統計情報を用いた日本語形態素解析

山田洋志（h-yamada@hum.cl.nec.co.jp）
NEC 情報メディア研究所

日本語形態素解析において辞書未登録語による誤りは解析誤りの中で大きな割合を占めている。未登録語がある場合,解析が失敗する場合だけでなく,短い単語の羅列として解析に成功する場合があり,未登録語箇所を推定する必要がある。本稿では,形態素解析結果中に含まれる未登録語の検出と範囲推定に,コーパスから抽出した統計情報を利用する方式について,性能及び従来方式との比較結果を報告する。本方式によって,解析時には検出できない未登録語の42～52%について,69～83%の精度で検出できた。

1 はじめに

日本語のテキストを単語に分割する形態素解析は,日本語を扱うアプリケーションの基礎となる技術で,盛んに研究開発が行われている。形態素解析の精度は,アプリケーション全体の性能に影響するため,より高い精度が求められている。

形態素解析技術には,辞書・文法を中心とした方式と,コーパス・統計を中心とした方式があり,それぞれ特長を持っている。現状では,辞書・文法ベースの方式の方が高い精度を実現しているが,次第に精度の向上が難しくなってきている。

筆者らは,従来開発してきた辞書・文法ベースの形態素解析の精度を向上させるために,局所的に統計処理を導入することを試みた。未登録語の解析誤りを回復するための規則を,学習用の解析結果から取り出し,従来の解析結果を修正するのに利用する。本稿では,未登録語推定規則の抽出方式と,評価結果を示す。

2 従来方式と問題点

従来の日本語形態素解析処理は,単語辞書と文法規則を基本とし,さらに単語の共起や構文規則などの知識を利用して90%を超える精度を実現しており,99%を超える高い精度を達成しているものもある[1]。従来システムでは,精度を上げようとすると各種辞書,文法の見直しが必要になる。しかし,精度が高くなるほど改良の効果が限定され,場合によっては逆効果になる。そのような背景から,保守性に着目したシステムも研究されている[2]。

一方,統計や確率を利用した形態素解析技術が開発されており,学習用のデータから情報を取り出して利用する。

学習に形態素解析したデータを使用する場合,解析結果から2語あるいは3語の連接確率を学習する[3, 4]。テキストを解析する際は,単語列の生起確率が最大になる単語分割を選択する。解析結果データの作成には人間による校正が必要なため,大量のデータを用意するのは困難である。それを補うため,品詞の生起確率と単語の生起確率を別に計算したり,低次のN-gramの組み合わせで高次のN-gramを補完したりする方式が使われている。

テキストを利用する場合は,文字や文字列の出現頻度を利用する[5]。解析したいテキストが与えられると,単語列の出現頻度の和が最大になる単語分割を選択する。この方式は,豊富にあるテキストがそのまま利用できるのが長所であるが,十分な精度が実現できていないのが現状である。

3 未登録語抽出規則の獲得

辞書・文法ベースの形態素解析の未登録語処理に局所的に統計処理を導入する方式を開発した。

未登録語は解析誤り原因の中で大きな割合を占めており,未登録語処理の改良は解析精度の向上に効果が大きい。辞書の改良による精度向上も行っているが,固有名詞や外来語など語彙は莫大な数にのぼり,しかも,常に変化しているため,未登録語の存在は避けられない。そこで,コーパスから獲得した規則を利用して,未登録語の箇所と単語範囲を判別し,解析結果を修正する。

未登録語を含むテキストを解析すると,解析に失敗する場合と,登録されている単語の組み合わせとして解析される場合とがある。前者に対して

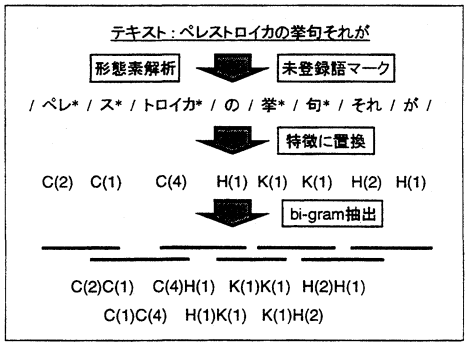

図 1: N-gramの抽出

は字種や前後の文字列から未登録語の範囲を推定する方式が提案されているが,後者では,まず,どこが未登録語であるかを推定する必要がある。以前,未登録語に特有のパタンを利用して解析結果を修正する方式を提案した[6]が,パタンの選択は経験と勘に基づいていた。今回の方式では,コーパスを利用して,未登録語推定の規則を作成する。

