1L5-OS-09b-2in1

クラウドソーシングを用いたカタカナ文字列の変換について

Conversion of a katakana string using a crowdsourcing system

寺岡 照彦[*1]
Teruhiko Teraoka

坪内 孝太[*1]
Kota Tsubouchi

[*1] ヤフー株式会社 Yahoo! JAPAN 研究所
Yahoo! JAPAN Research, Yahoo Japan Corporation

This paper proposes a method for converting defective katakana character string to correct Japanese words using a crowdsourcing system. It is difficult to convert a katakana string with missing characters often appeared on credit-card statements due to a character limit. Our methods give hints including Web search resluts of an original katakana string to workers on a crowdsourcing system. Experimental results show our method increases an accuracy of a conversion compared with a usual Kana-Kanji conversion.

1. 緒言

本稿では、不完全さを含む長いカタカナ文字列を、仮名漢字混じりの正確な文字列に復元することを目的に、クラウドソーシング[Doan 11, Amazon]を用いた手法について述べる。特に、固有名詞が多く、文字の欠損がある不完全なカタカナ文字列を復元することは、通常の仮名漢字変換を使っても難しく、クラウドソーシングを用いても、思い通りの回答が得られない。そこで、ウェブ検索結果の提示など、ヒントを工夫することで、復元精度の向上を目指す。実験により、提案手法の有効性を検証した。

2. 不完全さを含むカタカナ文字列の変換

2.1 カタカナ文字列の変換

カタカナ文字列の変換には、仮名漢字変換の手法を使うのが通例である。仮名漢字変換では、統計的仮名漢字変換[森99]が一般的であり、特に今回対象とするようなある程度変換対象が限られているような問題では、正解データを準備し、機械学習を利用する方法が有効と考えられる。但し、従来の手法では、対象とする変換前のカタカナ文字列は通常、文字列自体に誤りや欠損を含まないものが想定されている。

2.2 入力文字列：不完全さを含むカタカナ文字列

本研究において、復元対象の文字列は、クレジットカードなどの利用明細に記載されることになる、店舗名のカタカタ文字列とする。例えば「ABC 電気利用料金」という支払いに対して、利用明細のカタカナ文字列は「ABC デンキリヨウリヨウ」となる。

まず、「料金」をカタカナ文字列に直すと正しくは「リョウキン」であるが、システムの都合上小文字に対応できず、大文字になっている。また、データの保存領域の問題で「リョウキン」の「キン」は欠落してしまっている。その結果、「ABC デンキリヨウリヨウ」となる。このように、1)大文字と小文字の区別がない、2)保存できる文字数に制限があり、それを超える文字は欠落している、という 2 つの不完全さの特徴がある。

3. クラウドソーシングによるカタカナ文字列の変換

不完全さを含むカタカナ文字列の変換において、人の知恵を使うことが有効ではないかと考えた。そのため、クラウドソーシングを用いたカタカナ文字列の変換を提案する。

不完全さを含む文字や文章を、クラウドソーシングを使って正確にする研究として、SNS 上で使われる略語を正式な言葉にする研究[福島 15]、不完全な翻訳結果を正確な対訳にする研究[酒井 11]などがある。しかしながら、これらが対象にしているのは、提示されるデータが情報としては完全なものであり、本研究で対象にしている、欠落などで情報自体が不完全なものは、クラウドソーシングを普通に適用しても、復元が難しいと考えられる。

そこで、本研究では、クラウドソーシングのワーカに対して何らかのヒントを提示する事で、復元の精度が向上すると考えた。ヒントの提示方法として以下の 2 通りを考え、実証実験により、それらの効果を検証する。

- 仮名漢字変換処理後の候補を提示
- ウェブ検索結果へのリンクをヒントとして提示

4. 実証実験

4.1 実験の概要

実験は、Yahoo! JAPAN カードの利用対象となる店舗名データ 3000 件で行った。データのサンプルを表 1 に示す。表 1 の通り、大文字/小文字の区別がつかない、保存文字数制限を超える文字は欠落している、といった特徴が観察できる。

表 1. 実際のデータと、正しい変換例

変換前の文字列	正しい変換例
ケイデイデイアイゴリヨウリ	KDDI ご利用、 KDDI ご利用料金
トウキユウハンズシンジユク	東急ハンズ新宿
メツトライフセイメイホケンリヨ	メットライフ生命保険料
ニツケイアイデイケツサイ	日経 ID 決済

クラウドソーシングについては、Yahoo!クラウドソーシング[ヤフー]にタスクを依頼した。1 件の店舗名データに対し、5 人のワーカを割当てた。

4.2 比較手法（ベース・ライン）

比較手法として、1)統計的仮名漢字変換、2)ヒントを提示しないクラウドソーシング、の 2 つを用いて検証した。

連絡先：寺岡照彦，ヤフー(株) Yahoo! JAPAN 研究所，東京都港区赤坂 9-7-1，tteraoka@yahoo-corp.jp

統計的仮名漢字変換では、欠落したカタカナ文字列ではなく、仮名漢字混じりの店舗名文字列のデータが存在する別のカードサービスのデータを利用した。それぞれの店舗名の文字列を形態素解析[JUMAN]し、各単語とその読みを学習して変換を行った。なお、この仮名漢字変換では、N-best 方式で回答を出すことができる。ベースラインの手法としては一番スコアの高かった 1-best の結果を用いた。クラウドソーシングに出すヒントの提示として用いる場合には、5-best、つまりスコアの高い候補から順に 5 件の変換結果を採用した。

