月卒業

があるという問題がある。

そこで、本研究では津山市により公開されている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検討を実施する。具体的には、まず公開されているデータをもとに観光案内文を生成し、その際に有用な付加情報を与えることができないかを検討する。次に、生成された観光案内文や既存の観光案内文を多言語翻訳し、その際に有効な翻訳をおこなうことができないかを検討する。これらの技術を統合してデータを入力として多言語で観光案内文が生成されるために必要な技術を検討する。

結果として、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した。また、多言語翻訳への対応の前段階として日英での機械翻訳の手法を検討した。これにより、施設名等の固有表現について一意に翻訳されるよう検討したので本論文で報告したい。

２．観光案内文の生成

２．１　オープンデータの利用

総務省によると、オープンデータの定義は国、地方公共団体および事業者が保有する官民データのうち国民誰もがインターネット等を通じて容易に加工、編集、再配布等できるよう下記の要件を満たす形式で公開されたデータのことであるとされている[5)]。

- 営利目的、非営利目的を問わず二次利用可能なルールが適用されたもの
- 機械判読に適したもの
- 無償で利用できるもの

津山市では、いくつかオープンデータが公開されている[2)]。ここでは、津山市の施設の位置情報、会計情報等のデータが公開されている。表１に津山市の公開するオープンデータの種類を抜粋して示す。また、定期的に更新されるデータも公開されている。例えば、津山市内で開催されるイベント情報のように月ごとに更新されるデータがある。これらを利用することで、速報性の高い情報として活用することもできる。

表１　社会性の因子構造

コンテンツ名	ライセンス
津山市イベント情報	cc-by
津山市ごみの分別・収集	cc-by
津山市リージョンセンター利用状況	cc-by
津山市人口動態	cc-by

２．２　webAPIの利用

津山市では、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」を保有している[3)]。このシステムではwebAPI(webApplication Programming Interface)を構築しており、施設情報・目的言語を入力として施設情報の文章を得ることができる。よって、APIから得られる複数言語の情報を資源として利用することができる。本研究では、津山市よりAPIのアクセス権限を提供いただくことにより研究に利用することとした。

「つやま声ナビ」で構築されているwebAPIは、Salesforce社[6)]のシステム上で構成されており、APIアクセスのプロトコルにSOAPを使用している。また、APIによるデータ取得のためにはまず認証処理が必要である。具体的には、認証用APIの要求をすることでトークンを取得する。認証用APIからの応答で得られたトークンを付与してデータ取得用APIの要求をすることで施設情報等の文章を取得することができる。津山市からAPI・認証情報が記述されたWSDL(WebServices Description Language)ファイルを提供されたため、これをもとにAPIにアクセスした。表２に、APIから抜粋して取得した施設情報文章を示す。

表２　webAPIから抜粋して取得した施設情報文章

日本語	英語
旧津山扇形機関車庫は、昭和11年（1936）に建設されました。奥行22.1mで17線あり、現存するものでは京都の梅小路に次ぐ国内2番目の大きさです。全国で現存している扇形機関車庫は数少なく、県内では津山にあるだけです。	The Old Tsuyama Fan-Shaped Locomotive Depot was constructed in1936. The shed has a depth of22.1meters and there are17tracks. It is second in size only to Kyoto's Umekoji depot, ofthe railway roundhouses still existing in Japan. Railway roundhouses exist only in a small number oflocations around Japan, including Tsuyama,which is the only one in Okayama Prefecture.

使用するwebAPIは要求数に１日あたり30,000回との制約があるため、API応答によって得られたデータをローカルのデータベースに保存する処理を行った。データベースには MongoDB[7)]を使用した。これにより、研究に使用する施設情報等の文章をwebAPI の要求数を考えることなく実行することが可能となった。

