日本語テキストにおけるアルファベット文字列の読みクラス分類

浅野久子　永田昌明　阿部匡伸

日本電信電話株式会社　ＮＴＴサイバースペース研究所

1. はじめに

本研究は、日本語テキストに含まれるアルファベット文字列の読み精度向上を目的とする。ここで、アルファベット文字列とは、A-Z、a-z、'（アポストロフィー）からなる文字列とし、アルファベット文字列のみで構成される単語（形態素解析の認定単位[1]）をアルファベット単語とよぶことにする。

日本語テキストにはアルファベット単語が数多く含まれる場合がある。例えば、あるWebサイトの約8000店の「買う」というジャンルの店舗情報には、延べで約6400語、異なりで約2900語のアルファベット単語が存在する。これらのアルファベット単語は固有名詞も多く、すべてを辞書登録するのは非現実的であるため、テキスト音声合成で正しく読み上げるためには、アルファベット未知語に対して、読みを付与する必要がある。

Speech Synthesis Markup Language[2]では、テキスト中の部分文字列に対して、直接的な読み（例えば音素表記）ではなく、間接的なクラスを付与して読み方を指定するsay-as要素が存在する。例えば、英語テキスト内の"123"という文字列に対して、digitsというクラスでは"one two three"、ordinalというクラスでは"one hundred and twenty third"と読ませる。本稿では、このように読み方を指定するためのクラスを読みクラスとよぶ。日本語テキストにおけるアルファベット単語を対象とした読みクラスの標準は存在しないので、本稿では、各言語（ただし日本語はローマ字）読み、各言語のアルファベット読み（英語の場合、"A"=エー、"B"=ビー）、各言語の略語（英語略語の例："NY"＝ニューヨーク）と定義した。

アルファベット単語の読みクラス分類に関連する研究として、西欧言語の言語識別の研究[4][5]がある。これらは、あるテキストを、その中に含まれる複数の単語の綴り情報を利用して、1つの言語に識別するものである。読みクラスの分類は単語単位に行う必要があるため、これらとは処理単位が異なる。

従来のアルファベット単語に対する読み付与方法には、　文字列bigramを用いる方法[1]や、最長一致を用いる方法[2]がある。[1]は、英語（単一言語）を前提にした方法であり、英日外来語辞書での評価において、多くの誤りは人名（おそらく他言語由来の語）や頭文字語であったと報告している。すなわち、英語読みクラス以外の単語に対する誤りが多い。また、[2]は西洋人名辞典データを対象に、多言語混合と言語別での評価を行い、言語別（つまり、読みクラス別）の方が読み精度がよいと述べている。

本研究では、英語読み・ローマ字読み・フランス語読みなど、多種の読みクラスの単語が混在するWeb店舗情報を対象テキストとする。そこで、アルファベット未知語への読み付与として、まず、アルファベット未知語の読みクラスを推定し、次に、読みクラス別の読み付与規則（読みクラス別に学習した[1][2]等の従来技術やアルファベット表など）を適応して、読み付与を行うというアプローチをとる。本稿では、この読みクラスの分類において、Support Vector Machine（SVM）[3]を利用した手法について提案する。SVMを用いたのは、近年、様々な自然言語処理タスクに適用され、その有効性が報告されているためである。SVMは2値分類器であるため、読みクラス分類などの多値クラス分類を扱うために拡張が必要であるが、本稿では、SVMをpairwise法などにより多値分類に拡張した汎用TaggerであるYamCha0.2[3]を使用した。

2. 対象とする読みクラス

本稿では、Web店舗情報を対象として、アルファベット単語の読み精度の向上をめざす。そこで、前述の約8000店の店舗情報における店名と店舗紹介文を対象に、アルファベット単語の読みクラスの分布を調査した。この結果を表1に示す。ここで、略語クラスについては、綴りのみから読みを推定するのは困難であるため、辞書登録（未知語にしない）で対応することとし、読みクラス分類の対象とはしない。この略語クラス以外の上位5クラス、英語読み、アルファベット読み、ローマ字読み、フランス語読み、イタリア語読みを本稿での読みクラス分類の対象とする。

---

[1] 未知語の場合、1単語をどのように定義するかは、4節で検討する。

[2] http://www.w3.org/TR/speech-synthesis/

[3] http://cl.aist-nara.ac.jp/~taku-ku/software/yamcha/

表1 Web店舗情報「買う」における読みクラス分布

読みクラス	延べ単語数 (割合%)		異なり単語数 (割合%)	
英語読み	3883	(60.5%)	2023	(69.5%)
(英語)ｱﾙﾌｧﾍﾞｯﾄ読み	1366	(21.3%)	291	(10.0%)
ローマ字読み	486	(7.6%)	278	(9.5%)
フランス語読み	333	(5.2%)	199	(6.8%)
イタリア語読み	97	(1.5%)	56	(1.9%)
略語（全言語合計）	222	(3.5%)	37	(1.3%)
その他	30	(0.5%)	28	(1.0%)
計	6416		2911	

