2009－NL－190 (7)
2009／3／25

情報量基準に基づいた単語境界推定方式の提案

柳原 正[†1] 松本一則[†1]
池田和史[†1] 滝嶋康弘[†1]

自然言語処理で用いられる形態素解析において、品詞を特定できない文字列を未知語として分類する。しかし、これらの未知語は単語境界が正しく推定されていないことにより、結果的に品詞推定が正確に行えないことが多い。従来の解決方法では、文字間の接合する度合いを計るために、事前にコーパスから生成された n-gram 統計を使用していた。しかし、この手法では情報量の信頼性についての表現できなかったため、n-gram 統計の信頼性が低下してしまうという問題を抱えていた。そこで、本論文では、情報量の信頼性が保たれる情報量基準に基づいた単語境界推定方式を提案する。

Word Segmentation Estimation using Information Criteria

TADASHI YANAGIHARA,[†1] KAZUNORI MATSUMOTO,[†1] KAZUSHI IKEDA[†1]
and YASUHIRO TAKISHIMA[†1]

Morphological analysis used in natural language processing often find words which cannot be catagorized under any morphemes, which are often addressed as "unknown words". However, unknown words often occur due to incorrect estimation of word segmentation. Former methods suggest using a n-gram statistics created from a text corpus, but fail to present reliability within such information. In this paper, we propose a method based on information criteria, which guarantees reliability towards information found within such statistics.

1. はじめに

インターネットの発展によって、機械によって生成される大量のテキストデータが生成されている。このとき、テキストの意味を理解したり、他データと組み合わせるために加工したりなどを行うために、一般的には形態素解析が用いられる。これにより、これまで人手では難しかった大量のテキストから単語の抽出や品詞の推定が容易となった。

しかし、形態素解析を利用する前提条件として、形態素解析の際に用いる辞書に解析の対象となる文書の単語が多く登録されていることが必須である。もし辞書に登録されていない単語が含まれていた場合、これらの単語は形態素解析エンジンによって、単語境界及び品詞が特定できない未知語として検出されてしまう。未知語の数が多くなってしまうと、形態素解析を利用するサービスでは大きな問題となる。この問題に対し、一般的には未知語となった箇所を取り出し、人手によって形態素解析で用いる辞書へ手動で登録することで未知語を削減できるが、未知語の数が多くなるにつれ、対応が困難になるという欠点が挙げられる。

†1 株式会社 KDDI 研究所
KDDI R&D Incorporated

2. 提案手法

本論文では、情報量基準に基づいた単語境界の推定方式を提案する。具体的には、事前に与えられたコーパスにおいて、「任意の文字列 S において、n 文字目である文字 w_n が存在し、さらに $n+1$ 文字目に文字 w_{n+1} が存在したとき、w_n と w_{n+1} は接合すべきか」についての検証を行う。

情報量基準に基づいた単語間の接合度を計算するためには、Matsumoto[3] が提案するモデル検定の差分を利用した手法を採用する。具体的な内容としては、赤池情報量基準[2] に基づいたモデル検定を行った後に、独立関係と仮定した場合の AIC 値と従属関係と仮定した場合の AIC 値の差分を計算し、関連度の大きさとして扱う方式である。

この手法を用いて単語境界の推定を行うためには、以下の手順が必要となる：

1. AIC 値の差分を用いた文字間の接合度の計算結果である n-gram 統計の計算を行う。
2. 特定の条件に基づき、文字間のスコアが条件を満たした場合において、単語の区切りの判定を行う。

2.1 n-gram 統計の計算方法

文字間が接合するかを判断するために、文字間の関

連度を表す尺度が必要である。本論文では、赤池情報量基準に基づいた独立モデルと従属モデルと仮定したときのそれぞれの AIC 値の差分を文字間の接合度を表す尺度として表す。具体的な手順を以下に示す。

