

非内容語の N-gram に基づく英語母語話者性の推定

緒方 伸輔* 田中 省作** 富浦 洋一***

*九州大学大学院システム情報科学府

** 九州大学 情報基盤センター 研究部 外国語情報メディア部門

*** 九州大学大学院システム情報科学研究所

近年 Web からテキストデータを収集して言語処理や言語教育に利用しようという試みが盛んに行われている。その一つとして、英語の学習に必要な知識の獲得が挙げられる。抽出される知識の信頼性を考慮すると、知識を抽出するテキストデータは英語の母語話者によって書かれたテキストであることが望ましい。したがって、母語話者によって書かれた英語テキストを高い精度で抽出する必要がある。本稿では統計的言語モデルを用いた母語話者性の推定手法を提案する。また、小規模な推定実験により、提案手法の有効性を示す。

Estimation of the Nativeness of English Documents Based on N-Gram Model of Non-content Words

Shinsuke Ogata*, Shosaku Tanaka**, Yoichi Tomiura***

* Graduate School of Information Science and Electrical Engineering, Kyushu-University

** Computing and Communication Center, Kyushu University

*** Faculty of Information Science and Electrical Engineering, Kyushu-University

In recent years, it has been popular to collect text data from Web, for example, to use them for learning English. Those data are desired to be written by an English native because texts for learning English should be reliable as materials. Therefore, it is necessary to extract texts written by natives from Web with high precision. This paper proposes a statistical method to estimate the nativeness of English documents. The experimental result shows this proposed method is useful to estimate the nativeness.

1. はじめに

近年 Web からテキストデータを収集して言語処理や言語教育に利用しようという試みが盛んに行われている。その一つとして、英語学習に必要な知識の獲得が挙げられる。例えば英作文を例にとって考えてみる。日本語の単語 w の訳の候補は和英辞書で得られるが、 w が特定の日本語文中で現れた場合、適切な訳を辞書の例

文だけから特定することは難しい。さらに単純に和英辞書を活用するだけでは得られる訳の候補はどれも適切でない場合もある。そこで、Web 上の英文テキストを膨大な用例集と考え、辞書からは得られない知識を抽出することが考えられる[1]。Web から得られるデータには母語話者・非母語話者によって作成されたデータが混在しており、抽出される知識の信頼性を考えると、知識を抽出するテキストデータは英語の母語話者に

よって書かれたテキストであることが望ましい。したがって、母語話者によって書かれた英語テキストを高い精度で抽出する必要がある。本稿では統計的言語モデルを用いた母語話者性の推定手法を提案する。

2. 母語話者性推定手法

本手法では、母語話者性の言語モデル M_N 及び非母語話者性の言語モデル M_{NN} の両方を構築し、テキストの母語話者性推定を行う。本章ではこれらの手法について述べる。

2.1 非内容語による言語モデル

本研究で目的としている言語モデルは母語話者/非母語話者の性質を反映するような言語モデルである。したがって、記事の分野などの、テキストの内容に依存するような言語モデルの構築を避けなければならない。

内容依存を避けるための手段の一つとして、内容語の除去が挙げられる。しかし、内容語を完全にテキストから除去してしまうと完全に非内容語列のみから成るテキストとなってしまうため、構文的な要素などが言語モデルに反映されなくなってしまう。そこで、テキストの内容語部分をその語の品詞タグに置き換え、非内容語(機能語)と品詞タグから成るテキストから言語モデルを構築する。

一般的に内容語というと、名詞・動詞・形容詞・副詞を指すが、本研究では内容語をそのテキストに特有な単語、つまり多くのテキストには出現しない単語であると考えた。具体的には、単語 w が出現した文書数 DF_w (Document Frequency) が閾値 df 以下の w を内容語として品詞に置き換えた。

