文脈にもとづく未知語獲得における識別モデルの適用

鍛治伸裕　　喜連川優

東京大学 生産技術研究所
{kaji, kitsure}@tkl.iis.u-tokyo.ac.jp

1　はじめに

日本語テキストでは単語の境界が明示されない．そのため日本語処理においては，テキストを分かち書きすることが最も基本的で重要な解析処理となる．

一般的な日本語の分かち書き器は解析用の辞書を用いる．現在，分かち書き処理は，大半の単語が解析辞書に登録されているようなテキストであれば，約 99%という高い解析精度を達成できることが知られている [2]．しかしながら，その一方で，未知語の多いテキストは頑健に解析することができない．これは，例えばウェブテキストを対象とするときに無視できない問題となる．近年ではウェブが言語処理の対象として注目を集めているが，ウェブテキストには固有名詞や新造語などの未知語が頻繁に出現するため，高い精度で解析を行うことが難しい．

未知語の問題はこれまでにも多くの研究者の間で議論の対象となってきた [3]．特に最近では，言語学研究の支援を目的とした未知語獲得も議論されている [5]．未知語には「ググる」や「メタボる」などの新造語が存在するが，大規模コーパスを用いて新造語の言語学的分析を行うためには，そのような未知語を正しく解析する必要がある．

未知語の問題に対処する一つの方法は，コーパスから未知語を自動獲得して，既存の解析辞書を拡張することである．未知語の獲得は，もし未知語候補が与えられた場合には，分類問題と考えることができる．本論文では，未知語候補を効率的に生成して，識別モデルにもとづく未知語獲得を行う方法について述べる．

2　文脈にもとづく未知語獲得

2.1　問題設定

まずは本論文で扱う問題を整理する．本論文では，コーパスと辞書が与えられたとき，そのコーパスから

普通名詞，サ変名詞，母音動詞，子音動詞カ行 子音動詞サ行，子音動詞タ行，子音動詞ハ行 子音動詞マ行，子音動詞ラ行，子音動詞ワ行 子音動詞ザ変，イ形容詞，ナ形容詞

表 1: 未知語に割り当てる品詞

未知語 (辞書に登録されていない単語) を抽出して，それに適切な品詞を割り当てる問題を未知語獲得と呼ぶ．ただし，動詞などの活用語は，活用形によって表層形が変化するため語幹を考える．例えば，動詞「ググる」の場合は語幹「ググ」を抽出する．

未知語に割り当てる品詞は，先行研究 [7, 6, 4] を参考にして，表 1 の 13 種類とした．ここに記載されていない品詞は獲得対象としない．品詞の定義は，基本的には JUMAN 辞書の品詞細分類に従っているが，活用語に対しては活用型を品詞とみなしている．

2.2　文脈情報の利用

未知語獲得を行うための方法として，これまでに文脈情報の利用が提案されている [3]．これはいわゆる分布類似度と同様の考え方を未知語獲得に適用したものであり，ある文字列が未知語であるかどうかを判定するために，その前後に出現する文字列の分布を利用するというものである．具体例として，以下のテキストから普通名詞「ＸＯ醤」を抽出する場合を考える．

(1) a. たくさんのＸＯ醤をゲットしました．
　　b. ＸＯ醤などを軽くいためて香りを出す．
　　c. そのＸＯ醤は，１０年前に発明された．

「を」「などを」「は,」は普通名詞の直後に現れやすい文字 (列) である．さらに「たくさんの」「その」は普通名詞の直前に現れやすい．そのため「ＸＯ醤」は普通名詞であると推測することが可能である．例にあげた「たくさんの」や「を」などのように，ある品詞 t

の直前または直後に出現しやすい文字列のことを，品詞 t の弁別的文字列と呼ぶ．特に出現位置を区別するときには，弁別的先行文字列または弁別的後続文字列と呼ぶ．

3 識別モデルの適用

今，何らかの方法で，品詞 t の未知語候補である文字列が与えられたとする．そうすると未知語獲得とは，その候補が品詞 t を割り当て可能な単語であれば $+1$，そうでなければ -1 を出力する分類器 C_t を学習する問題と考えることができる．分類器は品詞の数 (表 1 参照) だけ構築する．