２．３　観光案内文の自動生成検討

次に観光案内文の自動生成を検討した。2.1節にて述べた津山市が公開するオープンデータからイベント情報に関するデータを取得する。データは

CSV形式で公開されている。このデータを使用することで、イベントに関する観光案内文を生成することができる。加えて、データに付加情報を加えることで速報性のある情報をもとに、より有用な観光案内文を生成することができないか検討した。図1に、観光案内文を生成するためのフロー概略を示す。具体的には、データからイベント名、開始日時、開始時間、開催場所を抽出することで観光案内文を生成することとした。まず、開催場所には施設名が記述されているため、この情報から住所・経緯度を取得する。住所・経緯度の取得にはGoogleMaps PlatformからGooglePlaces API[8]を使用した。次に、開催場所の経緯度、開始日時、開始時間からイベント開催時の天気予報を取得する。天気予報の取得にはDarkSkyAPI[9]を使用する。取得したデータをもとに定型文を生成する。定型文のフォーマットをもとに、イベント情報に関するオープンデータからイベント情報を案内する文章を生成するプログラムを作成した。表3に定型文のフォーマットと生成結果の例を示す。ただし、中括弧の要素は取得したデータ名であり、これを挿入することで定型文を生成する。これにより、入力としてオープンデータを与えることで各種 API を通して得たデータを付加して観光案内文を自動生成することができる。

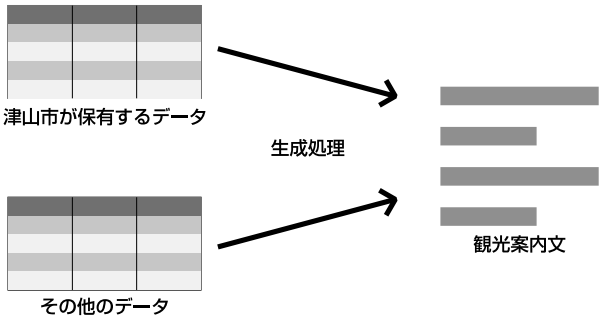

図1　観光案内文生成フロー

表3　観光案内文のフォーマット と生成結果の例

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。

また、生成された観光案内文をどのように利用者に提供するかについて、音声による出力を検討した。具体的には、テキスト音声合成システムであるOpen JTalk[10]を使用して音声出力をテストした。このとき、Open JTalk で使用している辞書の影響で固有名詞について正しい読みで出力されない問題が発生した。対策として、Open JTalkに出力する文章のうち固有名詞をひらがなに置換するよう処理した。読みの置換には、形態素解析エンジンであるMeCab[11]を使用した。また、辞書として mecab-ipadic-NEologd[12]を使用した。これにより、Open JTalkに限らず音声合成ソフトの固有名詞処理に依存しない文章を生成することができる。

3．機械翻訳の調査

3．1　人手による翻訳

観光案内文章を翻訳する手法として、まず人手による翻訳を考える。翻訳は複数の言語間での知識を要する作業であるため、知識を持たない人が翻訳することは困難である。したがって、翻訳知識を持つ人に依頼する必要がある。知識を持つ人により翻訳する場合、知識をもとに目的言語への自然な翻訳をすることができる。また、対訳辞書に沿って翻訳することで一意的な翻訳にすることもできる。しかし、これには金銭の面からコストが掛かることが想定される。つやま声ナビで提供されている多言語のデータは翻訳家によって翻訳されている。翻訳の際には、翻訳業者への委託費用が掛かった[4]。

また、人手による翻訳をする場合は時間の面でもコストが掛かることが想定される。よって、イベント情報などの速報性の高い文章を人手により翻訳して公開することはさらに困難である。したがって、観光案内文章の人手による翻訳は適切ではないといえる。

3．2　既存の機械翻訳システム

観光案内文章を翻訳する手法として、次に既存の機械翻訳システムによる翻訳を考える。翻訳システムを使用することで、翻訳に掛かる時間の面でのコストを下げることができる。翻訳の目的言語における流暢さについても、近年の研究により向上しているといえる[13]。しかし、施設情報を翻訳する際に固有表現をどのように翻訳することが適切であるかに疑問が残る。具体的には、固有表現を翻訳する際に目的言語との形態素単位での対訳が取れていることが重要であるか、公式の表現をもとにしてこれが言語中で統一されていることが重要であるかとする点である。観光案内文章を翻訳する際に重要となるのは、施設名等の固有表現は言語中で統一であることと仮定する。なぜならば固有表現に表記ゆれが存在した場合、異なった施設名を案内することと等しいからである。

既存の機械翻訳システムにおいて、一般的な翻訳単位は形態素である[14]。固有表現は、1 以上の形態素からなる。よって、固有表現が複数の形態素からなる複合語である場合がある。固有表現が複合語で構成されている場合、既存の機械翻訳システムは目標とする固有表現に翻訳されるとは限らない。よって、観光案内文章の既存の機械翻訳システムによる翻訳を考える際には固有表現を形態素に分解しない単位で対訳付けることが必要となる。