3. アクセント文字の扱い

フランス語とイタリア語にはアクセント文字（é, ù, ê 等）が存在する。しかし、前述の Web の店舗情報では、アクセント文字を使わずに、アクセントを省略した文字にする（é→e）、あるいはアポストロフィー（'）やグラーブ（`）付に分解する（é→e', ù→u`, u'）ことで表している[4]。ここで、グラーブ（ù）の分解においては、元々数が少ない上に、アポストロフィーとグラーブのどちらに分解するかのゆれがある。そこで、本稿では、グラーブ（`）はアポストロフィー（'）に置き換えて扱うことにする。

4. 処理単位

読みクラス分類は形態素解析の出力結果である単語を単位として行う。ただし、アルファベット未知語をどのような単位で1単語とすべきか検討が必要である。読みクラスの境界となりうる位置では、単語を分割すべきである。

アポストロフィーは、英語で所有格を表す "'s" の前に他の読みクラスがくる場合（例：Hanako's）があるため、読みクラス境界になりうる。ただし、この所有格などのように、直後の文字とつながりがある場合と、3節で述べたアクセント文字の分解（é→e'）など、直前の文字とつながりがある場合の双方がある。そこで、アポストロフィーの前、後、あるいは、前後両方のどれを単語境界とすべきか、検証が必要である。この検証は 6.5.1 節の実験で行う。

また、アルファベット小文字から大文字への境界も"CafeHanako"のように読みクラス境界になる場合がある。しかし、Web 店舗情報においては、"HaNaKo"、"DeeP"など、つながりが強いと思われる境界でも、小文字→大文字境界になる場合があり、これを単語境界とすることで逆に精度が低下する可能性もある。そこで、この境界を単語境界とすべきかについては、6.5.2 節の実験で検証する。

5. 文脈と属性

読みクラスの分類においては、推定対象となるアルファベット単語自身の情報に加えて、その前後の単語情報（文脈）が有効であると考えられる。例えば読みクラスの多義がある"Take"は、"Take it easy"では英語読み（テイク）、"Take ちゃん"ではローマ字読み（タケ）と、文脈から読みクラスが決定する。そこで、YamCha のデフォルトの設定通り、読みクラス推定対象となる単語に加えて、その直前、直後各2単語の情報、および、直前2つの推定された読みクラスを属性として利用することとする。

各単語の情報としては、以下の属性を検討した。それぞれの属性の有効性は、6.4 節で検証する。

- 綴り情報：綴りに関する情報は言語識別の研究でも利用されており、効果が期待される。ここでは、アルファベット単語の（一部の）文字または音節の表記を綴り情報として扱う。ただし、各言語の正しい音節境界を機械的に正しく得るのは困難であるため、便宜的に母音（AIUEO）から子音（AIUEO 以外）への境界を音節境界とする。また、アルファベット単語は大文字・小文字の組み合わせにより表記のバリエーションがあるが、文字・音節の表記は綴りのみに着目するためにすべて大文字に統一する。アルファベット単語以外では、この属性は利用しない（ダミーの値を入れる）。
- 文字数：単語の文字数。
- 文字タイプ：アルファベット単語は、すべて大文字、すべて小文字、先頭のみ大文字、その他に分ける。ひらがな、カタカナ、漢字、数字からなる語はその字種とする。それ以外の文字（記号や空白）は、その文字表記そのものとする。
- 品詞：主品詞（名詞、形容詞等）レベル。未知語に対しては、"未知語"という品詞を与える。

6. 実験

ここまでの検討を検証するために、6.1 節で示すコーパスを用いて実験を行った。実験はすべて、YamCha をデフォルトの設定で利用した。

また、ローマ字になりえない綴りの単語を検出するのは容易であるので（street：'tr'という綴りはない、't'は末尾にこない等）、全実験において、ローマ字になりえない綴りの単語は、YamCha の推定対象からあらかじめローマ字を除いて推定した。

