Vol.2012-NL-206 No.2
Vol.2012-SLP-91 No.2
2012/5/10

外国人の検索クエリに対する音訳手法の適用

辻理絵子† 木村健† 古宮嘉那子† 小谷善行†

外国人の商外国人が日本語のショッピングサイトを利用する際日本語の商品名が分からず，商品が見つけられないことがある．本研究では，実際に外国人が日本のショッピングサイトで商品検索の際に失敗したクエリとその正しいクエリの対のコーパスを用いて，非日本語圏のユーザによる検索クエリの音訳を行った．ペアコーパスには，意訳によって修正されたクエリのように，音訳するにはノイズとなるデータが含まれているため，文字種によるフィルタリングを行って学習データを絞り込んだ．さらに，BIGRAM,HMM, CRF の 3 つの機械翻訳手法を比較した結果，検索クエリの音訳では HMM 手法が最適であった．

Transliteration of Alphabet Queries in Japanese Shopping Site

RIEKO TSUJI† TAKESHI KIMURA†
KANAKO KOMIYA† YOSHIYUKI KOTANI†

There are some cases where the non-Japanese buyers are unable to find products they want through the Japanese shopping Web sites because it requires Japanese queries. We propose to transliterate the inputs of the non-Japanese user, i.e., search queries written in English alphabets, into Japanese Katakana to solve that problem. In this research, the pairs of the non-Japanese search queries which failed to get the right match obtained from a Japanese shopping website and its transcribed word given by volunteers were used for the training data. Since this corpus includes some noises for transliteration such as free translation, we used two different filters to filter out the query pairs that are not transliterated in order to improve the quality of the training data. In addition, we compared three methods, i.e., BIGRAM, HMM, and CRF, using these data to investigate which transliteration method is the best for query transliteration. The experiment revealed the HMM was the best.

1. はじめに

近年，多くの商品がインターネットを介して入手できるようになり，外国人が日本の商品検索サイトを利用する機会も増えている．外国人は日本語の商品名が分からない場合，英単語など，アルファベットの検索クエリを利用するが，そのクエリの検索結果が 0 件となった場合，その時点で外国人ユーザの商品購入は失敗する．このとき，検索クエリ中に現れるアルファベットの検索クエリを日本語商品名に適切に翻訳することができれば，ユーザビリティの向上に寄与する．そこで，我々は検索エンジン側から海外ユーザの検索クエリを修正し，提案することで外国人ユーザの勝因購入を支援することを考えた．実際の外国人の検索クエリのうち検索結果 0 件となったものには，次のように様々な要因がある．

1a. アルファベットで検索(英単語など)
[e.g.,lunch box→ランチボックス， mouse pad→マウスパッド]

1b. アルファベットで検索(日本語を書き表したもの)
[e.g., yukata→ゆかた， tanabata→七夕]

2. 翻訳しなければならないもの
[e.g., work dress→普段着]

3. 特殊系(特殊用語，固有名詞，アニメ関連用語)
[e.g., sanyo-GoPan→サンヨーGOPAN， K-On!→けいおん！]

しかし，この中でも音訳により解決する場合が多かったため，本研究では、音訳を用いてアルファベットから日本語クエリへと修正することを提案する。これまで音訳研究は整形されたデータセットに対して行われてきており (Min et al., 2011) [9]，我々が知る限り、ノイズを含むデータを用いて音訳を行った研究はない。本論文では，外国人が実際に日本の商品販売サイトで，情報検索を行った際に一致しなかったもの(以下，修正前クエリと呼ぶ)と，修正前クエリをボランティアの方が人手でマッチするように変更したクエリ(以下，修正後クエリと呼ぶ)が対となったコーパス(以下，ペアコーパスと呼ぶ)を使用することにより，クエリのマッチングにはどのような音訳が適しているかを調べた。 BIGRAM，HMM，CRF を比較したところ，HMM 手法が一番すぐれていた．

2. ペアコーパス

表 1 に実際のペアコーパスの例を示す．ペアコーパスにはこのようなものが 4574 レコード分含まれる．

† 東京農工大学
Tokyo University of Agriculture and Technology

表 1　ペアコーパスの例

修正前クエリ	修正後クエリ
fashion	ファッション
lolita	lolita
lolita -jeans	ロリータジーンズ
ポケモンカード	ポケモン カード
NARUTO 疾風伝	NARUTO 疾風伝
ayumi hamazaki	浜崎あゆみ
Ping Pong Club	稲中卓球部

