1C5-04

Support Vector Machineを用いた未知語の抽出

Extraction of Unknown Words with Support Vector Machine

倉田 岳人*1　　石塚 満*2
Gakuto KURATA　　Mitsuru ISHIZUKA

*1*2東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, University of Tokyo

More and more natural language resources are available these days. Because Japanese is characterized by no space between words, any application with natural language resources such as text summarization , information retrieval , question answering and so on follows after Morphological Analysis. Morphological Analysis tools we use today are essentially not good at extracting unknown words exactly. However, unknown words which represent new concepts, new people, new objects and so on are most important in the documents. So, failure in anlyzing unknown words is critical to any application.

This paper provides a new technique processing unknown words correctly with Support Vector Machine. Experiments show this method is superior to the approach with Decision Tree and AdaBoost.

1. はじめに

近年多くの情報が電子化され，我々はそれらを利用することができる．現在電子化された言語資源を利用して行われている処理としてはキーワード抽出，自動要約，情報検索，質問応答など多くが挙げられる．英語などの言語の場合は単語と単語の間に空白を入れるため，これらの処理を行う際に直接的に言語資源を用いることができる．それに対して日本語では単語境界を明示しないため，言語資源を用いるにあたり形態素解析が行われる．現在行われている形態素解析は辞書とルールに基づく処理を基本としている．この手法の問題点として未知語，つまり形態素辞書に記載されていない新しい単語に対する処理に脆弱である，ということが挙げられる．しかし，未知語は新しい概念や人物，組織などを現すものであり，文書中でもっとも重要な単語である場合が多い．このような観点から，日本語テキスト中から未知語を正しく切り出し，解析することは非常に重要であると考えられる．また，未知語を含む日本語テキストに対する形態素解析の精度を向上させることができれば，情報検索，質問応答などの様々な処理の精度向上に貢献できると考えられる．本報告では，機械学習の手法の一つである Support Vector Machine(SVM) を用いて高精度に未知語を抽出する手法を提案する．

2. 未知語の検出方法

人間が未知語をどのようにして同定しているか，ということを考えた場合，次の二つの観点から未知語を認識していると考えられる [1].

1. 未知語の表記が単語として尤もらしいか
2. 未知語が文章中で文法的に可能かどうか

この二つの観点を，計算機上で実装することを考えてみる．1 に示した観点は，例えば特定のドメインのコーパスからドメイン固有の文字列に関する特徴を学習し，ドメイン固有の複合語を検出する，という手法として実現される [2]．しかし，この手法は複合語を構成する文字列そのものに関する情報を用いており，辞書を用いる形態素解析と同様に，コーパス中に出現しない複合語の処理に関して脆弱であるという問題点を持つ．つまり，このアプローチにより未知語を検出しようとした場合，本質的に限界が生じる．

それに対して 2 の観点は，未知語そのものに関する情報を全く用いず，前後の環境から未知語を同定するというアプローチであり，1 の場合の様な限界はない．

以上の議論から，未知語を計算機上で扱う場合，2 の観点に従った手法を用いるべきであり，本報告でも 2 アプローチを採用する．

連絡先: 倉田 岳人,
東京大学大学院情報理工学系研究科,
113-0033 文京区本郷 5-25-16 石川ビル 11F,
03-5802-8213,
kurata@miv.t.u-tokyo.ac.jp

3. 分類問題への帰着

形態素解析は日本語単語分割と，各々の形態素に対する品詞の付与により実現される．ここで，日本語単語分割を分類問題に帰着する手法が提案されている [3]．提案されている手法の概要を図 1 に示す．

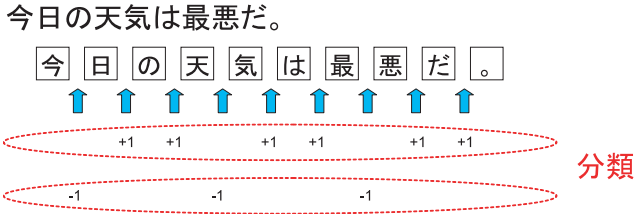

図 1: 日本語単語分割の分類問題への帰着

図 1 に示した様に，文字と文字の間に単語境界がある (+1) かないか (-1) の分類問題に帰着することで，様々な分類問題の手法を用いることができる．なお，提案されている手法では，単語境界があるかないかを判定するにあたり，判定する位置の前後の文字列，および前後の文字の文字種を用いている．