6.1. コーパス

学習用のコーパスとして、データ作成に手間がかかる Web 店舗情報（表1の約 8000 店舗情報：ジャ

[4] その他のアクセント文字（û,ö 等）では、分解は存在せず、すべて省略となっている。

ンル「買う」）と、データ量を補うために、データ作成が容易な辞書類の2種類を準備した。

各読みクラスに対応する辞書類として、英・仏・伊語読みについては、それぞれの言語辞書から約1万～1万2千語を抽出して利用した[5]。ローマ字読み、アルファベット読みについては、相当する辞書が存在しなかった。このため、ローマ字読みは日本人の姓と名、約1万2千語を機械的にローマ字変換した。アルファベット読みについては、手作業で約900語収集した。辞書類の合計は、約4万6千語となった。これらは、基本的に1語単位であるが、複合語など、ごく一部は複数語で1単位となっている。これら辞書類での1単位を1文として扱った。このため、辞書類には日本語の文脈情報は全くない。

また評価用のコーパスとして、Web店舗情報のジャンル「買う」（学習用コーパスとの重複はない）と「食べる」から、アルファベット単語を含む文を無作為に抽出した。読みクラス推定の対象はアルファベット未知語であるが、未知語の数は、アルファベット単語が単語辞書にどの程度登録されているかに依存する。そこで、本実験においては、学習用コーパスにおいて、出現頻度が5未満の単語[6]を未知語とみなした場合の正解率（推定正解数/推定対象単語数[%]）で評価した。評価用コーパスにおける全アルファベット単語数と、未知語扱いする単語数を表2に示す。ここでの単語数は、小文字→大文字境界を単語境界としない場合である。

表2 評価用コーパスのアルファベット単語数

アルファベット単語	「買う」	「食べる」
全単語	706	771
未知語(頻度5未満)	526	629

6.2. 初期設定

処理単位は、小文字→大文字境界で単語境界としない、アポストロフィーはその前後を単語境界とすることとした。各単語の属性は、第1,2,末尾-1,末尾音節表記[7]、文字数、文字タイプ、品詞の7つとした。

6.3. 学習データ源・量の検証

6.2節の初期設定で、学習データ源・量についての検証するため、(1)辞書類のみ、(2)Web店舗情報のみ、(3)辞書類+Web店舗情報、(4)(2)の半分のWeb店舗情報のみの4種類の学習データに対する評価を行った。その結果を表3に示す。

表3 学習データ源・量別の正解率

学習データ源・量	「買う」	「食べる」
(1) 辞書類	37.6%	34.8%
(2) Web店舗情報	87.1%	86.6%
(3) (1)+(2)	89.0%	86.8%
(4) (2)の半分	85.9%	82.4%

表3より、(1)の日本語の文脈情報を全く含まない辞書類のみは学習量は多いが効果がないことがわかる。(2)のWeb店舗情報のみは、学習量は約6400語のアルファベット単語+その前後の文脈情報のみで少ないが、正解率が格段に上昇する。(2)に辞書類を追加した(3)では、ジャンル「買う」では正解率が上昇するが、「食べる」ではほとんど変化がない。学習データ量が(2)の8倍以上である割には効果が低い。これより、Web店舗情報というドメインでは、そのコーパスのみを利用するだけでも十分であるといえる。(4)でWeb店舗情報のデータ数を減らしたところ、正解率が減少している。このため、Web店舗情報の学習データをさらに増やせば、より精度が向上する可能性がある。

以降の実験は、最高値の正解率を出した(3)辞書類+Web店舗情報を学習データとして行った。

6.4. 属性の検証

6.4.1. 初期設定属性の有効性

6.2節で設定した7つの属性を1つずつ除き、有効な属性を調査した。全属性を利用した場合との正解率の差分を表4に示す。マイナスの値が大きいほど、有効な属性となる。属性は、ジャンル「買う」における有効性の高い順に表記した。

表4 1属性除いた場合の正解率の差分

除いた属性	「買う」	「食べる」
(1) 末尾音節表記	-7.8%	-4.6%
(2) 第1音節表記	-6.7%	-4.0%
(3) 文字数	-3.4%	+0.5%
(4) 第2音節表記	-2.3%	-1.9%
(5) 文字タイプ	-2.1%	-0.5%
(6) 末尾-1音節表記	-0.8%	-1.1%
(7) 品詞	+0.2%	+0.5%

表4より、最も有効なのは、綴りの情報を直接表す音節表記であるといえる。第1・末尾音節と比較して、第2・末尾-1音節の有効性が低いのは、この属性は音節数が3または4以下の単語では意味のな

---

[5] アクセント文字は、"分解", "省略"の2つに展開した。ただし脚注4については"省略"のみに置換した。

[6] 文字数が1文字の単語、および、学習用コーパスで読みクラスの多義が存在した単語は頻度に関係なく未知語に含めた。また、本稿で推定対象外の読みクラスの語（例えばドイツ語や略語）は、未知語としていない。