3. 関連研究

これまで，自然言語処理において，英語の読み推定問題には音素や，つづり字(English Orthography)を用いる機械音訳の手法が研究されてきた．一方，ルールベースで行う手法や機械学習を用いる研究，それらを組み合わせた研究がなされている．例えば，英語と韓国語の読み推定問題において，音素を用いる手法，つづり字を用いる手法，ルールベースで行う手法の研究がある(Yu et al., 2011) [8].Mike ら[4]はつづり字を用いて，英語と中国語の機械音訳を行った．また機械翻訳を用いる手法では，羽鳥ら[3]が日本語の漢字仮名交じり文における読み推定問題を，Aramaki ら[1]は音訳のデコードの速さに焦点を当てた研究を行った．しかし，我々の知る限り，生データを用いて行われた研究は行われていない．本研究では実際に外国人ユーザによるノーマッチワードとなったデータを用いて音訳を行った．

4. クエリの修正

本研究では音素に着目し，音素を媒介として修正前クエリの音訳を行うことにより，マッチングをとった．音訳は次の５つのステップで行われる．

1. 音素と英単語の対応辞書を使い，修正前クエリを音素に変換する (3.1 節)
2. 修正後クエリにフィルタをかけ，ノイジのあるデータから音訳の際に学習データとなる日本語の修正後クエリを取り出す． (3.2 節)
3. 音素から日本語クエリへの翻訳確率を計算する (3.3 節)
4. 音素と日本語クエリの文字アライメントをとる (3.4 節)
5. 機械翻訳により，修正前クエリに対して該当する日本語を音訳する (3.5 節)

ステップ 1.からステップ 4.が学習データの作成段階であり，ステップ 5.が音訳段階である．

4.1 修正前クエリの音素への変換

音素への変換は音素と英単語の対応辞書である CMU Pronunciation Dictionary2(以下，CMUdict)に収録されているものを使用した．これにより，修正前クエリに音素対応する英語が含まれるものを対象とする．この時，音素が存在する英語クエリは，全 2833 レコードであった．

4.2 フィルタ

英語のクエリを音訳するにあたり，ペアコーパスにはノイズとなるデータが含まれているので，音訳の精度を上げるために文字種による学習データの絞り込みを行った．その際，学習データの質と量を適切に調節するため，2 種類のフィルタを用いた. 表 2 にペアコーパス中の音訳と意訳，文字種の例を示す．本来は学習データとして音訳対だけを使用したいが，ペアコーパスには音訳対(L)，意訳対(T)の両者が含まれており（表 2)，その判別は容易ではないことが分かる．

表 2　ペアコーパス中の音訳と意訳，文字種の例

Alphabet Query	Correct Query	transliteration(L) or translation(T)	Type of Characters of Correct Query
Doraemon	ドラえもん	L	Katakana, Hiragana
Miyazaki	ジブリ	T	Katakana
AKB48 poster	AKB48 ポスター	L	Katakana, Alphabet
Ufm rod	Ufm ロッド	L	Katakana, Alphabet
Tokyo adidas	東京 adidas	L	Chinese character, Alphabet
Dress Tokyo	原宿 ドレス	L, T	Chinese character, Katakana

そこで，音訳対の修正後クエリには，平仮名，片仮名が多く含まれ，また意訳対の修正後クエリには，漢字が含まれることに注目し, 文字種によるフィルタリングを行った．これにより，ペアコーパスから出来るだけ多くの音訳対を取り出し，学習データの質を高められると考えられる．

しかし，表 2 の 5 行目，tokyo-東京の例が示すように，修正後クエリが漢字でも音訳をしているケースがある．こ

2 http://www. speech.cs.cmu.edu/cgi-bin/cmudict

の様に，音訳と文字種の対応は完全に一致してはおらず，完全な判別はできない. また, フィルタリングを厳しくし，学習データの質を高めることは，すなわち学習データの量の減少という問題と直結する．そこで，次の 2 種類のフィルタを比較実験し，本実験での検索クエリの音訳に最適な選別条件を探った．

1. Chinese character filter (CF)
2. Chinese character and alphabet filter (CAF)

1 つ目のフィルタ Chinese character filter (CF)は，修正後クエリに，一文字でも漢字が含まれていたら，学習データから取り除くフィルタである．2 つ目のフィルタ Chinese character and alphabet filter (CAF)は， 修正後クエリに，1 文字でも漢字または英字が含まれていたら学習データから取り除くフィルタである．この時，英語クエリ全 2833 レコード中，CF により絞り込まれたデータは 2223 レコード (78.5%)， CAF により絞り込まれたデータは 714 レコード (25.2%)であった．