3.1 未知語候補の生成

未知語の候補を生成する単純な方法は，コーパスに出現した全ての部分文字列を未知語候補とすることである．しかし，大規模なコーパスを対象する場合，そのような方法は現実的ではない．

そこで，品詞 t の弁別的先行文字列と弁別的後続文字列を用いて，分類器 C_t に与える未知語候補を生成する．まず各品詞について長さ n の弁別的文字列を用意する．そして，品詞 t の弁別的先行文字列と弁別的後続文字列に囲まれて出現した全ての文字列を品詞 t の未知語候補とする．n の値をある程度大きな値に調整することによって，大規模なコーパスからでも効率的に候補を列挙することが可能となる．これに加えて，発火している素性 (次節を参照) の数が少ないものは，そもそも正しく分類するための情報が少ない可能性が高いので，素性数が σ 以下のものは候補から外す．

品詞 t の弁別的先行文字列は，辞書を用いてコーパスから自動獲得する．まず文字列 p に対して次のスコアを定義する．

$$coverage(p) \quad = \quad |\{w \in \mathcal{W}_t | 0 < f(pw)\}|$$

式中の $f(pw)$ は文字列 pw の頻度であり，$\mathcal{W}_t$ は品詞 t が割り当てられている辞書登録語 (活用語の場合は語幹) の集合である．$coverage(p)$ は，p が何種類の辞書登録語の直前に出現したかを表しており，この値が大きいほど p は品詞 t に特徴的と言える．そのため $coverage(p)$ が閾値[1]を越える p を品詞 t の弁別的先行文字列とする．弁別的後続文字列についても同様である．

この処理だけでも，各品詞の弁別的文字列を取得することは可能であるが，活用語である動詞と形容詞の処理には，次の 2 つのヒューリスティクスを追加することができる．まず，活用型によって弁別的先行文字列が大きく異なるとは考えにくいので，弁別的先行文字列を取得するときには，動詞と形容詞の活用型は区別しないで考える．次に，弁別的後続文字列を獲得するさいには，その活用規則を制約として使う．

3.2 動的絞り込み

このように生成された未知語候補をさらに絞り込む．まずは以下のテキストを考える．

(2) これまでは心配だったのですが．

このテキストからは「心配」がナ形容詞語幹の候補と抽出される．しかし「たのですが」という文字列が母音動詞に特徴的な後続文字列であるため，同時に「心配だっ」も母音動詞語幹の候補として抽出されてしまう．このように重複する候補が同時に抽出された場合，少なくともどちらか一方の候補は誤っている．もし，ナ形容詞語幹「心配」が既知の単語であれば，後者のほうが誤って抽出されていると判断することができる．このとき，誤った候補の抽出に使われた弁別的文字列対ことを不適格な弁別的文字列対と呼ぶ．

上記のことを利用すると候補の絞り込みを行うことが可能である．ある候補 c がコーパスから抽出されたときに使われた弁別的文字列対の集合を考える．その集合の中で，不適格な弁別的文字列対の割合がある閾値[2]を上回っていれば c を候補から取り除く．この処理は未知語獲得の過程で動的に行う．重複する候補のどちらかが既知語であるかを調べるさい，辞書登録語だけでなく，これまでの処理で未知語であると判定された候補も既知語に含めて考える．この方法では，候補を分類器に適用する順番が最終的な精度に影響を与えるため，より短い文字列ほど単語になりやすいとう考えにもとづき，長さの短い候補から優先的に分類器に適用する．

[1] $\sqrt{|\mathcal{W}_t|}$ とした．

[2] 0.5 とした．

3.3 素性

未知語候補 c を分類器に適用するさいには，コーパスから c を抽出するときに使われた弁別的文字列を 2 値素性として使う．しかし，弁別的文字列の長さ n が大きい場合，素性ベクトルが疎になって事例間で素性が共有されにくくなるので，弁別的先行文字列の全ての接尾文字列も素性とする．例えば，弁別的先行文字列「たくさんの」からは接尾文字列「くさんの」「さんの」「んの」「の」も素性として利用する．同様に，弁別的後続文字列の全ての接頭文字列も素性とする．

