

SVMによる学習とタイプ分類パターンの組み合わせによる固有表現抽出手法

尾田俊弘[†] 福本 淳一[‡]

立命館大学大学院 理工学研究科[†] 立命館大学情報理工学 メディア情報学科[‡]

1 はじめに

固有表現抽出とは、人名、地名、組織名、日時など情報の単位として一定のまとまりのある重要な表現を文書中から抜き出す技術であり、情報抽出システムや質問応答システムなどの重要な基礎技術である。また、様々なタイプの質問に回答することが求められる質問応答システムには、例えば、人名のタイプから派生した、政治に関する人名、会社に関する人名、といった詳細なタイプ分類が正確な回答を得るために必要である。固有表現抽出手法としては、人手で作成した抽出規則を対象文書に適用するパターンを基にした手法と、あらかじめ固有表現抽出を行っておいたテキストから機械学習を用いて抽出規則を自動的に学習する手法がある。そして、機械学習を用いた固有表現抽出手法には、Support Vector Machine(SVM)を用いた手法が提案されており、日本語を対象とした手法としては、山田らといった先行研究があり高い精度を挙げている。[3]

しかし、機械学習のみを用いた固有表現抽出手法には、固有表現タイプごとに学習と固有表現抽出を行う必要があること、および、構文情報を用いたタイプ分類など、学習のみではタイプ分類の難しい固有表現が存在することが問題として挙げられる。我々は、機械学習を用いて固有表現範囲の同定を行い、タイプ分類と固有表現抽出のための後処理に人手でパターンを書く手法を用いることで固有表現の詳細なタイプ分類を実現する。

2 提案手法

2.1 固有表現範囲の同定

固有表現抽出の対象範囲を決定するために、まず、学習による固有表現範囲の同定を行う。学習には、対象データを正例、負例に二値分類するSVMを利用するが、本手法では、固有表現範囲に含まれる場合を正例、含まれない場合を負例として分類する。

Named Entity Extraction Method Combined with Machine-Learning Using Support Vector Machine and Pattern for Type Classification

Toshihiro ODA[†] and Junichi FUKUMOTO[‡]

Graduate school of Science and Engineering, Ritsumeikan University[†] and Department of Media Technologies, Ritsumeikan University[‡]

2.1.1 固有表現範囲規則の学習

SVMを用いて固有表現範囲を同定するために、固有表現範囲規則を学習する。学習では、固有表現範囲の境界を単語単位とし、学習データ中出现する人名、地名、組織名、国名、言語名、製品名、施設名、作品名、法律名、イベント名とタグ付けされた固有表現を正例として、その他の固有表現、タグ付けされていない単語を負例として学習した。規則の学習には、文頭から順番に固有表現範囲を同定する右向き解析の手法を取り、単語ごとの素性を調べ学習に用いる。例えば、参照中の単語に関する素性には、2つ前の単語から2つ後の単語までの計5単語の単語の字面、品詞、文字種を用いる。そして、素性は、単語の字面を計41366種類、文字種を計16種類、品詞を計38種類に分類し出現する位置も考慮に入れ、SVMへの入力のため、0と1の2値で表したバイナリベクトルに変換する。バイナリベクトルの例を表1にて示す。また、一度解析した単語の素性は処理時間の短縮のために、次の解析にてそのまま用いる。

¹ 学習時のパラメータは [2], [3] の評価実験にて、最も高い精度が得られた数値 $d = 2$, $C = 0.1$ に設定した。

表 1: 学習データのバイナリベクトルへの変換

2つ前の単語が「points」	...	1
2つ前の単語が「,」	...	0
2つ前の単語が「Mr.」	...	0
...		
2つ前の品詞が「名詞複数形」	...	1
2つ前の品詞が「コンマ」	...	0
2つ前の品詞が「固有名詞」	...	0
...		
2つ前の文字種が「アルファベット小文字」	...	1
2つ前の文字種が「コンマ」	...	0
2つ前の文字種が「キャピタライズ」	...	0
...		

¹ パラメータ d は、単語自身、品詞など任意の素性の組み合わせの数であり、パラメータ C は正則化の度合を制御するための数値である。

表 2: 固有表現のタイプ分類パターンの例

タイプ	パターン
人名	〈固有表現範囲〉 + ‘前置詞 of’ or ‘前置詞 at’ + 組織名
地名	‘前置詞 in’ or ‘前置詞 near’ + 〈固有表現範囲〉
組織名	‘役職名の名詞’ + ‘前置詞 of’ or ‘前置詞 at’ or ‘前置詞 for’ or ‘前置詞 in’ + 〈固有表現範囲〉
名詞に準じたタイプ	‘単語リスト中の名詞’ + ‘such’ + ‘as’ + 〈固有表現範囲〉
名詞に準じたタイプ	〈固有表現範囲〉 (同格) → 単語リスト中の名詞
名詞に準じたタイプ	‘単語リスト中の名詞’ + 〈固有表現範囲〉
...	...