4.3 翻訳確率の計算

4.1 節で変換された音素を source language， 4.2 節でフィルタリングされた日本語クエリを target language として翻訳確率を計算した．この時，GIZA++[3] toolkit (Och and Ney, 2003)を使用した．

4.4 アライメント

機械翻訳をするために 1 クエリ毎に，音素と日本語クエリの文字対応付けを行った．まず， 4.1 節で変換された音素をラティスの X 軸に，それに対応する 4.2 節でフィルタリングされた日本語クエリを Y 軸に設定する．その時の部分経路のコストを，4.3 節で求めた翻訳確率の値とする．また，垂直方向と水平方向のコストは，翻訳確率からは計算できないので 10^{-20} とした．このラティスに対し，ダイクストラ法を用いて確率を最大にする経路を求め，その組み合わせを音素と修正後クエリの文字対応付けとする. また, 水平方向の文字対応付けは NULLJ, 垂直方向の文字対応付けは NULLP というタグに設定した．

例えば，英語クエリに対応する音素が“D AAl K Y AH0 M EH0 N T”であり，それに対応する日本語クエリが“ドキュメント”である場合のラティスは，次の図 1 である．そしてこの時に得られるアライメントは図 2 となった．

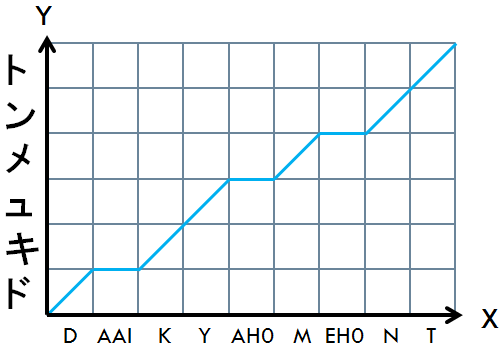

図 1 アライメントの例

[D - ド(do)]
[AAI - NULLJ]
[K - キ(ki)]
[Y - ュ(yu)]
[AH0 - NULLJ]
[M - メ(me)]
[EH0 - NULLJ]
[N - ン(n)]
[T - ト(to)]

図 2 “D AAl K Y AH0 M EH0 N T”と“ドキュメント”のアライメント結果

4.5 Machine Learning

機械学習をもちいて，クエリの音訳を行った．本研究では次の 3 つのモデルによる音訳を行った．

1. BIGRAM: 4.4 節で得られたアライメントをもとに Bigram Model を適用する．
2. HMM: 4.4 節で得られたアライメントをもとに隠れマルコフモデルを適用する．
3. CRF: 4.4 節で得られたアライメントをもとに CRF(条件付き確率場)を適用する．

本研究において，BIGRAM モデルと HMM モデルでは NLTK[4] を用い，CRF モデルでは CRF++[5] toolkit を用いた.CRF モデルにおいては unigram, bigram, trigram の素性を用いた．その素性を次に示す.

- Unigram: s-2, s-1, s0, s1, and s2
- Bigram: s-1s0 and s0s1
- Trigram: s−2s−1s0, s−1s0s1, and s0s1s2

またパラメータの値は，本研究においては素性の種類が多

3 http://www.fjoch.com/GIZA++.html

4 http://www.nltk.org/
5 http://crfpp.googlecode.com/svn/trunk/doc/index.html

表 3　修正前クエリ”ファブリック”に対する出力と評価例

method	BASE	BIGRAM	HMM	CRF
system output	フャブーンク	ファブリック	ファブリック	フブック
evaluation	1	3	3	2

くなるため f は 50 とし，c は 2 とした.
また, ベースラインとして, 次の方法と各手法を比較した.

- BASE:各音素に対し，最も高い翻訳確率となる日本語を出力する

以上の手法について，ペアコーパスを用い 5 分割交差検定を行った.

5. 評価

システムの評価を，母語を日本語とする 20 人によって行った．評価方法は，修正前のクエリに対しシステムが推定した読みが妥当かどうかを，3 段階(3 が高評価，1 が低評価)で主観判定した． 表 3 は修正前クエリが”ファブリック”であった時のシステムの出力と，その時についた評価の例である．この表において，理想的な出力は“ファブリック”となることと想定される．ここで，“precision 3”と “precision 3 or 2” を次のように定義する.

precision3＝3 の評価を得たシステムの出力の総数/クエリペアの総数

precision2 か 3＝3 の評価を得たシステムの出力の総数/クエリペアの総数

表 4，表 5 に，厳しめの評価である“precision 3”と 優しめの評価である“precision 3 or 2”の値による各手法のスコアを示す.

