ドメイン依存知識を動的に切り替える固有表現抽出手法について

鈴木　伸哉†　　桝井　文人‡　　河合　敦夫‡　　椎野　努‡
†三重大学大学院工学研究科　　‡三重大学工学部

1　はじめに

近年，文書中の主要な情報を取り出すことを目的とした，情報抽出技術が注目されている [1]．文書中に含まれる，組織名・地名・人名などの固有名詞や，日付・時刻・金額などの数値表現を抽出する固有表現抽出技術は，情報抽出において基本的かつ主要な技術である．これまでにも，人手による抽出規則を用いる手法や，機械学習によって抽出規則を獲得する手法が提案されている [2]．

現在のところ，固有表現抽出技術では特定のドメインに対してチューニングを行えば，高い抽出精度を得られることが確認されている．しかしながら，以前より，ドメインが異なれば言語現象も異なることが指摘されており [3]，特定のドメインから獲得した抽出知識が他のドメインでコンフリクトを起こし，抽出精度の低下を招くことが考えられる．

そこで本論文では，ドメイン固有の知識を動的に切り替えることによって，抽出精度を低下させずに，複数のドメインに適用できる固有表現抽出手法を提案する．固有表現抽出に用いられる知識のうち，固有表現の末尾に用いられる接尾辞や接辞（以下まとめて接辞と呼ぶ）に注目する．まず，あらかじめ分類されたドメインに属する文書集合からドメイン固有の接辞を獲得しておき，接辞辞書に登録する．対象文書のドメインに応じて，動的にドメイン固有接辞辞書を切り替えることでドメイン固有接辞によるコンフリクトを抑制し，抽出精度の低下を防ぐことができる．

以下，２章ではドメイン依存知識について詳述し，３章で接辞辞書の構築について述べる．４章で処理の概要を述べた後，５章で実験及び評価を行い，６章で考察を行う．最後に７章で今後の展望を述べる．

2　ドメインに依存する知識

固有表現抽出でよく用いられる手法に，固有表現に付属する接辞の照合がある．この手法では，接辞辞書と照合した語を固有表現の一部とみなし，意味の切れ目を特定することで固有表現抽出を行う．例えば，接辞「大学」を用いることで「三重大学」という組織名が抽出できる．

接辞には，あるドメインにしか使われない，ドメイン固有のものがある．これらを他のドメインに適用すると，ノイズとなり，精度低下の原因となる場合が多い．例えば，スポーツのドメインに属する記事中において「花籠部屋」は組織名であり，「部屋」が接辞である．しかし，他のドメインに属する記事にこの接辞を適用すると普通名詞の「勉強部屋」や「子供部屋」などが過抽出されてしまう．

以上の問題は，記事が属するドメインによって，接辞辞書を切り替えることで対応でき，その結果，精度低下を抑制できると考えられる．即ち，ドメイン固有の接辞が原因となって起こった過抽出が減少すれば，抽出システム自体の精度は向上するはずである．先の例を用いると，スポーツのドメインと判定された記事に対しては接辞「部屋」が用いられ，他のドメインの記事では適用しないため「勉強部屋」などの過抽出を防ぐことができる．

さらに，特定のドメインにしか機能しないために，接辞辞書に登録できなかった接辞をドメインに限定して用いることができるようになり，より多くの接辞を抽出処理に利用することができる．その結果，抽出できる固有表現数が増加する．例えば，スポーツのドメインに有効な接辞「部屋」を全てのドメインの記事に対して適用すると全体の精度が低下する場合を考える．接辞辞書の切り替えを行わない抽出手

法では精度の低下を防ごうとすると辞書に登録できないが，辞書を切り替える手法を用いれば精度を低下させることなくドメイン固有の接辞辞書に登録できる．

3　接辞辞書の構築

本章では，固有表現に付属する接辞辞書を構築する手法について述べる．概要を図１に示す．

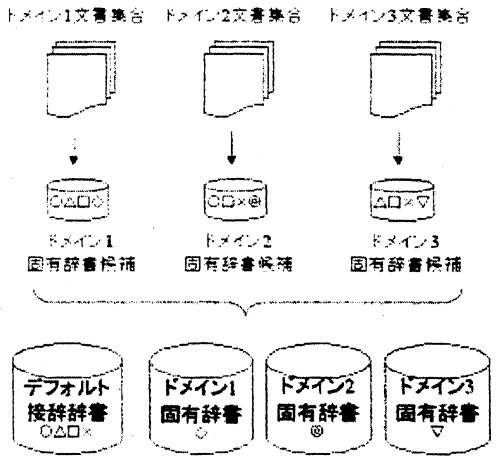

図 1: 接辞辞書の構築

まず，あらかじめ人手により固有表現に対しタグ付けされた新聞記事をドメイン毎に分類し，ドメイン文書集合を作る．次に，文書集合に含まれる固有表現を取り出し，形態素解析を行い，末尾の形態素をドメイン固有接辞の候補として得る．例えば，「津市」という地名の形態素解析結果は「津／市」であり，末尾の「市」が接辞候補として得られる．ただし，「アメリカ」のような単独の形態素から成る固有表現は，接辞が付属しないものとし，除外する．また，数字から構成される接辞候補が得られた場合についても，数値表現とコンフリクトを起こしやすいため，対象から除外しておく．例えば，「松山町１９」からは番地を示す「１９」が接辞候補となるが，「学生１９人」という表現に対して適用すると，人数を抽出対象にしていない場合，「学生１９」が誤って抽出される．