2.2 統計的言語モデルによる母語話者性の推定法

母語話者性の推定問題を、テキスト d の属するカテゴリ $c(c \in \{\text{母語話者}(N), \text{非母語話者}(NN)\})$ を推定する統計的なパターン認識の問題として捉えるならば、テキスト d の属するカテゴリ $Cat(d)$ は以下のようにして求まる。

$$Cat(d) = \begin{cases} N & ; P(NN|d) < P(N|d) \\ NN & ; P(NN|d) \geq P(N|d) \end{cases}$$

ここで、 $P(c|d)$ はテキスト d が与えられたとき、そのカテゴリが c である条件付確率である。つまり、

$$\frac{P(N|d)}{P(NN|d)} = \frac{P(N)}{P(NN)} \frac{P(d|N)}{P(d|NN)}$$

と1との大小関係に基づき、テキスト d の $Cat(d)$ を推定する。 $P(d|c)(c \in \{N, NN\})$ は、カテゴリ c が与えられたとき、テキスト d が発生する条件付確率、すなわち言語モデル M_c に基づく d の発生確率である。

$$J(d) = \frac{P(d|N)}{P(d|NN)}$$

とおくと、

$$Cat(d) = \begin{cases} N & ; J(d) > \frac{P(NN)}{P(N)} \\ NN & ; J(d) \leq \frac{P(NN)}{P(N)} \end{cases}$$

と判定することになり、これが M_N, M_{NN} に基づいた場合の母語話者/非母語話者の判定の正解率が最も高い判定法である。ただし、本研究の目的を考慮するならば、母語話者テキストの抽出の再現率は多少犠牲にしても適合率を高くすべきである。この場合は上記の判定に代え、

$$J(d) > \theta \Rightarrow Cat(d) = N \quad \dots(1)$$

と判定する。 θ は実験的に定める閾値である。

本研究では言語モデル M_N 及び M_{NN} を n -gram モデルとした。文 s の生起確率 $P(s)$ 及びテキスト d の生起確率 ($P(d|c)(c \in \{N, NN\})$) の定

義を以下に示す。

$$P(d|c) = \prod_{i=0}^n P(s_i|c) \quad (d = s_0 s_1 \dots s_n)$$

$$P(s) = \prod_{j=0}^{m+1} P(w'_j | w'_{j-2} w'_{j-1})$$

$$w'_i = \begin{cases} w_i \text{の品詞} & (w_i \in \text{内容語}) \\ w_i & (w_i \in \text{非内容語}) \end{cases}$$

$$s = w'_0 w'_1 \dots w'_m$$

$$w'_{-2} = w'_{-1} = \langle S \rangle$$

$$w'_{m+1} = \langle E \rangle$$

2.3 N-gram モデルの学習

本実験では内容語を品詞に変換したテキストから N-gram モデル(本実験では特に N=3 とした)を学習する[2]。しかし、学習テキストの量が十分でないため、データスパースに陥る恐れがある。そこで、bi-gram 及び uni-gram により tri-gram モデルの補間を行う。補間された tri-gram $P(w_n|w_{n-1}w_{n-2})$ の定義を以下に示す。

$$\begin{aligned} P^{<c>}(w'_n | w'_{n-1} w'_{n-2}) = & \\ & \lambda_3 P^{<c>}_{ML}(w'_n | w'_{n-1} w'_{n-2}) \quad \dots (2) \\ & + \lambda_2 P^{<c>}_{ML}(w'_n | w'_{n-1}) \\ & + \lambda_1 P^{<c>}_{ML}(w'_n) \end{aligned}$$

$$P^{<c>}_{ML}(w'_n | w'_{n-1} w'_{n-2}) = \frac{\text{Freq}(w'_n w'_{n-1} w'_{n-2})}{\text{Freq}(w'_{n-1} w'_{n-2})}$$

$$P^{<c>}_{ML}(w'_n | w'_{n-1}) = \frac{\text{Freq}(w'_n w'_{n-1})}{\text{Freq}(w'_{n-1})}$$

$$P^{<c>}_{ML}(w'_n) = \frac{\text{Freq}(w'_n)}{N}$$

$$C \in \{N, NN\}$$

¹ 文の始め及び終わりに出現する単語の条件付確率を考慮するため、文の最初と終わりにそれぞれダミータグ<S>と<E>を付加し、文の生起確率を求めた。

$\text{Freq}(w_n w_{n-1} \dots w_{n-k})$: 単語列 $w_n w_{n-1} \dots w_{n-k}$ の出現頻度

N : 出現した単語の総数

λ_i : 重み係数(但し、 $\sum_{i=1}^3 \lambda_i = 1$)