３．３　単語アライメント

対訳コーパスをもとにした対訳表現の抽出方法としてヒューリスティックに基づくモデル、統計モデルが挙げられる。ヒューリスティックに基づくモデルとしては、$Dice$係数を

$$Dice(X,Y) = \frac{2 \cdot f_{XY}}{f_X + f_Y}$$
$$(0 \leq Dice(X,Y) \leq 1) \qquad (1)$$

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。

統計モデルとしては、IBM モデルがよく知られている[15]。IBM モデルは，統計的機械翻訳によるモデルである。

３．４　機械翻訳の自動評価

機械翻訳システムを自動評価する手法としては様々なものがある。自動評価手法として必要となるのは、原言語文を翻訳者が実際に翻訳した参照訳である。参照訳と翻訳文を比較してどれだけ近いかを計算することで評価する[16]。このときの文の近さは、表層的な観点からみた単語等の一致である場合もあれば意味的な観点からみた類似である場合もある。

自動評価手法の例として、適合率に基づく手法である BLEU、再現率に基づく手法である翻訳編集率がある。BLEU は、基本的な考えとしては翻訳文の n-gram のうちどれだけが参照訳の n-gram と一致するかという適合率をもとに測定するものである。翻訳編集率(Translation Edit Rate; TER)は翻訳文を参照訳に編集する際にどれだけ編集する必要があるかということを編集率として表すものである[16,17]。具体的には、編集の種類として単語単位の挿入、削除、置換、単語あるいは句単位でのシフトを定義する。参照訳をr、翻訳文をeとしてこれらの回数をそれぞれ$ins(r,e)$、$del(r,e)$、$sub(r,e)$、$shift(r,e)$とする。式(2)に，翻訳編集率の式を示す。ただし、$|r|$は参照訳の単語数である。

$$TER(r,e) = \frac{ins(r,e) + del(r,e) + sub(r,e) + shift(r,e)}{|r|} \qquad (2)$$

翻訳文が複数存在する場合は、これを全体に渡って計算して相加平均を取る。翻訳編集率は、翻訳を人手により編集することを前提に考えたときの評価尺度として用いられる。

４．観光案内文の翻訳

４．１　目的とする機械翻訳の整理

機械翻訳は、生成した観光案内文や既存の観光案内文について目的言語に翻訳するために利用する。観光案内文の機械翻訳についてまず評価すべき点は、施設名等の固有表現について一意に翻訳されることであるといえる。次に、翻訳文で原言語文の内容の一部または全部が欠落することなく流暢に目的言語に翻訳されていることである。しかし、既存の翻訳システムでは固有表現について一意に翻訳されることが保証されていない。したがって、固有表現を考慮した翻訳システムの検討が必要であるといえる。

対訳辞書の構築に基づく翻訳と比較して異なるのは、対訳コーパスを用いる点である。「つやま声ナビ」で構築されている webAPI のように、津山市には対訳コーパスとして利用できるデータが存在している。よって、対訳辞書を人手により構築することに比べて対訳コーパスを用意することの方がコストは低い。したがって、目標は対訳コーパスをもとに固有表現を考慮し翻訳を獲得する機械翻訳を構築することであるといえる。

４．２　機械翻訳の設計

次に機械翻訳のシステムフローについて設計した。図 2 に設計したシステムフローを示す。従来の機械翻訳と異なり、機械翻訳システムに入力する前に固有表現について目的言語の対訳に置換する。これにより、機械翻訳システムを改良することなく固有表現を考慮した翻訳ができると考えられる。しかし、目的言語の利用しやすい文章に翻訳されるかについては、実験をもとにした評価が必要である。ここでの利用しやすさとは、施設名等の固有表現について翻訳された文章を修正する必要がないといった後編集のしやすさとする。

機械翻訳のシステムフローから、まず必要となるのは文への分割処理である。これは、各言語における文の特性を用いる。具体的には、日本語における

句点、改行に該当する特性を用いる。次に必要となるのは、固有表現の対訳への置換処理である。これは、固有表現の対訳を必要とする。よって、対訳コーパスから固有表現の対訳を抽出する手法を検討する必要がある。最後に必要となるのが、機械翻訳処理である。これは、既存の機械翻訳システムを用いる。