4.3 評価方法

クラウドソーシングを行った結果、5 人のワーカからの回答が集まる。その回答に対して、3 人以上が同じ回答をしていた場合のみその回答が正しいかを評価した。全員バラバラの回答の時や、2 人だけが同じ回答をしているケースについては、「そもそもクラウドソーシングでは変換ができず、回答を絞り込めない、誤り」とした。

最後に仮名漢字変換やクラウドソーシングにより得られた回答を複数人の目視判断で正解/不正解のラベルをつけた。なお、目視の判定者の割当てについて、次の注意を払い評価を進めた。すなわち、対象文字列 1 件につき、4 つの手法により比較を行うため、最大 4 通りの回答が集まるが、それら 4 つの回答は同じ判定者によって正解/不正解を判断するようにした。目視判断にはある程度ルールはあっても、個人差が考えられるため、評価を公平にするためにそのような工夫をした。

4.4 実験結果

実験の結果を表 2、図 1 に示す。表 2 は、各手法における変換の正解率、図 1 はクラウドソーシングを用いた場合に何人が同じ回答をしたか、という割合を示す。同じ回答を多くの人がした場合の方が、クラウドソーシングの結果として、より信頼のおける回答が導出されているといえる。

仮名漢字変換のみだと 56.6%の精度であるが、クラウドソーシングを用いるだけで 73.2%にまで向上した。クラウドソーシングと本研究の不完全さを含むカタカナ文字列の変換が相性が良いことが分かる。

表 2. 実験の結果の概

	正解率
仮名漢字変換のみ	56.6%
クラウドソーシングのみ	73.2%
クラウドソーシング （仮名漢字変換によるヒント）	74.2%
クラウドソーシング （ウェブ検索によるヒント）	78.0%

仮名漢字変換の結果（5-best）をヒントとして提示する手法ではクラウドソーシングだけの結果よりも、若干の改善が見られる程度となった。これは仮名漢字変換の精度に課題があるためと考えれる。もともと 56.6%の精度しか出ない自動変換であり、ヒントとして有効に効果を発揮できていないのではないかと考察できる。

一方、同じヒントを提示する方法として、ウェブの検索結果を表示する方法は、クラウドソーシングのベースラインよりも 4.8%の向上が見られ、とても有用であることがわかった。図 1 の回答の信頼度のグラフを見ても、ウェブ検索をヒントとして用いることで「5 人全員が正解」というケースが増えており、より信頼のおける変換が実現できていることが分かる。ウェブ検索は曖昧さも加味した検索結果を提示するため、その効果が本問題においてはワーカにとって有効な作業支援になったと考えられる。

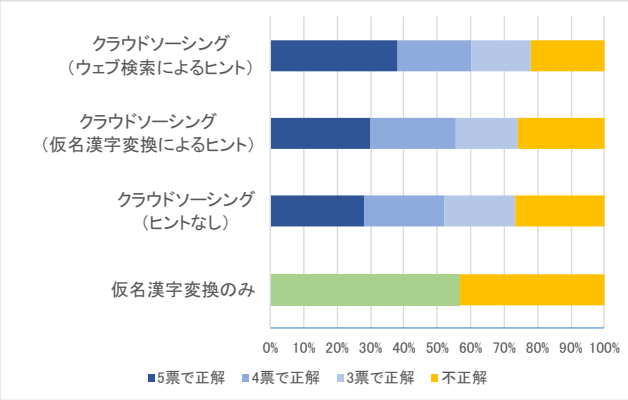

図 1. クラウドソーシングの信頼度の比較結果

5. 結言

不完全さを含むカタカナ文字列の変換として、ヒントを提示するクラウドソーシングを用いた変換を提案し、実験によりその有効性を示した。

大文字/小文字の区別がつかない、文字に欠損がある、ような不完全な文字列は、機械的に変換するには限界がある。それに対し、ワーカの知恵を使うクラウドソーシングは、ある程度の不完全さは補完でき、効果的な変換が実現できる。クラウドソーシング単体でも有効ではあるが、ウェブ検索結果をヒントとして提示する方法により、さらなる精度向上を確認できた。

今後は、本研究で確認した結果をもとに、クラウドソーシングにより、自動的に不完全さを含むカタカナ文字列を変換するシステムの開発を行う。具体的には、統計的に信頼のおける回答が集まるまでクラウドソーシングで回答を集め続け、信頼のおける結果が得られた時に、それを回答とする方法である。

参考文献

[Doan 11] A. Doan, R. Ramakrishna, and A. Y. Halevy: Crowdsourcing systems on the World-Wide Web, CACM, Vol.54, No.4,pp86-96, 2011

[森 98] 森,土屋,山池,長尾: 確率的モデルによる仮名漢字変換, 情報処理学会論文誌, Vol.40, No.7, pp.2946-2953, 1999

[酒井 11] 酒井,芦川,廣川: Crowdsourcing System を用いた略語の推定手法の提案, 信学技報 NLC2011-36, 2011

[福島 15] 福島,吉野: 翻訳パズル:クラウドソーシング上における不完全な翻訳を用いた対訳作成手法, 情処研資 GN-93(37), 2015

[JUMAN] 日本語形態素解析システム JUMAN, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[Amazon] Amazon Mechanical Turk, https://aws.amazon.com/jp/mturk/

[ヤフー] Yahoo! クラウドソーシング, http://crowdsourcing.yahoo.co.jp/