なお、以下の説明において、A と B は以下のように定義する：

- A: n 文字目に w_n が存在するという現象
- B: $n+1$ 文字目に w_{n+1} が存在する現象

AIC 値を計算するための 2×2 分割表の作成

任意の文字列 S において、n 文字目に文字 w_n があり、さらに $n+1$ 文字目に文字 w_{n+1} が存在したとき、2×2 分割表を作成する際に利用する変数を以下のように定義する：

- $a = N_{11}$: n 文字目に w_n が存在したとき、$n+1$ 文字目に w_{n+1} が存在した事例数
- $b = N_{12}$: n 文字目に w_n が存在したとき、$n+1$ 文字目に w_{n+1} が存在しなかった事例数
- $c = N_{21}$: $n+1$ 文字目に w_{n+1} が存在したとき、n 文字目に w_n が存在しなかった事例数
- $d = N_{22}$: $n+1$ 文字目に w_{n+1} が存在せず、n 文字目に w_n も存在しなかった事例数

このとき、以下の式が成り立つ：

- $i = a + b$: n 文字目に w_n が出現する事例数
- $j = c + d$: n 文字目に w_n が出現しない事例数
- $k = a + c$: $n+1$ 文字目に w_{n+1} が出現する事例数
- $l = b + d$: $n+1$ 文字目に w_{n+1} が出現しない事例数
- $z = a + b + c + d$: 全事例の総数

以下の表 1 において、それぞれの変数を記載した表を示す。

	w_{n+1}	$\neg w_{n+1}$	合計
w_n	a	b	i
$\neg w_n$	c	d	j
合計	k	l	1

表 1 2×2 分割表の実例

独立モデルと仮定したときの AIC(IM) の計算

モデル検定における独立モデルでは、A と B という現象が観測されたとき、これらの現象を決定付ける 2 つのパラメータである p と q が存在すると仮定する。このとき、p と q によって起こりうる現象をすべてを 1 としたとき、$p = 1-q$ 及び $q = 1-p$ で求められる、と仮定する。これにより、A という現象で変化があった際に、B という現象における値の変化に影響を与えないこととなる。以下の表 2 において、独立モデルを仮定したときの 2×2 分割表を示す。

本論文内では、独立モデルが成立するときは、「$n+1$ 文字目に w_{n+1} が存在したことによって、n 文字目に w_n が存在したこととは関連がない」と見なし、「接合すべきではない」と判定する。

従属モデルの AIC 値である $AIC(IM)$ を以下の式に

	B	$\neg B$	合計
A	pq	$p(1-q)$	p
$\neg A$	$(1-p)q$	$(1-p)(1-q)$	$1-p$
合計	q	$1-q$	1

表 2 独立モデルにおける 2×2 分割表の実例

基づいて計算する。

$$MLL = {}_i\log_i + {}_j\log_j + {}_{(z-i)}\log_{(z-i)} + {}_{(z-j)}\log_{(z-j)} - {}_z\log_z \quad (1)$$

$$AIC(IM) = -2 \times MLL + 2 \times 2 \quad (2)$$

従属モデルと仮定したときの $AIC(DM)$ の計算

モデル検定における従属モデルでは、A と B という現象が観測されたとき、これらの現象を決定付けるパラメータとして $p_{11}, p_{12}, p_{21}, p_{22}$ が存在すると仮定する。これらの 4 つの値により起こりうる現象をすべて観測が可能であると仮定したとき、$p_{11}+p_{12}+p_{21}+p_{22}=1$ という式が成り立つ。このとき、1 つの p は他の 3 つの p を 1 から差し引くことで値が決定される。このため、パラメータ数は 3 つ存在すると仮定できる。以下の表 3 において、従属モデルを仮定したときの 2×2 分割表を示す。

	B	$\neg B$	合計
A	p_{11}	p_{12}	p
$\neg A$	p_{21}	p_{22}	$1-p$
合計	q	$1-q$	1

表 3 従属モデルにおける 2×2 分割表の実例

本論文内では、従属モデルが成立するときは、「$n+1$ 文字目に w_{n+1} が存在したことによって、n 文字目に w_n が存在するかどうかと関連する」と見なし、「接合すべき」と判定する。