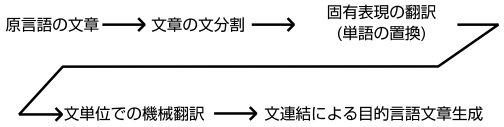

図 2　機械翻訳システムフロー

４．３　対訳単語対抽出手法の検討

固有表現の対訳単語対を抽出する手法について検討する。単語アライメントを*Dice*係数によっておこない、対訳単語対を抽出する。このために、*Dice*係数の計算によって得られた係数値から、設定した閾値を超えるものについて対訳単語対とみなすこととする。また、単語の出現頻度が言語間で低い場合は相対比の関係で係数値の取りうる値の間隔が広い。よって、対訳でないとすべき単語についても係数値が上昇してしまう可能性がある。これにより、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正確に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。

また、対訳コーパスの単語アライメントは分かち書きされた単位でおこなう。したがって、施設名等の固有表現については分かち書きしないよう処理する必要がある。このため、本研究において分かち書きは形態素を複数含むことがあり得る。

４．４　対訳コーパスを用いた対訳単語対自動抽出の実験

実験では OS として macOS　10.13.6[18]を開発言語として Python 3.5.2[19]を使用し、対訳コーパスとして Wikipedia 日英京都関連文書対訳コーパス [20]を用いた. 対訳コーパスには, 15 のカテゴリについて日本語と英語の文アライメント済みで約 50 万文対収録されている。本実験では、このうち鉄道(交通関連)のカテゴリに属する対訳文を先頭から 100 文対選択して入力する。 表 4 に抽出した対訳文対を示す。また、対訳コーパスを増やすために用いる対訳コーパスは、鉄道(交通関連)のカテゴリ に属する対訳文から 14,002 文対である。

表 4　対訳コーパスから抽出した対訳文対

日本語	英語
そのために主要観光地へは京都市営地下鉄東西線に乗り換えるか、 地下鉄の各駅から京都市営バス・京都バスなどの路線バス利用となることが多い。	So they usually change over to the Kyoto Municipal Subway Tozai Line or take local buses such as Kyoto City Bus, Kyoto Bus,etc from each subway station.
湖西線 ※正式には近江塩津駅- 山科駅間だが、すべての列車が当駅に直通している。	Kosei Line*Although this line officially covers the section between Omi-Shiotsu Station and Yamashina Station,all trains come directly to this station.

対訳単語対を抽出する前に, 対訳コーパスに事前処理をする。具体的には、まず、URI (Uniform Resource Identifier)、電話番号、括弧等の表現について削除する。次に形態素解析をおこない、日本語については分かち書きする。分かち書きには MeCab[11]を使用した。また、辞書として mecab-ipadic-NEologd[12]を使用した。次に、日本語・英語のコーパスについて自立語のみを抽出する。最後に、施設名等の固有表現と推定されるものについてタグを付けて結合する。具体的には、日本語については専門用語抽出システム [21]によって抽出された用語について分かち書きした単語を連結する。英語については、名詞句を連結する。名詞句の検出には、NLTK(Natural Language Toolkit)[22,23]を使用した。事前処理をしたコーパスを、*Dice*係数をもとに 4.3 節で示す手法で単語アライメントをおこない、対訳単語対を求めた。このとき、*Dice*係数の閾値を変数として測定した。表 5 に対訳単語対の抽出数を示す。ただし、コーパスを増やして計算した場合とコーパスを増やすことなく計算した場合を比較して示す。

表 6 に例として*Dice*係数の閾値を 0.6 とした場合の実験に用いたデータと抽出結果の関係を示す。ただし、正解数を求めるために人手による評価を実施した。評価基準は、コーパスのドメインをもとに辞書に記載されている意味的に類似しているとい