本モデルでは w' が以下のいずれかになる。

- w
- $H(w)$ (w の品詞)
- $\langle S \rangle$ (文の先頭を示すシンボル)
- $\langle E \rangle$ (文の終わりを示すシンボル)

これらは学習データにおけるスパースさの程度が異なるため、tri-gram を構成する w' がそれぞれ上記のいずれになるかで重み付けパラメータ λ を分類した。つまり、式(2)の定義を以下のように変更する。

$$\begin{aligned} P^{<c>}(w_n | w_{n-1} w_{n-2}) = & \\ & \lambda_3^{<H_{n-1} H_{n-2}>} P^{<c>}_{ML}(w_n | w_{n-1} w_{n-2}) \\ & + \lambda_2^{<H_{n-1} H_{n-2}>} P^{<c>}_{ML}(w_n | w_{n-1}) \\ & + \lambda_1^{<H_{n-1} H_{n-2}>} P^{<c>}_{ML}(w_n) \end{aligned}$$

$$H_n = \begin{cases} 0 & (w_n \text{が内容語}) \\ 1 & (w_n \text{が非内容語}) \\ 2 & (w_n = \langle S \rangle) \\ 3 & (w_n = \langle E \rangle) \end{cases}$$

モデルのパラメータは削除補間法[2](学習データの分割数 4)によって求めた。 λ_i の求め方は次の通りである。

1. $\lambda_i (i=1 \sim 3) = 1/3$ とする。
2. 学習データ L を 4 個の部分データ $L_1 \dots L_4$ に分割する。
3. 各部分データ $L_1 \dots L_4$ を L_j として、ステップ 4,5 を行う。
4. 学習データから L_j を削除し、残りのデータから N グラムのパラメータを求める。
5. 削除されたデータ L_j を用いて、補間係数 λ_i を再推定する。 L_j から求められた補間係数

を $\lambda_i^{(j)}$ とする。

$$\begin{aligned}\lambda_3^{<H_{n-1}H_{n-2}>^{(j)}} &= \sum_{s \in L_j} \sum_{n=1}^{\text{length}(s)} \frac{\lambda_3 P_{ML}(w'_n | w'_{n-2} w'_{n-1})}{P(w'_n | w'_{n-2} w'_{n-1})} \\ \lambda_2^{<H_{n-1}H_{n-2}>^{(j)}} &= \sum_{s \in L_j} \sum_{n=1}^{\text{length}(s)} \frac{\lambda_2 P_{ML}(w'_n | w'_{n-1})}{P(w'_n | w'_{n-1})} \\ \lambda_1^{<H_{n-1}H_{n-2}>^{(j)}} &= \sum_{s \in L_j} \sum_{n=1}^{\text{length}(s)} \frac{\lambda_1 P_{ML}(w'_n)}{P(w'_n | w'_{n-2} w'_{n-1})}\end{aligned}$$

$\text{length}(s)$ は文 s の長さ (単語数) である。

6. $\lambda_i^{(j)}$ を部分データ全体について平均し、 λ_i を更新する

$$\lambda_i = \frac{1}{m} \sum_{j=1}^m \lambda_i^{(j)} \quad (i = 1, 2, 3)$$

7. 各 λ_i が収束するまで²ステップ 3～6を繰り返す。

以上の処理を両話者による学習テキストに対してそれぞれ行い、 M_N, M_{NN} を構築した。

3. 実験

3.1 学習・評価データ

本実験では母語話者及び非母語話者によって書かれたテキストとして表1のような内訳のテキスト292本を収集し、実験データとして使用した。但し、実験テキスト中には数式や表など文以外の要素も含まれている。これらは母語話者性の推定に不要であると考え、テキストに出現する文³のうち動詞を含まない文及び数値を含む文を