3.4 訓練事例

分類器 C_t の訓練事例を作成するためには，未知語候補 c とそれに品詞 t を割り当て可能であるかどうかを示す正解タグの組が必要となる．正例の作成には，品詞 t が割り当てられている辞書登録語を利用することができる．負例の場合も同様に，t 以外の品詞が割り当てられている辞書登録語を使うことができる．しかし，これだけでは単語ではない文字列が負例に含まれないため不十分である．そこで，候補の絞り込みの場合と同様に，辞書登録語と重複して抽出された候補を負例に含める．

3.5 オンライン最大マージン学習

提案手法では，分類器の学習に任意の学習アルゴリズムを適用することができるが，ここでは高速なオンライン学習手法である Passive Aggressive (PA) アルゴリズムを用いた [1].

PA アルゴリズムは，素性ベクトル $\boldsymbol{x}$ と正解ラベル $y(=\pm 1)$ の組を観察するごとに，重みベクトル $\boldsymbol{w}$ を逐次更新する．重みベクトルは $\boldsymbol{w}_1 = (0, \ldots, 0)$ で初期化される．そして，i 番目の訓練事例 $(\boldsymbol{x}_i, y_i)$ が与えられるたびに，重みベクトル $\boldsymbol{w}_i$ を以下の最適化問題の解 $\boldsymbol{w}_{i+1}$ に変更する．

$$\boldsymbol{w}_{i+1} = \arg\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w} - \boldsymbol{w}_i||^2 + C\xi$$
$$\text{s.t. } l(\boldsymbol{w}; (\boldsymbol{x}_i, y_i)) \leq \xi \text{ and } \xi \geq 0$$

C はスラック変数 ξ の影響を調整するためのハイパーパラメータであり，$l(\boldsymbol{w}; (\boldsymbol{x}, y))$ は損失関数である．

$$l(\boldsymbol{w}; (\boldsymbol{x}, y)) = \begin{cases} 0 & y(\boldsymbol{w} \cdot \boldsymbol{x}) \geq 1 \\ 1 - y(\boldsymbol{w} \cdot \boldsymbol{x}) & \text{otherwise} \end{cases}$$

この制約付き最適化問題を解くと以下の更新式が得られる．

$$\boldsymbol{w}_{i+1} = \boldsymbol{w}_i + y_i \tau_i \boldsymbol{x}_i$$

ただし τ_i は以下の値である．

$$\tau_i = \min\{C, \frac{1 - \boldsymbol{w}_i \cdot \boldsymbol{x}_i}{||\boldsymbol{x}_i||^2}\}$$

3.6 句判定

文脈にもとづく未知語獲得の欠点の一つは，単語と句を明確に区別できないことである．例えば名詞句「ＸＯ醤の味」は，普通名詞と似た文脈に出現しやすいことが考えられるため，誤って普通名詞であると判断されてしまう可能性がある．

この問題を避けるために，全ての未知語候補の処理が終わったのちに，分類器が $+1$ を出力したものに対して，それが句であるかどうかの判定を行う．現在のところ，この処理は単純な辞書引きを行っている．分類器によって未知語と判断されたものを全て辞書に追加し，その拡張された辞書を使って，未知語候補を複数語に分割可能であるかどうかを調べる．ただし，過分割されるのを防ぐために，京大コーパスから学習した品詞 2-gram を制約として使う．

4 実験

4.1 設定

実験には 1.7 億文のウェブテキストと JUMAN 辞書を用いた．JUMAN 辞書は，表 1 に記載された品詞に分類されている単語だけを取り出して，訓練用と評価用に分割した．訓練用辞書は，弁別的文字列の獲得と訓練事例の生成のために使い，評価用辞書は分類器の適合率と再現率を求めるために使った．