表 4　“The precision 3”による各手法のスコア

	CF	CAF
BASE	0.036	0.044
BIGRAM	0.029	0.071
HMM	0.062	**0.121**
CRF	0.064	0.046

表 5　“The precision 3 or 2”による各手法のスコア

	CF	CAF
BASE	0.323	0.209
BIGRAM	0.190	0.270
HMM	**0.448**	0.373
CRF	0.316	0.199

6. 考察

今までの English to non-Japanese Language の音訳研究では CRF モデルを用いる手法の適合率が高い(Shishtla et al 2009)と報告されてきたが，本研究では表 4，表 5 が示すように HMM モデルを用いたものの適合率が最も優れていた. この一因として，本実験では CRF の素性に trigram を採用したことがあげられる．ペアコーパス中の英語クエリは複数の単語からなる例が多いが，本システムではその区別ができず，一つづきの語とみなすので，その影響で性能が低下していると考えられる．例えば，"super mario"という英語クエリと，それに対応する音素として” S UW1 P ER0 M AA1 R IY0 OW0”がある．この” S UW1 P ER0 M AA1 R IY0 OW0”中の，音素”M”の音訳を出力するとき，本システムの CRF モデルでは，trigram の素性により” S UW1 P ER0” の中の”P”との関係も考慮に入れるが, 本来, mario の音訳に”P”は関係ない. このような要因が CRF モデルの結果を悪くしていると考えられる.

他に， 4.2 節で言及したように，CF では学習データの数が 714 レコードと CAF の 2223 レコードより少なかったが，厳しい評価を表す表 4 の結果では，CAF のほうが優れていた．このことから，より正確な音訳が求められる際には，学習データの数が少なくなっても，データの質を高めることがより重要と考えられる.

逆に，表 5 に見られるように，優しめの評価では CF のほうが優れている．つまり，ある程度の精度でよく，より多くの検索クエリの音訳を行いたい場合は，学習データの質よりも量を重視することが重要と考えられる.
まとめると, より正確な検索クエリの音訳を求めるならば, 学習データの量よりも質を高めることが重要であるが，おおざっぱな音訳でよいならば，コーパスの質を高めることよりも量が重要である．実際に検索クエリのマッチングを

とることを考えたときには，まず少量で良質なコーパスからの厳しめのデータを用いた結果を示した後，第二候補として多量でおおざっぱな音訳を示すなどの対応が考えられる．

表 4，表 5 が示すように，HMM 手法と CRF 手法はいつも BASE より優れているが，BIGRAM は BASE より劣っていることがある．この結果から BIGRAM モデルは検索クエリの音訳には適していないことが分かる．

次に，表 6 に HMM モデルかつ CAF という条件のもとで，単語が学習データの修正前クエリ中に現れる回数と，その単語数，スコアの平均値を示す．例えば，学習データ中に 1 回出現する単語の語彙数は 417 語あり，その平均スコアは 1.77 である．

表 6　学習データ中の英単語の頻度とスコアの関係

頻度	修正前クエリ中の単語の数	平均スコア
1	417	1.77
2	124	1.780
3	57	1.790
4	21	1.670
5	14	1.87
6	13	1.87
7	4	2.13
8	4	1.84
9	3	1.75
10	5	1.81
11	2	2.36
12	3	1.65
13	1	1.85
14	1	2.21
16	1	2.00
17	1	1.29
20	2	1.86
29	1	2.48
34	1	2.79

表 6 から，単語の出現頻度が高くなるとその平均スコアも高くなる傾向がわかる．CAF は 714 レコードを学習データとしており，頻度が 7(おおよそ全体の 10%)を超えたとき，初めて平均スコアが 2 に到達した．

また図 3 は表 6 中の，頻度と平均スコアの関係を表すグラフである．このように，出現頻度が増えるほど，音訳がより正確にできると考えられるので，学習データの総数を増やすことにより，より高い適合率を得ることができると考える．