推定規則の取り出し方は以下のとおり(図1)。

1. コーパス(学習データ)の作成
   テキストを形態素解析し,解析結果に含まれる未登録語箇所を人手でマークする。
2. 解析結果から単語のN-gramを取り出し,含まれる単語の特徴で分類する。今回の実験では,システムへの依存の度合いが少なく,一般的な特徴として,単語の字種と単語長を用いた。
3. 各分類に含まれるN-gramごとに未登録語であるかどうかをマークで判定し,分類に含まれているときに未登録語になる条件付き確率を計算する。
4. 確率が大きいものから分類を選び,未登録語推定規則とする。

取り出した推定規則は以下のように利用する。

1. 形態素解析を実行する。
2. 解析結果と推定規則と照合する。
3. 規則にマッチした箇所の単語区切りを修正し,未登録語属性を付加する。

4 未登録語推定実験

前節の方式に基づいて未登録語の箇所を推定する実験を行った。

表 1: 実験用テキスト

	文数	字数	語数	未登録語箇所	
				既知	未知
1	1,500	70,754	41,323	418	709
2	1,500	70,881	41,144	436	723

4.1 実験手順

1. コーパスの作成
   日本語テキスト3000文(表1)を形態素解析し,結果に含まれる未登録語をマークした。表1で'既知'は解析時に検出された未登録語数,'未知'は見かけ上解析に成功した未登録語数である。テキスト1を規則抽出に使用し,テキスト2は性能評価のみに使用する。

   実験に使用したシステムは,未登録語によって解析に失敗した場合,失敗した位置の文字を未登録語として登録し,次の文字位置から解析を再開する。同一字種の未登録語が連続した場合はひとつにまとめる。

2. 解析結果から単語のbi-gram(2単語の連続)を取り出し,分類する。分類の条件としては,単語の字種(漢字,ひらがな,カタカナ,英字,数字,句読点,その他の文字)と単語長(文字数),bi-gramの前後の単語の特徴を用いた。分類の細かさを変えて以下の5レベルに分類した。

   (1) bi-gramの字種だけで分類
   (2) bi-gramの字種と長さで分類
   (3) bi-gramと直後の単語の字種で分類
   (4) bi-gramと直前の単語の字種で分類
   (5) bi-gramと前後の単語の字種で分類

3. 各分類に含まれるbi-gramが未登録語である条件付き確率を計算する。

$$P(\text{未登録語}|\text{分類}) = \frac{\text{未登録語の}bi-gram\text{数}}{\text{分類中の}bi-gram\text{数}}$$

4. 未登録語推定規則の抽出
   確率のしきい値として,0.7,0.75,0.8の3通りを使用した。レベル1の分類から順にしきい値以上の確率を持つ分類を規則として選択した。下位レベルから選ばれる分類がすでに

表 2: 抽出規則の例

規則名	level	内容
xCCxx12	2	1-2文字のカタカナ語連続
xKKPx11	3	1字漢語連続の後に句読点
HKKxx11	4	1字漢語連続の前に平仮名
KKKSx11	5	1字漢語の3連続の後に記号

選択されている分類の一部の場合は省いた。また,5個以下のbi-gramしか含まない分類を除外しての実験も行った。 実際に抽出された規則の一部を表2に示す。

5. 形態素解析への規則の導入
テキスト1とテキスト2とを解析し,抽出規則に一致するパタンが出現した場合に,未登録語属性を付加した。複数の規則が当てはまる場合は独立に適用し,ひとつでも当てはまる場合は未登録語とした。

4.2 実験結果

実験結果を表3,表4に示す。表中の'対象','検出','もれ','過剰'は,テキスト中の未登録語数,正しく未登録語と判定した語数,未登録語と判定できなかった語数,未登録語ではないのに未登録語と判定した語数である。システムが辞書引きや接続検定の段階で未登録語と判定した箇所は対象としていない。'再現率','適合率'は以下の式で求める。

$$\text{再現率} = \frac{\text{検出}}{\text{対象}} \qquad \text{適合率} = \frac{\text{検出}}{\text{検出}+\text{過剰}}$$