4. 提案手法

上述した日本語単語分割を分類問題に帰着するという観点から考えると，未知語を検出するということは，文字列中の文字と文字の間が未知語の開始時点であるかないか，もしくは終端地点であるかないか，ということを分類することにより実現することができる．提案手法の概要を図 2 に示す．

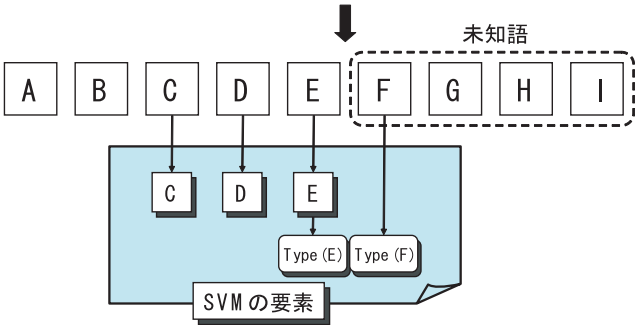

図 2: 提案手法の概要

提案手法では，未知語の開始時点と終端時点を別々に検出する．図 2 では開始時点を検出する場合の概要を示した．

4.1 Support Vector Machine

提案手法では，分類問題に帰着した後の分類手法として Support Vector Machine(SVM) を用いる．SVM は統計的学習理論に基づく 2 クラスのパターン認識手法であり，少ない学習量で高い認識精度が得られる方法である．しかし，学習データ量が増えるとモデルの学習にかかる時間が膨大になる，という問題点もある．そのため適切な要素を選択することが重要となる．

4.2 SVM の要素

図 2 に示したように，提案手法では未知語の開始位置を検出するために，未知語の前の三文字 (C,D,E)，および未知語の直前の文字の文字種 (type(E))，未知語の最初の一文字の文字種 (type(F)) を要素とした．同様に未知語の終端位置を検出する場合，未知語の直後の文字三文字，および未知語の最後の一文字と直後の一文字の文字種を要素とした．要素をこのようにとることにより，文字種の変化を SVM の要素に取り入れることができるとともに，未知語そのものの情報を用いることなく，未知語の検出を行うことができる．なお，本研究で用いた文字種は以下の表 1 の通りである．

表 1: 文字種

文字種	
記号 1	句読点,「」等
記号 2	その他の記号
平仮名	あいうえお
カタカナ	アイウエオ
アルファベット	ABCDE
漢数字	一二三四五
漢字	漢字

4.3 処理の流れ

今回行った処理の流れを以下の図 3 に示す．なお，今回は未知語として未知の名詞を対象とした．

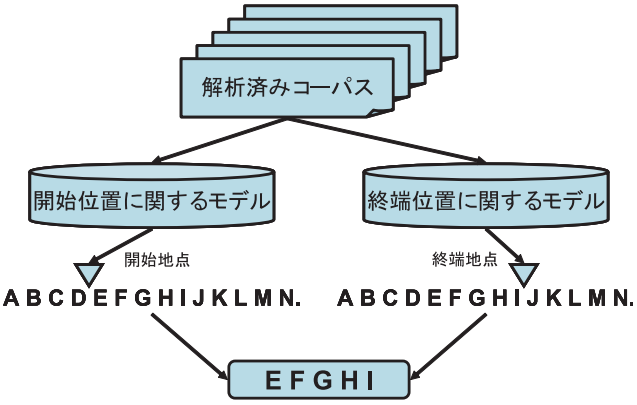

図 3: 処理の流れ

1. 解析済の新聞データから，未知語の開始時点に関するモデル，および未知語の終端に関するモデルを構築する．
2. 解析されていない文章のすべての文字と文字の間に関して，未知語の開始位置であるかないか，および終端位置であるかないかを判定する．
3. 開始位置，および終端位置の情報から未知語を検出する．

5. 評価実験

5.1 評価実験の条件

評価実験の実験条件を表 2 に示す．

表 2: 実験条件

学習データ	
京都大学コーパス	1 日分
文書数	1129
データ数	42631
テストデータ	
京都大学コーパス	1 日分
文書数	687
データ数	27146