未知語候補の生成では弁別的文字列の長さを $n = 5$ とした．実験で獲得した先行文字列と後続文字列は 299,576 と 153,583 であった．表 2 に具体例を示す．素性数に対する閾値 σ は $1, 16, 32, 64$ を試した．PA アルゴリズムのハイパーパラメータ C は 1.0 とした．また PA アルゴリズムではカーネルトリックを使うことができるので 2 次の多項式カーネルを用いた．

品詞	先行文字列/後続文字列 (上段/下段)
普通名詞	またとない, など 重要な, どあらゆる があります, はもちろん, に対しても
子音動詞ラ行	あっけなく, いい感じに, から何かが るでしょう, ろうとした, りましょう
イ形容詞	ヒジョーに, たまらなく, よりもっと かったので, いのですが, くなったの

表 2: 弁別的文字列の例

普通名詞	兄ィ，腐女子，音樂，特盛
サ変名詞	逆ギレ，怪演，マターリ
子音動詞ラ行	ポシャる，帰える，ハショる
イ形容詞	甘っちょろい，ヤヴァい，ムズ痒い
ナ形容詞	ぴっかぴかだ，マッドだ，みょーだ

表 3: 獲得した未知語

4.2 実験結果

分類器の性能を調べるため，評価用辞書を用いて適合率と再現率を求めた．評価用辞書に登録されている単語には，そもそも未知語候補 (3.1 節) に含まれていないものもある．そのため，分類器が全ての候補を正しく分類できた場合の再現率も調べた (表 4)．次に，コーパスでの出現頻度と再現率の関係を調査した．コーパスを Juman で解析して評価用辞書に含まれる単語の頻度を求め，一定以上の頻度を持つ単語のみを対象として再現率を求めた (表 5)．全体の再現率は 77.1%であるが，獲得漏れの原因の一つは低頻度の単語であり，ある程度高い頻度のものに限定して見ると，より高い再現率を達成できていることが分かる．コーパスからは全部で 12,823 の未知語を獲得することができた ($\sigma = 64$)．ランダムに 100 個を選んで調べたところ 81 個が適切なものであった．表 3 に獲得した未知語の例を示す．

5 おわりに

本論文では識別モデルを用いて未知語獲得を行うため手法について述べた．本研究の成果は言語学研究の支援技術として活用していく予定である [5]．今後は，手法の改良を進めるとともに，固有名詞や副詞など，今回の実験では対象としなかった品詞にも同様の手法が適用可能であるかを調査する．

σ	適合率	再現率	再現率の上限
1	80.2	77.1	89.0
16	86.7	75.5	84.9
32	89.1	73.2	80.4
64	90.2	69.9	75.5

表 4: 適合率と再現率

頻度	$1 \leq$	$10 \leq$	$20 \leq$	$50 \leq$	$100 \leq$	$200 \leq$
再現率	77.1	82.7	85.0	87.5	89.0	90.4

表 5: 出現頻度と再現率の関係 ($\sigma = 1$)

参考文献

[1] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shawartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, Vol. 7, pp. 551–583, 2006.

[2] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of EMNLP*, pp. 230–237, 2004.

[3] Shinsuke Mori and Makoto Nagao. Word extraction from corpora and its part-of-speech estimation using distributional analysis. In *Proceedings of COLING*, pp. 1119–1122, 1996.

[4] Yugo Murawaki and Sadao Kurohashi. Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proceedings of EMNLP*, pp. 429–437, 2008.

[5] 宇野良子, 鍛治伸裕, 喜連川優. 新動詞の認知言語学的分析：大規模時系列ウェブコーパスと言語処理技術が可能にする言語のダイナミズム研究. 言語処理学会第１５回年次大会, 2009.

[6] 桑江常則, 佐藤理史, 藤田篤. 後続ひらがな列に基づく語の活用型推定. 情報処理学会研究報告 NL-186-2, pp. 7–12, 2008.

[7] 福島健一, 鍛治伸裕, 喜連川優. 機械学習を用いたカタカナ用言の獲得. 言語処理学会第１３回年次大会, pp. 815–818, 2006.