えるものについて正解と評価している。また、単語の訳抜けについては正解ではないと評価している。

表5 対訳単語対の抽出実験の結果

*Dice*係数閾値	コーパスを増やした場合	コーパスを増やさない場合
0.2	65	79
0.4	59	81
0.6	54	82
0.8	33	74

表6 実験に用いたデータと *Dice*係数の閾値を0.6とした場合の抽出結果からの評価

言語	コーパスを増やす処理	分かち書き数	抽出数	正解数	適合率[%]	再現率[%]
日本語	あり	1,036	54	28	**51.9**	**5.21**
日本語	なし	1,036	82	31	**37.8**	**7.92**
英語	あり	806	54	28	**51.9**	**6.70**
英語	なし	806	82	31	**37.8**	**10.2**

適合率について着目すると、*Dice*係数の閾値を0.6とした場合ではコーパスを増やす処理をすることが向上に寄与していると考えられる。これは、単語の出現頻度についてコーパスを増やすことによってそれぞれ増加させることができていたためと考えられる。よって、本手法は適合率を上昇させる可能性があることを示唆している。再現率について着目すると、コーパスを増やすことで再現率が下がることがわかる。先行研究[15]においては閾値の設定によって再現率が変化することが知られている。本研究においては施設名等の固有表現を抽出できればよいため、再現率を上昇させることが必ずしも目的を満たすとは限らない。先行研究[15]においては、適合率と再現率はトレードオフの関係であることが知られている。よって、今後は翻訳文が利用しやすくなる方向に寄与する閾値を選択することが望ましい。

ところで、分かち書きの部分については、言語特性への依存をしている。よって、分かち書きについて対訳言語間で同等の意味を持つ単位で処理することが必要であるといえる。これは、分かち書きについて対訳言語間で同等の意味を持つ単位で処理することができれば、特定の言語によらない複数の言語間での翻訳を構築できることを示唆している。

4．5　自動抽出した対訳単語対を用いた機械翻訳の実験

4.3節で示す手法で対訳単語を抽出し、4.2節で設計した手法に沿って機械翻訳を実験する。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。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[20]のうち鉄道(交通関連)のカテゴリから14,002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 URI 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[24]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7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。

4.2節で定義した翻訳文の利用しやすさを評価するために、翻訳の質を評価する手法のひとつである翻訳編集率を用いて評価する。表8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。

表7 *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
原言語	本日は津山まなびの鉄道館へご来館いただき誠にありがとうございます.
本システム	Thank you very much for visiting Tsuyama Railroad Educational Museum today.
他システム	Thank you very much for visiting Tsuyama Manabi’s railway hall today.
参照訳	Thank you for visiting the Tsuyama Railroad Educational Museum today.

表8 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

翻訳の例文から、 参照訳のうち施設名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。

５．おわりに

本研究は、津山市により公開されているオープンデータおよび webAPI 等を利用し、サービスの質が高い情報提供をするシステムのための技術的検討を目的としている。

公開されているデータをもとに観光案内文を生成する際に有用な付加情報を与えることについては、イベント情報の観光案内文に開催日の天気予報と開催場所の住所を与えて生成することができた。これにより、データをもとに与える付加情報をさらに検討することでより情報提供のサービスの質を高めることができる。また、情報提供のために観光案内文を人手により作成することに代わって、自動的に生成することができた。これにより、人員の面でのコストを削減することが期待できる。観光案内文の利用者への出力方法としては、音声による出力を検討した。音声による出力の場合、音声合成ソフトによって固有名詞の読みを誤る場合があることが確認できた。この対策として、入力文章の固有名詞について辞書をもとにひらがなに置換処理した。これにより、音声合成ソフトの用いる固有名詞の表現に依存しない出力ができた。

ただし、利用するデータの種類を変えることでイベント情報に限らない情報の提供ができることが望ましい。よって、情報提供に利用可能なデータについて引き続き調査する必要がある。

観光案内文のための翻訳に必要な要件について検討することで、施設名等の固有表現を重視する翻訳が必要であることを仮定した。また、施設名等の固有表現を重視する翻訳のシステムフローについて検討した。加えて、施設名等の固有表現の対訳を対訳コーパスを用いて自動的に抽出する手法について検討して実験した。これにより、対訳辞書を自動的に構築することができる。多言語に向けた翻訳の前段階として、日英による翻訳手法として検討して実験した。また、検討手法を多言語に展開する際の問題点について言語特性に依存する点がどこであるか検証した。これにより、多言語翻訳の可能性を検討することができた。

今回研究した各要素技術によって、観光案内文を多言語に生成するシステムを実際に構築することがある程度可能であることが確認できた．　図 3 に今後目指すべき観光案内文の生成と翻訳による全体のフローを示す。