従属モデルの AIC 値である $AIC(DM)$ を以下の式に基づいて計算する。

$$MLL = {}_a\log_a + {}_b\log_b + {}_c\log_c + {}_d\log_d - {}_z\log_z \quad (3)$$

$$AIC(DM) = -2 \times MLL + 2 \times 3 \quad (4)$$

文字間の接合度 $E(w_n)$ の計算

$AIC(IM)$ 及び $AIC(DM)$ が求まり次第、それらの値の差分を計算する。モデル検定では、適合度が高いモデルはより小さいスコアが求める性質を持つ。この性質を利用し、独立モデルと比べ、従属モデルが小さい事例では高いスコアが計算されるようにする。以下の計算式により、接合度 $E(w_n)$ を求める。

$$\frac{a}{a+b} > \frac{c}{c+d} \rightarrow E_{w_n} = AIC(IM) - AIC(DM) \quad (5)$$

$$\frac{a}{a+b} < \frac{c}{c+d} \rightarrow E_{w_n} = AIC(DM) - AIC(IM) \quad (6)$$

これにより、任意の文字列 S において：

- 「n 文字目が w_n であり、$n+1$ 文字目が w_{n+1} である場合は、互いに接合する可能性が高い」と判定されるため、E_{w_n} は大きい値を示す。
- 「n 文字目が w_n はなく、$n+1$ 文字目が w_{n+1} ではない場合は、互いに接合する可能性が低い」と判定されるため、E_{w_n} は大きい値を示す。

という E_n が求まる。

$n=2$ 以上の n-gram 統計情報の構築方法

これまで示してきた例では、任意の文字列である S のうち、文字 w_n と文字 w_{n+1} の間の関連性についてのモデル検証を行った。uni-gram だけでなく、他に $n \geq 1$ である n-gram 統計モデルを使う場合も考えられる。この場合は、文字 w_n を文字列 s_n、文字 w_{n+1} を文字列 s_{n+1} と見なすことで、bi-gram 以上の統計モデルが構築可能となる。

2.2 単語境界の決定方法

各文字 w_n とそれらの文字に後続する文字との接合度 $E(w_n)$ が求まったあと、以下の 2 方式のうちのいずれかの手順に従って、単語境界を推定する。

A. 閾値による推定方式

B. 隣接する文字との相対的な関係による推定方式

A. 閾値による推定方式

閾値である α を定義し、文字間の関連度を表す接合度 $E(w_n)$ が α よりも大きい場合[*1]、文字間は接合すると判定する方式が考えられる。事前に学習データから単語境界が観測可能な事例を選び出し、それらの例における文字間の $E(w_n)$ を求めたあと、例に付加された観測可能な単語境界を尤もらしく接合する α を決める。あとは α の値に基づき、入力となる文字列に対し、単語境界の推定を行うことができる。

B. 隣接する文字との相対的に関係による推定方式

文字 w_n があったとき、その前後に隣接する文字である w_{n-1} の $E(w_{n-1})$ 及び w_{n+1} の $E(w_{n+1})$ と比較し、

$$E(w_n) < E(w_{n-1}) \cap E(w_n < E(w_{n+1})$$

という条件を満たす場合に限り、「w_n は w_{n+1} と接合する」と判定する方式も考えられる。本方式のメリットとしては、α の閾値が不要となることが挙げられる。