5.2 評価実験の結果

実験結果として，開始位置，終端位置の検出精度，および開始，終端とも正しく検出され正確に名詞が抽出された精度を次頁の表 3 に示す．

表 3: 抽出の精度

	Precision(%)	Recall(%)
開始位置	84.2	82.8
終端位置	91.5	91.1
抽出精度	**74.1**	**78.3**

6. 抽出できた文字列に関する考察

6.1 文字種の変化

日本語特有の単語分割手法として，文字種区切りを利用する手法がある．文字種の変化を利用することは非常に有用であり，本手法でも SVM の要素として取り入れている．しかし，様々な文字種から構成される単語も存在する．このような単語に関しては，文字種区切りのみに着目した場合，検出することができない．提案手法により複数の文字種を含むが正しく検出できた例を表 4 に示す．

表 4: 抽出できた例

２５トン大型トラック	2＋5＋トン+大型+トラック
AFP 通信	AFP + 通信
ホテル日航大阪	ホテル + 日航 + 大阪
アントシアン系色素	アン+トシ+アン+系+色素

以上より，提案手法は文字種の変化を要素として用いているが，未知語の前後の文字そのものも要素とすることにより，単純な文字種区切りによる手法よりも頑健であるということができる．

たとえば,「アントシアン系色素」を ChaSen で解析した場合,「アン」と「トシ」が人名として ChaSen の辞書に登録されているため，過分割が生じていると考えられる．しかし，人間が「アントシアン系色素」を過分割することなく，人間の同定方法に従った提案手法の有効性が現れていると考えられる．

6.2 複合語

複合語の抽出というタスクを考えた場合，まず「どの単位を以て複合語とみなすか」という問題がある．この定義に関しては様々な議論があり，明確に定義し，それに基づき正解集合を定義することは困難であると考えられる．しかし，質問応答，係り受け関係の解析等を考えた場合，複合語の抽出ということも重要となる．以下の表 5 に提案手法を用いて抽出できた複合語の例を示す．

表 5: 複合語の例

医薬品販売業	医薬品 + 販売 + 業
護送船団方式	護送 + 船団 + 方式
国連平和維持軍	国連 + 平和 + 維持 + 軍
国連安全保障理事会	国連 + 安全 + 保障 + 理事 + 会

提案手法は名詞の前後の環境を用いているため，本質的に複合語の抽出に向いている手法であると言うことができる．具体的には，学習コーパス中に,「国連」という単語があり,「国」と「連」の間を開始位置，もしくは終端位置と学習している場合は非常に少ないと考えられるからである．しかし，このアプローチは辞書を用いたアプローチと同じ問題を持つ．よって本報告では複合語に関する議論については深く行わないこととする．

7. まとめ

関連研究として決定リストと AdaBoost を用いて日本語単語分割を行っている研究事例が挙げられる [3]．この手法を用いると非常に高い精度で日本語単語分割が実現されているが，未知語の検出率は 6 割程度にとどまっている．この原因としては，未知語そのものの情報を決定リストの学習時に利用していることが挙げられる．この手法では前述した辞書に基づく処理と同じ問題点があり自ずと限界が生じる．

これに対して本報告での未知語の検出精度は，5.2 で示した通り関連研究の精度を大きく上回っており，提案手法の有効性が示されたと考えられる．この要因としては，開始位置と終端位置を別々のモデルで検出することにより，未知語そのものの情報を用いずに抽出を行っていることが挙げられる．

今後の課題としては以下の様なことが挙げられる．

適切な要素の選択 提案手法では SVM の要素数は２万５千程度となった．そのため，モデルの作成の時間を考慮すると，数日分の新聞データからしか学習することができなかった．さらに適切な要素を考案することにより，抽出精度のさらなる向上，および，処理の短時間化をはかりたいと考えている．

形態素解析ツールとの融合 今回提案した手法を，現在使われている形態素解析ツールによる手法の前処理として用い，形態素解析の精度向上，およびその他の処理の精度向上につなげたいと考えている．

マージンを考慮した学習手法の提案 SVM の学習ではソフトマージンという手法が用いられている．ここに boosting の様な考え方を導入することにより，より性能の高い学習手法を提案したいと考えている．

参考文献

[1] 内元清貴, 関根聡, 井佐原均. “最大エントロピーモデルに基づく形態素解析–未知語の問題の解決策–”. 自然言語処理, Vol. 8, No. 1, pp. 127–141, Jan. 2001.

[2] Nobesawa, S., et al. “Segmenting Sentences into Linky Strings Using D-bigram Statistics”. *Proceedings of Coling-96*, No. 586-591, 1996.

[3] 新納浩幸. “決定リストを弱学習器としたアダブーストによる日本語単語分割”. 自然言語処理, Vol. 8, No. 2, pp. 2–17, 2001.