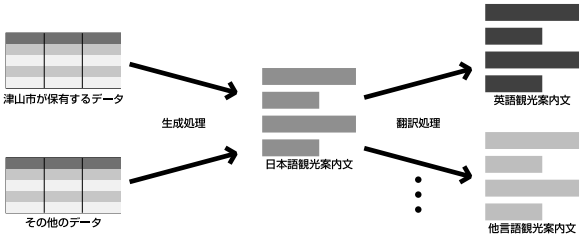

図 3　機械翻訳システムフロー

本研究は、津山市のデータに基づいて津山市の観光向け情報提供のサービスの品質を向上させる手法の検討を行った。また、本研究は同様の手法はデータさえ提供されていれば別の自治体等のデータに応用することも可能である。これにより、汎用的なシステムとして展開することも期待できると考えている。

謝　辞

データの提供や実験へのご協力ご討論頂いた株式会社ワードシステムの北村森夫氏、津山市役所の黒瀬英生氏、旦政宏氏、小坂宏美氏、葛原充洋氏に感謝する。

参　考　文　献

1) 訪日外客統計|日本政府観光局 (JNTO) :https://www.jnto.go.jp/jpn/statistics/data_info_listing/index.html (2019.2.1).
2) 津山市:https://www.city.tsuyama.lg. jp/life/index2.php?id=5530.
3) 多言語音声ガイドアプリ 「つやま声ナビ」運用開始:https://www.city.tsuyama.lg.jp/schedule/detail.php?id=15939(2019.2.1).
4) 津山市多言語音声ガイドシステム導入事業に係る公募型プロポーザルの審査結果:https://www.city.tsuyama.lg.jp/common/photo/free/files/10166/20171012110609099078.pdf (2019.2.1).
5) 総務省｜ICT 利活用の促進｜オープンデータ戦略の推進:http://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/index.html(2019.2.1).
6) Salesforce-セールスフォース・ドットコム:https://www.salesforce.com/jp/ (2019.2.1).
7) Open Source Document Database|MongoDB:https://www.mongodb.com/ (2019.2.1).
8) Google Maps Platform-Geo-location API|Google Maps Platform|Google Cloud:https://cloud.google.com/maps-platform/?hl=ja(2019.2.1).
9) Dark Sky:https://darksky.net/dev(2019.2.1).
10) Open JTalk: http: //open-jtalk.sourceforge.net/

(2019.2.1).

11) MeCab: Yet Another Part-of-Speech and Morphological Analyzer:http://taku910.github.io/mecab/(2019.2.1).

12) neologd/mecab-ipadic-neologd:Neologism dictionary based on the language resources on the Web for mecab-ipadic:https://github.com/neologd/mecab-ipadic-neologd(2019.2.1).

13) 中澤敏明:機械翻訳の新しいパラダイム:ニューラル機械翻訳の原理，情報管理，Vol.60，No.5，pp.299-306(2017).

14) 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).

15) Brown,Peter F.,Pietra,Vincent J.Della,Pietra,Stephen A.Della,et al.:The Mathematics of Statistical Machine Translation :Parameter Estimation, Computational linguistics, Vol.19,No.2,pp.263-311(1993).

16) 渡辺太郎，今村賢治，賀沢秀人，Neubig,Graham，中澤敏明:機械翻訳，pp.1-60，コロナ社(2014).

17) 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,Vol.200,No.6 (2006).

18) macOS High Sierra- 技術仕様：https://support.apple.com/kb/SP765?locale=ja_JP (2019.2.1).

19) Welcome to Python.org: https://www.python.org/(2019.2.1).

20) Wikipedia 日英京都関連文書対訳コーパス :https://alaginrc.nict.go.jp/WikiCorpus/index.html (2019.2.1).

21)“専門用語（キーワード）自動抽出システム”のページ :http://gensen.dl.itc.u-tokyo.ac. jp/(2019.2.1).

22) Natural Language Toolkit―NLTK3.3documentation:https://www.nltk.org/(2019.2.1).

23) 7.Extracting Information from Text:https://www.nltk.org/book/ch07.html(2019.2.1).

24) Google 翻訳:https://translate.google.com/intl/ja/about/(2019.2.1).