具 体 例

以下に方式 A. 及び方式 B. の動作手順を「マジでヤバい」を文字列 S として使って説明する。初めに、それぞれの s_n についての N_{11}〜N_{21} についての統計情報を求める。本例では uni-gram 以上の統計情報を扱うため、文字 w_n を文字列 s_n として表し、また、n-gram の統計情報のすべての項目は s_n で始まる項目順にソートされているものとする。(例：マ→マジ→マジで…)

s_n	s_{n+1}	N_{11}	N_{12}	N_{21}	N_{22}	E_{s_n}
マ	ジ	10,052	182,059	121,505	129,815,060	60,898.60
マジ	で	3,743	6,309	2,264,286	132,139,900	17,503.73
マジで	ヤ	84	3,659	58,396	129,940,319	492.15
マジでヤ	バ	79	5	176,391	130,061,973	1,003.57
マジでヤバ	い	49	30	3,506,426	133,392,013	253.79
ジ	で	6,295	125,262	2,261,734	132,015,843	5,107.49
ジで	ヤ	85	6,210	58,395	129,937,766	413.16
ジでヤ	バ	79	6	176,391	130,061,972	998.1
ジでヤバ	い	49	30	3,506,426	133,392,013	253.79
で	ヤ	668	2,267,361	57,812	127,675,449	138.9
でヤ	バ	241	427	176,229	130,061,227	2,309.03
ヤ	バ	8,751	49,729	167,719	129,994,905	66,780.17
でヤバ	い	122	119	3,506,353	133,391,778	564.29
ヤバ	い	4,549	4,202	3,501,926	133,378,841	21,444.05
バ	い	4,627	171,843	3,501,848	133,211,044	0.57

表 4 S = "マジでヤバい" のときの n-gram 統計情報に関する一覧表

方式 A. では、閾値 α を設定してから、単語境界の推定を行う。例えば、$\alpha = 6000$ と定義したとき、以下の動作手順に従う。

1. 「マ」⇒「ジ」: $E_{\text{マ}} = 60,898.60 > \alpha \rightarrow$ ○、「マ」と「ジ」は接合すると仮定し、s_n = 'マジ' と見なしたときの $E_{\text{マジ}}$ を調べる。
2. 「マジ」⇒「で」: $E_{\text{マジ}} = 17,503.73 > \alpha \rightarrow \times$、「マジ」は一つの単語と見なす。
3. 「で」⇒「ヤ」: $E_{\text{で}} = 138.9 < \alpha \rightarrow \times$、これにより、「で」は一つの単語であると見なす。
4. 「ヤ」⇒「バ」: $E_{\text{ヤ}} = 66780.17 > \alpha \Rightarrow$ ○
5. 「い」: s_{n+1} が存在しないため、「い」は一つの単語であると見なす。

以上の手順により、「マジ / で / ヤバ / い」という単語境界が得られる。

方式 B. では、s_n を与えられたとき、E_{n-1} と E_{n+1} のそれぞれの値を E_n と比較し、$E_n < E_{n-1} \cap E_n \mathrel{¡} E_{n+1}$ が成り立つときに接合すると判定する。

1. 「マ」: $E_{\text{マ}} = 60898.6$、$E_{s_{n-1}}$ は存在せず、$E_{\text{ジ}} = 5107.49$。このとき、$E_{s_{n-1}} \neq E_{s_n} > E_{s_{n+1}}$ となり、単語境界は「ない」とみなすことができる。
2. 「ジ」: $E_{\text{ジ}} = 5107.49$、$E_{\text{マ}} = 60,898.60$、$E_{\text{で}} = 138.9$。このとき、$E_{s_{n-1}} > E_{s_n} > E_{s_{n+1}}$ となり、単語境界は「ない」とみなすことができる。
3. 「で」: $E_{\text{で}} = 138.9$、$E_{\text{ジ}} = 5107.49$、$E_{\text{ヤ}} = 66780.17$。このとき、$E_{s_{n-1}} > E_{s_n} < E_{s_{n+1}}$ となり、単語境界は「ある」とみなすことができる。
4. 「ヤ」: $E_{\text{ヤ}} = 66780.17$、$E_{\text{で}} = 138.9$、$E_{\text{い}}$ は存在しない。このとき、$E_{s_{n-1}} > E_{s_n} \neq E_{s_{n+1}}$ となり、単語境界は「ない」とみなすことができる。