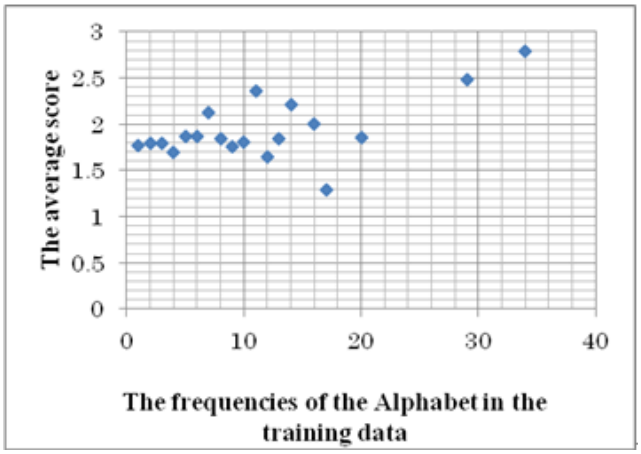

図 3　英単語の出現頻度とその平均スコアのグラフ

例えば，単語“figure”は全ての学習データの英語クエリ中，102 回出現したが，そのほぼ全てで評価 3 であった．

また，日本語出力が音素数に比べ極端に少なくなる例が見られた．これは 4.4 節のアライメントをとる段階において，音素の多くが NULLJ とアライメントされたことによると考えられる．よって NULLJ のコストをより少なくすることによって，改善できると考えられる．

7. おわりに

本研究では，実際に外国人が日本のショッピングサイトで商品検索の際に失敗したクエリとその正しいクエリの対のコーパスを用いて，非日本語圏のユーザによる検索クエリの音訳を行った．ペアコーパスには，意訳によって修正されたクエリのように，音訳するにはノイズとなるデータが含まれているため，文字種によるフィルタリングを行い学習データを絞り込んだ．この際，CF と CAF という 2 種類のフィルタを使い，学習データの質と量の調節を行った．本実験では，学習データの量や質，その目的によってフィルタを使い分けることが好ましいと分かった．

さらに，BIGRAM，HMM，CRF の 3 つの機械翻訳手法を比較した結果，検索クエリの音訳では HMM 手法が最適であった．HMM が CRF より優れていた理由の一つには，CRF の素性で trigram を使用したからだと考えられる．修正前クエリは複数の単語を含み，それらがエラー引き起こしたものと考えられるからである．

最後に，本実験結果を改善するためには，

- 学習データ量を増やす
- アライメントをとるときに NULLJ となるコストの

設置を変える

などが考えられる.

謝辞

本研究を行なうにあたり、ジェイグラブ株式会社様にデータをご提供頂きました．このような機会を賜りましたことを心から感謝します.

参考文献

1) Eiji ARAMAKI and Takeshi ABEKAWA. 2009. Fast decoding and Easy Implementation:Transliteration as Sequential Labeling, Proceedings of the 2009 Named Entities Workshop , ACL-IJNLP 2009, pages 65-68.
2) Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python: Analyzing Text with The Natural Language Toolkit, O'Reilly.
3) 羽鳥潤, 鈴木久美: 機械翻訳手法に基づいた日本語読み推定, 言語処理学会, 第 17 回年次大会, pp.579-582 (2011).
4) Mike Tia-Jian Jiang, Chan-Hung Kuo, Wen-Lian Hsu. 2011. English-Chinese Machine Translitaration using accessor Variety Features of Source Graphemes. Proceedings of the 2011 Named Entities Workshop, IJCNLP 2011, pages 86-90.
5) Canasai Kruengkrai, Thatsanee Charoenporn, Virach Sornelertlamvanich 2011. Simple Discriminative Training for Machine Transliteration. Proceedings of the 2011 Named Entities Workshop, IJCNLP 2011, pages 28-31.
6) Franz Joseph Och, Hermann Ney 2003. A sys-tenmatic comparison of various statistical alighnment models. Association for Computational Linguistics, ACL 2003, 29(4):417-449.
7) Praneeth Shishtla, Surya Ganesh V, Sethuramalingam Subramaniam, and Vasudeva Varma. 2009. A Language-Independent transliteration Schema-Using Character Aligned Model At NEWS 2009, Proceedings of the 2009 Named Entities Workshop , IJNLP 2009, pages 40-43.
8) Yu-Chun Wang, Richard Tzong-Han Tsai. 2011. English-Korean Named Entity Transliteration Using Statistical Substring-based and Rule-based Approaches. Proceedings of the 2011 Named Entities Workshop, IJCNLP 2011, pages .32-35.
9) Min Zhang, Haizhou L, A Kumaran and Haizhou Li. 2011. Report of NEWS 2011 Machine Transliteration Shared Task. Proceedings of the 2011 Named Entities Workshop, IJCNLP 2011, pages 1–13.