3.1　ドメイン固有接辞候補の分類

前節の手法で得られたドメイン固有接辞の候補には，ドメイン固有でない接辞も含まれる（以下では，ドメイン固有でない接辞をデフォルト接辞と呼ぶ）．ドメイン固有接辞候補を分類し，デフォルト接辞辞書とドメイン固有接辞辞書に登録する手順を以下に示す．

1. 複数のドメインに出現するドメイン固有接辞候補は，デフォルト接辞辞書に登録する
2. 上記以外を，出現したドメインの固有接辞辞書に登録する

4　処理の概要

本章では，ドメイン固有接辞を動的に切り替える，固有表現抽出処理について説明する．処理の流れは図２のようになる．以下，各処理について述べる．

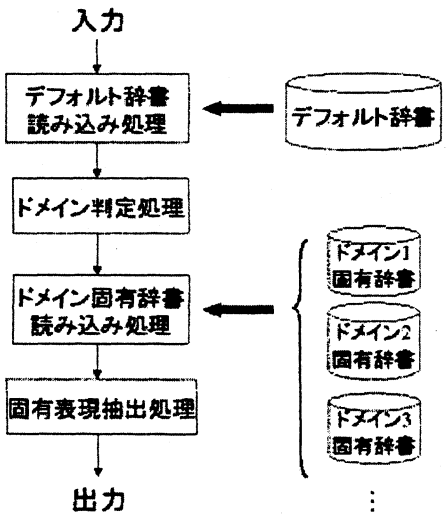

図 2: 処理の流れ

- デフォルト辞書読み込み処理
  デフォルト接辞辞書を読み込む．デフォルト接辞辞書は全てのドメイン共通に用いる．

表 1: 構築した辞書のみを用いた結果

	再現率	適合率	F 値
A	67.82	66.63	67.22
B	68.63	71.84	70.20
C	70.19	71.85	71.01

A： 4 ドメインから得られた接辞を全て使用

B： ドメイン毎に得られた接辞辞書を切り替えて使用

C： B から構築したデフォルト接辞辞書と，切り替え用のドメイン固有接辞辞書を使用

表 2: NExT 付属辞書に適用した結果

	再現率	適合率	F 値
D	69.97	71.63	70.79
E	70.46	68.25	69.34
F	71.81	71.54	71.67

D： NExT 付属の接辞辞書を使用

E： NExT 付属の辞書に，今回獲得したドメイン固有接辞辞書を加えて使用

F： NExT 付属の辞書に，今回獲得したドメイン固有接辞辞書を切り替えて使用

- ドメイン判定処理
  文書に対して形態素解析を行い，ドメインの判定を行う．ドメインの判定には，tf·idf 値を用いた (式 1).

$$tf \cdot idf = tf \times \log \frac{N}{df} \qquad (1)$$

  N は文書総数であり，df は名詞の出現文書数である．tf は各ドメイン文書集合における名詞の頻度とし，各ドメイン毎に名詞の tf·idf 値を得た．文書から名詞を取り出し，ドメイン毎に tf·idf 値の和を算出し，各ドメインのスコアとする．最も高いスコアを得たドメインを，文書の属するドメインと判定する．

- ドメイン固有辞書読み込み処理
  判定された文書のドメインに従い，ドメイン固有の接辞辞書を選択し，読み込む．

- 固有表現抽出処理
  各形態素を接辞辞書と照合し，接辞であれば固有名の範囲を特定する．範囲が定まると，接辞に対応した固有表現のラベルを与える．

5 実験と評価

5.1 実験データと評価手法

前章で述べた固有表現抽出手法の有効性を検証するために，以下の実験を行った．訓練コーパスとして，CRL 固有表現データを用いた [4]．まず，tf·idf 値の算出とドメイン固有接辞の獲得に用いるため，毎日新聞記事の CD-ROM データに収録されている掲載面種別コードに従って，国際，経済，スポーツ，社会の 4 ドメインについて 50 記事，計 200 記事を選出した．また，これとは別に，テストコーパスとして 4 つのドメイン各 30 記事，計 120 記事を選出した．次に 200 記事から訓練コーパスを用いてドメイン毎に固有表現のリストを作成し，固有表現に付属する接辞を獲得した．

3 章で述べた手法で接辞を分類し，デフォルト接辞辞書と各ドメイン固有接辞辞書を得た．得られた接辞辞書を用いて，テストコーパスに対して固有表現抽出処理を行い，評価を行った．固有表現抽出には，汎用のツールとして開発・公開されている NExT[5] を用い，接辞辞書のみを今回獲得したものと置き換えた[1].