---

*1 接合度の計算方法によっては、小さい場合も有り得る。

3. 関連研究との比較

人手に頼らず、自動的に単語境界を推定する方式として、森と長尾[4] が提案する手法が挙げられる。この手法では、事前にコーパスから作成した文字単位の n-gram 統計を基に、文字列を与えられると、文字の前後との文字の尤度を推定する方式を提案している。

しかし、異なる点として、文字間の関連度を計算する際には、以下の確率の出現回数を使用している。具体的には、文字間が同時に成立する確率である $P(A;B)$ は以下のように計算している。

$$P(A;B) = \frac{n_{11}}{n_{11}+n_{12}} \tag{7}$$

確率を用いた場合では、値の正規化が行われるため、n-gram 統計を作成する際の情報の信頼性が失われてしまう。例えば、学習データを 10 倍増やし、n_{11}〜n_{21} のそれぞれの値がちょうど 10 倍ずつ増えたとする。学習データが増加した場合、統計の信頼度が向上しているが、$R(A;B)$ の値は変わらないという現象が発生してしまう。このため、情報量の信頼性を示すための計算方法を用いることが望ましいと言える。

4. 評　　価

本方式の有効性を示すため、インターネットから取得したブログ記事を学習データとした n-gram 統計 $(1 \leq n \leq 5)$ を構築し、さらに別途のブログ記事から形態素解析エンジンによって未知語と判定された文字列を評価データとした。これらのデータを基に、単語境界の推定に対する精度を計測する。

4.1 使用データ

本実験を行う際に使用した学習データと評価データは以下の通りである。

- **学習データ**：インターネット上で公開されているブログ記事 計 50 万件 (約 1.3 億文字)
- **評価データ**：インターネット上で公開されているブログ記事 計 10 万件に対し、mecab[1]+ipadic (2.7.0-20070801) を使い、抽出された未知語 100 件

なお、実際に利用した評価データは論文の付録として添付した。単語境界が推定されるべき箇所に対し "||" という目印を入れた。

4.2 評 価 尺 度

精度を計測するために、一般的に情報検索で用いられる適合率 (*precision*) と再現率 (*recall*), F 値 (*f-measure*) を利用する。

- N_T = 推測された単語境界の数
- N_E = 実際に存在した単語境界の数
- $N_C = N_T$ のうち、N_E を満たす単語境界の数

それぞれの尺度の計算方法については、以下のように定義する。

$$precision = \frac{N_C}{N_E}$$

$$recall = \frac{N_C}{N_T}$$

$$f\text{-}measure = \frac{2 \times precision \times recall}{precision + recall}$$

但し、区切りが初めから含まれていない文字列であり、且つ、単語境界が存在しないと判定した場合 $(N_T = 0 \cap N_E = 0)$ に限り、*precision* と *recall* はともに 1 と見なす。

4.3 実 験 結 果

以下に、評価データに対し、単語境界を推定した後、ラベリングした単語境界の箇所との一致した数から求まった *precision* 及び *recall* を表 5 に、グラフ化した結果を図 1 に示す。

method	*precision*	*recall*	*f-measure*
AIC-T (α = 1000)	0.1383	0.9067	0.1519
AIC-T (α = 2000)	0.1208	0.92	0.1381
AIC-T (α = 3000)	0.1256	0.965	0.1468
AIC-T (α = 4000)	0.1231	0.965	0.1440
AIC-T (α = 5000)	0.1264	0.97	0.1474
AIC-V	0.54	0.8383	0.4966