[7] 4音節未満の単語では、第1,末尾, 第2音節の順に音節表記を入れ、空いた部分にはダミーの表記を入れた。

い属性（相当する表記がない）であり、音節数がそれ以下の単語の割合が高いためだと考えられる。

品詞は、わずかではあるが利用しない方がよいという結果になった。このため、以降の実験では、品詞を属性から除く。

6.4.2. 音節と文字の比較

綴り情報として音節を用いた場合（＝表4の(7)）と、それを文字表記に置き換えた場合の正解率を表5に示す。綴り情報として、文字より音節表記を用いるほうが、非常に有効であるといえる。

表5 音節利用と文字利用の正解率

綴り情報の単位	「買う」	「食べる」
音節	89.2%	87.3%
文字	83.7%	82.2%

6.5. 処理単位の検証

6.5.1. アポストロフィーの扱い

アポストロフィーの前後を単語境界とした場合（＝表4の(7)）、アポストロフィーの前のみ、後ろのみを境界とした場合の正解率を表6に示す。出現頻度が低いためか、正解率の差はあまりなく、ジャンルにより有効性の順序にばらつきがある。

表6 アポストロフィーの境界位置別の正解率

アポストロフィーの単語境界位置	「買う」	「食べる」
前後境界	89.2%	87.3%
前のみ境界	86.9%	87.4%
後のみ境界	87.3%	87.1%

6.5.2. 小文字→大文字境界の扱い

小文字から大文字への境界を単語境界としない場合（＝表4の(7)）と、単語境界とした場合の正解率を表7に示す。正解率は、両者とも小文字→大文字境界を単語境界としない単位で求めた。評価データで小文字→大文字境界が読みクラス境界となった単語は、ジャンル「買う」では1語だけであり、「食べる」では存在しなかった。このためか、正解率にほとんど差はみられない。

表7 小文字→大文字境界の扱い別の正解率

小文字→大文字境界の扱い	「買う」	「食べる」
単語境界としない	89.2%	87.3%
単語境界とする	89.2%	87.6%

6.6. 最適ケースと誤りの分析

属性の有効性の順序が、ジャンル「買う」と「食べる」の評価で入れ替わる部分があり、厳密に最適なケースを選ぶのは難しいが、正解率が最も高い値となったのは、単語境界=アポストロフィーの前後、属性＝音節表記、文字数、文字タイプとしたもの（表4の(7)）で、ジャンル「買う」における正解率89.2%であった。処理単位は、Web店舗情報においては、どれを選択しても、あまり差はないと考えられる。

以下では、この最高正解率のケースにおいて、誤りの内容を分析する。誤りの代表例を表8に示す。

表8 読みクラス誤りの代表例

No.	誤りｸﾗｽ(正解ｸﾗｽ)	具体例=下線部
(1)	英語（仏語）	La Boutique Montres a Paris
(2)	英語（伊語）	カフェ Capanna
(3)	ｱﾙﾌｧﾍﾞｯﾄ（英語）	B-A-R

誤りの内訳としては、英仏語間の誤りが最も多く、「買う」の全誤り57件のうち29件（51%）、「食べる」の全誤り80件のうち34件（43%）を占める。

ジャンルについては、学習データと同じ「買う」の方が、全実験を通して精度がよかった。これは、イタリア語読みが正解である語の数が「食べる」の方が多く、その誤りが多くなるのが、主な原因であった（「食べる」：30語中13語誤り、「買う」：17語中6語誤り）。また「食べる」では、表8の(3)のように、本来は1語扱いとすべき単語が1文字ずつ扱われ、誤りとなった事例も存在した。

7. おわりに

本稿では、SVMを用いた読みクラス分類手法を提案し、その有効性を確認した。今後は、読みクラス別の読み付与規則を適用し、読みの正解率を評価していく予定である。

参考文献

[1] 齋藤，篠原，永田，小原：音声制御ブラウザVCWebの英日シームレス化，人工知能学会論文誌，Vol.17, No.3, 2002

[2] 増田，梅村：人名辞書から名前読み付与規則を抽出するアルゴリズム，情報処理学会論文誌，Vol.40, No.7, 1999

[3] V. Vapnik: The Nature of Statistical Learning Theory, Springer, 1995

[4] G. Kikui: Identifying the Coding System and Language of On-line Documents Using Statistical Language Models, 情報処理学会論文誌, Vol.38, No.12, 1997

[5] W. B. Cavnar, J. M. Trenkle: N-Gram-Based Text Categorization, Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994