また，NExT 付属の辞書を用いた実験を行った．NExT 付属の接辞辞書は人手によって獲得されたもので，約 400 の接辞が収録されている．この接辞辞書をデフォルト接辞辞書とし，先に獲得したドメイン固有接辞辞書を適用した[2].

IREX[4] を評価基準とし，評価尺度には再現率，適合率，F 値を用いた．算出式を以下に示す．

$$Recall = \frac{\text{システムが抽出した正解数}}{\text{正解数}}$$

$$Precision = \frac{\text{システムが抽出した正解数}}{\text{システムの抽出数}}$$

[1] NExT は人工物名の抽出に対応していないため，これを評価対象外とした．

[2] 数値表現については，NExT 付属の接辞辞書で十分にカバーできたため，除外した．

$$F = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

5.2 評価結果

今回構築した接辞辞書を用いた，テストデータに対する評価結果を表1に示す．また，NExT付属の接辞辞書を用いた結果を表2に示す．

6 考察

表1において，AとBの結果を比較すると，Bにおいて適合率の向上度が高く，ドメインを判定して接辞辞書を切り替えることによって，過抽出を抑制できたことがわかる．例えば，経済面の記事で，普通名詞である「デジタル通信」が誤って組織名として抽出されていたが，接辞「通信」を国際のドメイン固有辞書に限定して登録することで抽出を抑えることができた．また，BとCを比較すると，再現率の向上度が高く，デフォルト接辞によって，より多くの固有表現が抽出されたことがわかる．例えば，経済のドメイン文書集合とスポーツのドメイン文書集合で得られた「グループ」という接辞が，社会のドメインに属する記事に対しても有効に働き，「水俣病研究グループ」が組織名として正しく抽出できた．

一方，表2の結果からは，NExT付属の辞書に対しても，今回獲得したドメイン固有辞書を切り替えることで有効に作用することがわかる．Dに対してEでは精度が低下したが，切り替えを行なったFでは精度を向上させることができた．例えば，Eでは社会ドメイン固有の接辞として得られた「目」(適用例：5丁目)がスポーツドメインの記事に対して適用されたことにより「W杯2勝目」が誤って抽出されたが，Fでは辞書の切り替えを行なったため抑止することができた．その結果，FにおいてDの精度を上回ることができた．

また，表1と表2を比較すると，接辞切り替えを行なわなかったAではDに対しF値で約3.5の差があったが，切り替えを行なったCではDを上回り，Fに対しても約0.6まで差を縮めることができた．以上のことからも，接辞辞書の切り替えの有効性が確認できた．

7 おわりに

本論文では，ドメイン固有の知識を動的に切り替えることによって，複数のドメインに適用できる固有表現抽出手法を提案した．ドメイン文書集合毎に接辞を獲得する手法について述べ，ドメイン固有接辞とデフォルト接辞を分類し，接辞辞書を構築する手法ついても述べた．提案手法を検証するため，新聞記事に対して実験を行った．評価を行った結果，接辞辞書を切り替えない場合よりも切り替えた場合の方が高精度が得られ，接辞辞書をドメインによって切り替えることの有効性を示すことができた．また，既存のシステムで用いられていた接辞辞書についても同様に実験を行い，精度を向上させることができた．

今回は，ドメイン判定にtf-idf値を用いたが，SVMやブースティングなどの学習アルゴリズム[6]を導入することも可能である．また，スポーツに対して野球，サッカーなどのジャンルを用いて，ドメインを階層化することでさらに本手法を精緻化することも考えている．

参考文献

[1] 福本淳一, 関根聡, 江里口善生: MUC-7, Tipster参加報告, 1998

[2] 関根聡，江里口善生：IREX-NEの結果と分析，言語処理学会第6回年次大会ワークショップ論文集, pp.25-32, 2000

[3] 関根聡：コーパスからの自動学習と人手での規則作成を融合させた形の英語品詞タガー，言語処理学会 第7回年次大会 発表論文集, pp.14-17, 2001

[4] IREX実行委員会編：IREXワークショップ予稿集, 1999

[5] 桝井文人, 鈴木伸哉, 福本淳一：テキスト処理のための固有表現抽出ツールNExTの開発，言語処理学会, 2002

[6] 永田昌明, 平博順：テキスト分類 ー学習理論の「見本市」ー，情報処理 Vol.42 No.1, pp.32-37, 2001