表 5　各方式における *precision* 及び *recall* に関する表

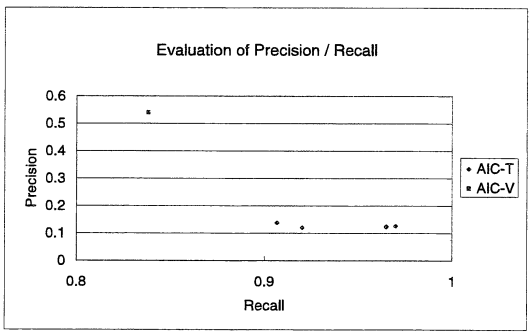

図 1　評価結果

AIC-T の α を増加させた場合、*recall* が 1 に近づくと共に、*precision* の低下があまり確認されなかった。これにより、α を増加させた場合のペナルティは少ないため、影響力を受けにくいと言える。しかし、総合的に α の値に関係なく、*precision* の値は低いため、有効な手法であるとはいえない。一方、*f-measure* の尺度においては、方式 B. である AIC-V を用いた場合が良

い結果を示した。これにより、現行方式では、AIC-Vがより有効であると言える。

5. ま と め

本論文では、情報量基準を使った単語境界推定方式の提案を行った。情報量基準を使ったモデル検定によって、文字間の接合度が求めることができ、この結果によって *n*-gram 統計が構築可能である。2 種類の単語境界方式を提案し、そのうち、対象語の前後に存在する文字列との接合度を比較することで、単語境界を推定する上で有効であることがわかった。今後は単語境界の決定方式をさらに検証することで、より精度の高い単語境界の推定方式を確立する。

参 考 文 献

1) MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/.
2) Hirotsugu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory*, 1973.
3) Kazunori Matsumoto and Kazuo Hashimoto. Schema Design for Causal Law Mining from Incomplete Database. In *Discovery Science, Second International Conference*, Vol. Lecture Notes in Computer Science 1721 Springer, pp. 92–102, 1999.
4) 森信介, 長尾眞. *n* グラム統計によるコーパスからの未知語抽出. 社団法人情報処理学会論文誌, Vol.95, No.69, pp. 7–12, 1998.

付録: 評価用データとして使用した未知語 100 件の一覧

ア・ラ・リ	フアッション
マイチェン	ハシカ
エエエェエエ	カコヨス
スノー||ラメ	Ｌｅｏｎ
スカポンタン	ダブルデート
レピッシュ	ヂーパン
ヨーカン	キラヤマト
ＣＤＢＯＸ	ライム||ピール
（；≧	Ｔｅｅｎ
マーゴン	リアクリョン
Ｌｉｃａ	ハジメル||ヨ
ぉぃしぃ||ょん	ぐはははは
シャリーン	ガンモドキ
ｐｏｓｃａｍ	イイー||ネッ
フーセン	ャンキー
ＺＡＫＺＡＫ	ズゴック
モエ||レ	ＢＬＡＮＫＳ
ぁぁぁあああ	ナン||ダ||ソレ
ドラエモン	ｉＴｕｎｅ
ワザワザ	ロングヘアー
マジ||ショック	ワーイ||ワーイ
アリャー	ＬＬＬｌ
スポンジ||デブ	Ｃｏｒｄ
タモ||レシピ	アリガタヤー
ＢｉＲＴＨ	ぱしょこん||の
ハンゾー	マー||クン
ＩＭＤｂ	ホリエー
バザーイ	キャワイ
マジ||モン	ぃ||た||ょう||です
￣ ┌)	ぉもしろ
ナイトメアー	アピーリング
マタドガス	ぁ||ない||って
ｂｌＯｇ	ベンケー
ドルルルルン	メンドウ
バレネーヨ	ピーサン
ヨク||ナイ||ケド	ドッグ||カフェ
ジョギーング	マシ||タ
ジァグラー	ドス||ビス||カス
コイビト	∀。□)
チル||アウト	ヴー||ヴー
−）−□□	イー||ペーコー
トィ||トィ||サン	ジーパン
チクショウ	グォー||グォー
イコライザ	コービン
おおおおお	ぴぇぇ
ジオング	ウネリ
セシル	ヘヴィ
ＮＰＣ	ＡｉＭ
ウラヌス	ラチェット
ＤＯＯＲ	□≦)