'差分'は'検出'と'過剰'の差で,未登録語推定規則によって得られる差し引きの正解数である。

比較のため以下の従来方式での結果も示す。

従来方式1 解析時の未登録語の前後に同一字種の単語がある場合に未登録語とする。ただし,1-2文字のカタカナと1文字の漢字のみを対象とした。

従来方式2 1-2文字のカタカナあるいは1文字の漢字が連続する場合に未登録語とする。

5 まとめ

実験結果のまとめと検討課題をあげる。

表 5: 未知語推定誤り例

推定前	推定後
/細胞/内/小/器官/	→ /細胞/内小/器官/
/鈴木/氏/不/支持/の/	→ /鈴木/氏不/支持/の/
/鎖/状/に/	→ /鎖状/に/

テキスト2に対しては,テキスト1と比較して10-20%精度が落ちている。過剰に検出する原因としては接辞を未知語と誤る例が多い(表5)。特に,出現頻度で規則を絞り込まなかった場合(表4下段)に精度が悪く,頻度が少ないと規則の信頼性が低く精度も悪くなることが実証されている。逆に,今回の方式のままでも,より大きなコーパスから規則を抽出することで精度が上がる可能性がある。今回の評価ではひらがなの未登録語に対して有効な規則が見つからなかった。今後の課題とする。

従来方式との比較では,従来方式1は解析時の未登録語箇所を手掛かりとするため再現率が非常に低くなる。ただし適合率が高さを利用するため,規則の条件として解析時の未登録語情報を取り入れることを検討したい。従来方式2は,使用した条件が本方式のしきい値70%で抽出したときの規則と類似していたために似た結果になっているが,よりしきい値を高くした場合には本方式の方が高い精度を達成している。

未登録語抽出精度向上のための手段として以下を検討する。

- コーパスの増強
今回,学習に使用したコーパスは1500文と少ないため各規則の信頼性が低くなっている。また,テキストの分野による性能の変化についても評価する必要がある。

- 規則に使う特徴の追加
現在使用している特徴に加え,品詞や解析時の未登録情報を利用する。bi-gramをtri-gram以上に拡張しての評価も実施したい。

- 規則抽出条件の決定方法の検討
各種しきい値の決定方法や条件付き確率に変わる統計量の導入など。

表 3: 実験結果(テキスト1)

しきい値			未登録語				検出精度		
頻度	確率	規則数	対象	検出	もれ	過剰	再現率	適合率	差分
6以上	0.7	7	709	468	241	181	66.0	72.1	287
	0.75	12		429	280	100	60.5	81.1	329
	0.8	11		410	299	74	57.8	84.7	336
1以上	0.7	50	709	515	194	188	72.6	73.3	327
	0.75	51		479	230	107	67.6	81.7	372
	0.8	51		459	250	76	64.7	85.8	383
従来方式1			709	152	557	12	21.4	92.7	140
従来方式2			709	356	353	137	50.2	72.2	219

表 4: 実験結果(テキスト2)

しきい値			未登録語				検出精度		
頻度	確率	規則数	対象	検出	もれ	過剰	再現率	適合率	差分
6以上	0.7	7	723	374	349	165	51.7	69.4	209
	0.75	12		335	388	78	46.3	81.1	257
	0.8	11		304	419	60	42.0	83.5	244
1以上	0.7	50	723	388	335	204	53.7	65.5	184
	0.75	51		349	374	122	48.3	74.1	227
	0.8	51		318	405	99	44.0	76.3	219
従来方式1			723	171	542	7	23.7	96.1	164
従来方式2			723	300	421	107	41.5	71.7	193

6 おわりに

未登録語の推定のために,学習用データから抽出した規則を使用する形態素解析方式の提案と性能評価を行った。本方式によって,解析時には検出できない未登録語の42〜52%について,69〜83%の精度で検出できた。

今後,規則抽出方式に改良を加えることで,精度を向上させるとともに,より大きなコーパスからの学習による評価を行う。また,未登録語を判別するだけでなく,品詞推定処理とも組み合わせたい。さらに,同一の方式を未登録語以外の解析誤りにも応用することも検討する。

参考文献

[1] 白井,横尾,池原,奥山,宮崎: “多段解析法による日本語形態素解析の精度”, 情処50回全国大会,1R-2 (1995)

[2] 渕,松岡,高木: “保守性を考慮した日本語形態素解析システム”, 自然言語処理研究会,NL-117-9 (1997)

[3] 永田: “前向きDP後向きA*アルゴリズムを用いた確率的日本語形態素解析システム”, 自然言語処理研究会 NL-101-10 (1994)

[4] 森, 長尾: “形態素bi-gramと品詞bi-gramの重ね合わせによる形態素解析”, 自然言語処理研究会 NL-112-6 (1996)

[5] 中渡瀬: “正規化表現による形態素境界の推定”, 自然言語処理研究会 NL-113-3 (1996)

[6] 福島ほか: “校正支援システムSt.WORDSの文書検査機能”, 情処46回全国大会,3L-3 (1993)