² 具体的には、対数尤度の平均変化量 <0.1 となるまで各 λ_i の学習を繰り返した。

³ 後述する TreeTagger によって品詞タグ付けを行い、文終端タグ‘SENT’によってテキストを文に分割した。但し、分割された文のうち先頭が大文字のアルファベットで始まっていないものは除去した。

非文として取り除く、簡単なフィルタリングを行った。

3.2 内容語から品詞への変換

本実験では実験に用いたテキスト 292 本を用いて単語 w の df を求めた。後で述べるように、本実験では実験テキストを4分割し、母語話者・非母語話者学習テキストセットをそれぞれ4セット(計8セット)作成した。非内容語の性質を考えると、これら8個のテキストセットすべてに出現している単語を非内容語の候補とするのが妥当である。そこで、8個の学習テキストを用いて非内容語の候補を選択し、そのうち df の高い上位 K 個の語を非内容語として残した。その他は内容語として品詞に置き換えた。テキスト中の単語に対する品詞タグ付けには TreeTagger[3]を使用した。

表 1 学習・評価データ内訳

母語話者 テキスト	海外の国際会議で発表された論文で著者の所属機関が USA であるもの(229 本、662991 文、1935468 語)
非母語話者 テキスト	日本の国際会議で発表された論文で著者が日本人であるもの(63 本、60266 文、195690 語)

3.3 実験結果及び考察

2.3 節で述べたスムージングが施された N-gram を構築し、モデルの評価を行った。評価は次のようなクロスバリデーションによって行った。

1. 収集した母語話者による論文 229 本及び非母語話者による論文 63 本をそれぞれ 4 個

- のグループ $L_{N1}, \dots, L_{N4}, L_{NN1} \dots L_{NN4}$ に分割する。
2. $i=1, \dots, 4$ に対して 3.~4.を行い、それぞれの再現率・適合率の平均を求める。
 3. L_{Ni} 及び L_{NNi} を除いたグループを用いて $M_N^{(i)}, M_{NN}^{(i)}$ のパラメタ推定 及び、 $P_C(N)$, $P_C(NN)$ の推定を行う。
 4. 各 $d \in L_{Ni} \cup L_{NNi}$ に対し、 $J(d)$ を求め、各 $\theta \in \Theta$ に対する(1)による判定の適合率・再現率を求める。 Θ は-1000~1000 の範囲で 0.5 刻みとした。

上記の処理を $K=0, 500, 1000$ に対してそれぞれ行ったときの適合率と再現率の関係を図 1 に、適合率を固定したときの K と再現率の関係を図2に示す。

図1及び図2から、品詞タグと少数の非内容語によってモデルを構築した場合 ($K=500$) に高い再現率・適合率を示している。適合率を 90%以上として母語話者のテキストを収集した場合には、再現率が 100%となり、非常に高い割合で母語話者のテキストを収集できることがわかる。また、適合率を 95%以上としても再現率は 96%を超えており、本研究の目的を十分に達成できた。

K が小さい場合、ほとんどの単語を内容語として品詞に変換するため、母語話者性を反映するような機能的な語まで品詞に変換される。つまり、推定に必要な情報が欠落するため、推定精度が下がると考えられる。一方、非内容語として扱う単語の異なり数を増やしてモデルを構築した場合、テキストは母語話者性を反映するのに有効に働く機能語と品詞タグの列に変換される。しかし、 K を大きくしすぎると、母語話者性を反映するのに有効でない単語が品詞タグに変換されずにテキスト中に残るため、推定精度が下がると考えられる。