※矢印は係る先, +は隣接, () で囲まれた部分は係り先との関係, {} で囲まれる表現は固有表現範囲でありタイプの固有表現として抽出される, ‘ ’ 内は文字列

2.2 人手でパターンを作成する手法に基づく処理

従来のSVMを用いた固有表現抽出手法では, タイプ分類もSVMによる処理に含まれるのだが, 単語自身や品詞, 文字種など表層的な情報のみを素性として使用しているため, 場合や単語の係り先など文脈の情報により判断できるタイプ分類は困難である. また, 固有表現範囲の同定にて正しく範囲として得られなかった部分が存在した. そのため, 人手でパターンを書く手法にてタイプ分類と抽出のための後処理を行う. タイプ分類の際には, 人名, 地名, 組織名といった上位タイプから派生したサブタイプも含め計99種類のタイプに分類する.

2.2.1 固有表現のタイプ分類

タイプ分類は, 人名, 地名, 組織名などのタイプから派生したサブタイプも含め計99種類のタイプへの分類を行う. 例えば, 固有表現範囲に存在する接尾語, 接頭語といった接辞情報によって判断する場合, ‘Co.’ は会社に関する組織名の分類に, ‘Dr.’ は役職に関する人名の分類に用いられる. 以上の接辞情報によるタイプ分類を基本的な処理とし, 単語自身や単語間の前後関係, 品詞解析結果, 構文解析結果を基に書かれた計79種類のパターンとタイプごとに登録された単語リストをタイプ分類に用いる. タイプ分類パターンを例を表2に示す.

2.2.2 固有表現範囲のための後処理

後処理では, 固有表現範囲の同定では対象としなかった数量表現の抽出を対象にする. また, SVMを用いて得られた固有表現範囲には誤りや見落としが存在したことより範囲の修正を行うパターンを記述し, より正確に固有表現を抽出することを目的とする. 範囲の修正を行うパターンとしては, 「組織名のタイプ分類にて, ‘序数’, ‘前置詞 of’, ‘前置詞 for’, ‘前置詞 in’, ‘所有格’s’, ‘等位接続詞 and’, ‘記号-’ が固有表現範囲の間に現れた場合, 固有表現範囲に含む」といったパターンがある.

3 実験・考察

実験では, 学習に Tiny-SVM0.09 を用い, 学習データには, Linguistic Data Consortium の BBN Pronoun Coreference and Entity Type Corpus を使用し

た. BBN Pronoun Coreference and Entity Type Corpus は, Wall Street Journal の 1989 年度の記事 2282 件に対し, Penn Treebank Project で定義された品詞と構文情報が付与され, 固有表現にはタグ付けがされている. その中から, 学習データに用いていないデータでテストを行ったところ, 組織名, 人工物名, その他の固有表現の抽出にて失敗が多く見られた. これは, タイプ分類のパターンに無い固有表現が存在したこと, 人名, 地名といった他の固有表現の特徴と異なる特徴を持つことが多かったため, 固有表現範囲の同定にて良い結果が得られなかったこと, が主な理由として挙げられる.

4 おわりに

本稿では, 英語テキストを対象に, 学習と人手でパターンを作成する手法とを組み合わせた固有表現抽出手法を提案した. 本手法により, タイプごとに行う必要のあるSVMの処理を一度に行うことができ, さらに, 機械学習を用いた固有表現抽出のみでは難しかった固有表現の詳細なタイプ分類を実現することができた. 評価では, 詳細なタイプ分類の評価を人手での判断で行ない, 適合率と再現率で精度を測り, 評価する予定である.

参考文献

- [1] 福本淳一, 榊井文人, 鈴木伸哉: 固有表現抽出ツール NExT の精緻化とユーザビリティの向上, 第8回言語処理学会年次大会発表論文集, pp.176-179., 2002.
- [2] Isozaki, Kazawa: Efficient Support Vector Classifiers for Named Entity Recognition, COLING '02 Proceedings of the 19th international conference on Computational linguistics, Vol.1, 2002.
- [3] 山田, 工藤, 松本: Support Vector Machine を用いた日本語固有表現抽出情報処理学会論文誌, Vol.43, No.1, pp.44-53, 2002.