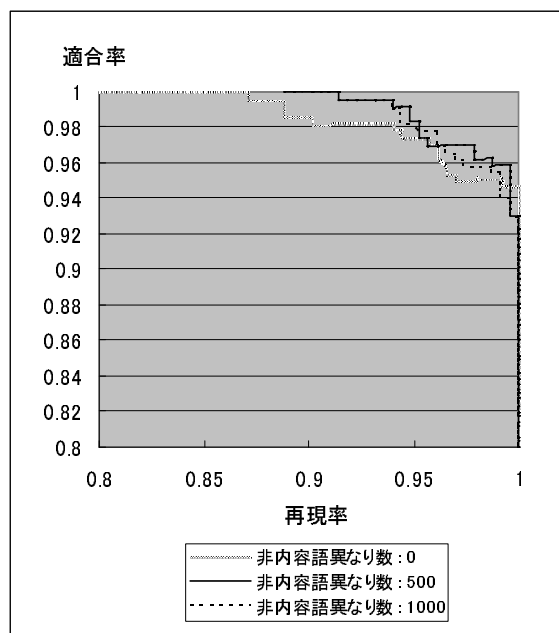


図 1 適合率・再現率

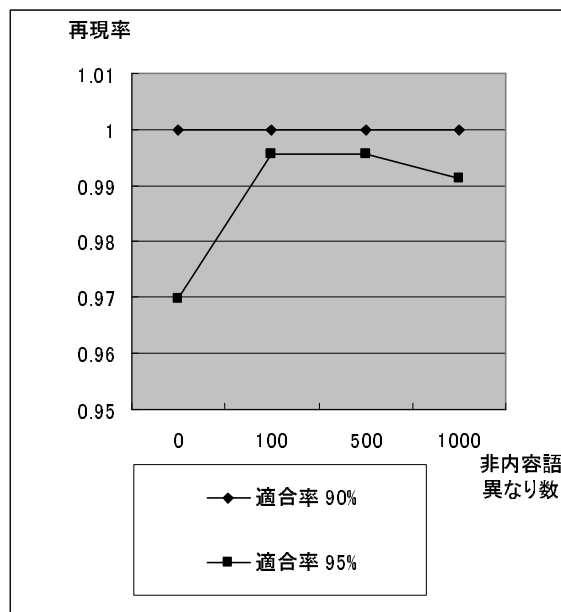


図 2 再現率(適合率固定)

4. 終わりに

英語母語話者性を推定する言語モデルの提案

を行い、評価実験を行った。実験の結果、少量の学習テキストでも比較的高い再現率・適合率を得られることが示された。今後学習データを増やして検討を行うと共に、今回収集したテキストの分野以外のテキストに対しても評価実験を行う予定である。

現在本モデルで利用していない情報として、

- 別々の文に出現した機能語の依存関係⁴
- 係り受け情報

といった、離れた単語間の依存関係に関する情報が挙げられる。これらの情報は今回利用した N-gram モデルと独立しており、これらの情報を利用することで精度の向上が期待できる。これらの情報を反映するモデルを構築して母語話者の推定を行うことが今後の課題である。

The Japan Times (2001).

参考文献:

- [1] Masahiro Shibata, Yoichi Tomiura, Shosaku Tanaka: A Method for Retrieving Translations of Collocation in Web Data. Asian Symposium on Natural Language Processing to Overcome Language Barriers, accepted.
- [2] 北研二: 確率的言語モデル. 東京大学出版会 (1999).
- [3] TreeTagger. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>
- [4] 齊藤俊雄, 中村純作, 赤野一郎: 英語コーパス言語学. 研究社 (1998).
- [5] 伊藤サム: ネイティブに通じる英語の書き方.

⁴ 例えば日本語母語話者の場合は「起承転結」の構成に基づいて文章を構成する傾向がある一方、英語母語話者の場合は「結論」を最初を書く傾向がある。そこで、機能語の一つである接続詞同士の依存関係を推定のための要素に取り入れることで、両話者の文書構成の違いを検出できる。

正誤表(NL・160・18)

	誤	正
p1 l4	情報基盤センター 研究部 外国語情報メディア 部門	情報基盤センター 外国語情報メディア研究 部門
p4 l20	229 本、662991 文、1935468 語	229 本、70277 文、2073349 語
p4 l24	63 本、60266 文、195690 語	63 本、7353 文、191473 語