辞書に基づく組織名抽出における辞書整備の影響

髙橋 寛治　奥田 裕樹
Sansan 株式会社 DSOC
{ka.takahashi, okuda}@sansan.com

1　はじめに

固有表現抽出 (Named Entity Recognition; NER) とは，文中に含まれる固有表現 (Named Entities; NEs) を抽出する技術である．テキスト中の組織名や人名，地名などを固有表現と呼ぶ．

任意の固有表現抽出に取り組む際に，辞書と対象のテキストは所持するがアノテーション済みデータが無い場合がある．辞書を用いたパターンマッチ手法や，辞書に基づく教師なし固有表現抽出手法は，アノテーション作業を行わずに固有表現抽出器を開発することができる．辞書とルールに基づく手法 [2] や，辞書とタグなしコーパスによる手法 [1, 3] が提案されている．

辞書に現代日本語書き言葉均衡コーパスに出現した固有表現を形態素解析辞書に追加することで，再現率の向上が見られる [4]．組織名抽出に限定したタスクでは，語彙を追加すると再現率が向上し，適合率が低下した [5]．

それでは，抽出対象となる属性の語彙を可能な限り網羅した辞書により解析するとどうなるであろうか．また，辞書中の語彙の増やし方は性能にどのように影響するのだろうか．本稿では，辞書に登録する語彙を意図的に制御し，固有表現抽出に与える影響を調査する (図 1)．具体的には，パターンマッチの手法，形態素解析辞書への追加，教師なし固有表現抽出に対して，辞書の統制を行う．可能な限り網羅的に辞書を作成可能な組織名（法人格が株式会社の組織）に絞り，調査する．本稿での調査内容を次に示す．

- 辞書に含める組織を無作為に選び、辞書の規模を変化させた際の影響を示す
- 文字数による辞書中語彙の制御を行った場合の固有表現抽出への影響を示す
- 形態素解析辞書中の組織を除く名詞と重複する固有表現の有無が固有表現抽出に与える影響を示す

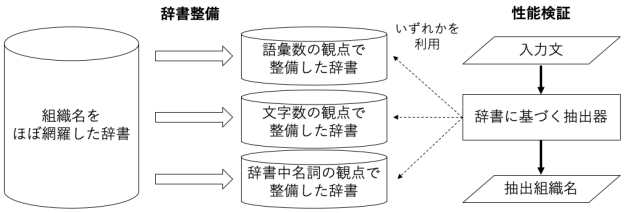

図 1: 辞書整備の影響調査

2　組織名抽出タスク

本稿での組織名抽出タスクとは，Web 上のニュース記事中の文から組織名を抽出することである．本稿での組織名の定義と対象とするニュース記事について説明する．

2.1　組織名

組織とは，法人番号が付与された法人のことである．法人番号は 1 法人に 1 番号付与され，その商号や所在地については，国税庁が公開する法人番号公表サイト[1]で調べることができる．

ここでの組織名は，組織の商号（登記上の名前）および法人格を省略したものとする[2]．略称については取り扱わない．

同一の組織名の異なる法人には，それぞれ法人番号が割り当てられている．ただし，今回のタスクでは組織名文字列の抽出であるため法人の曖昧さの考慮は不要である．すなわち，収集した組織名でユニークなものを辞書登録の候補とする．

2.2　ニュース記事

Web 上で公開されているニュース記事のタイトルと冒頭部を対象とする．おおよそ RSS フィードの概

[1] https://www.houjin-bangou.nta.go.jp/
[2] 「Sansan」と「Sansan 株式会社」が組織名となる．

要と同程度の文量である．記事の例 (内容は架空) を示す．

見出し Sansan が自然言語処理領域に参入

本文冒頭 法人向けクラウド名刺管理サービス「Sansan」を提供する Sansan 株式会社は，自然言語処理領域の研究開発に取り組むことを発表しました．名刺管理に加え，ニュース機能が強化されます．

サービス名 (*Sansan*) と社名 (*Sansan*) が同一文字列だったり，社名の手がかりとなる法人格 (株式会社) が省略されたりする．

3　組織名抽出の手法

本稿で検証対象とする組織名抽出手法について説明する．辞書とコーパスを持っている場合に実施できる手法としてパターンマッチ，形態素解析辞書への追加，辞書に基づく教師なし固有表現抽出を対象とする．

3.1　パターンマッチ

辞書中の語彙と文中の形態素列の表層形が一致する場合に組織名として認定する．文頭から最長一致の組織名を抽出する．形態素解析には MeCab [3]，形態素解析辞書には IPA 辞書を用いる．

3.2　形態素解析辞書への追加

形態素解析辞書 IPA 辞書に組織名[4]として追加し，MeCab による形態素解析の結果から組織名を取り出す．

MeCab IPA 辞書を基本の辞書とし，そこに任意の組織名辞書を追加する．コストは MeCab のコスト自動推定機能を用いる．

3.3　辞書に基づく教師なし固有表現抽出

辞書に基づく教師なし固有表現抽出に Tie or Break 手法 [1] を用いる．Distant Supervision による教師データ作成の際に，辞書を利用する．実装は AutoNER[5]を用いる．

[3] https://taku910.github.io/mecab/
[4] IPA 辞書中の組織名は ”名詞, 固有名詞, 組織”.
[5] https://github.com/shangjingbo1226/AutoNER

4　辞書整備

4.1　基本となる組織名辞書

国税庁で公開されている情報を元に収集した辞書を用いる．日本の法人をほぼ網羅した辞書である．

法人のうち，法人格「株式会社」を先頭もしくは末尾に含む組織名を今回の対象とする．該当する組織名の文字列の種類数は，約 160 万である．このうち，先頭もしくは末尾の「株式会社」を削除した際の組織名の種類数は約 152 万である．上記の 2 種類を結合した約 312 万語彙を基とし，辞書整備を行い効果検証する．

4.2　辞書の語彙数の整備

語彙数が増えれば増えるほど再現率が向上するが，無作為に辞書を作った際にどのような影響があるかを調査する．

収録対象とする語彙は無作為に選ぶため，評価用データセットや開発用データセットに存在する語彙かどうかは問わない．1 万社から 9 万社まで，1 万社ずつ増やして調査する．それぞれで無作為に語彙を選ぶため，1 万社と 5 万社で同じ語彙が含まれているとは限らない．N 万社取得後，法人格つきの組織名をまず辞書に加える．次に，法人格を削除したリストを作成し，重複を削除した上で辞書に追加する．これは，「株式会社 HOGEHOGE」と「HOGEHOGE 株式会社」が候補として選ばれた場合に「HOGEHOGE」が重複するためである．

4.3　文字数による整備

組織名として辞書に登録するかどうかを文字数により決める．パターンマッチや形態素解析辞書への追加においては，短い文字数の語彙が辞書に含まれることで，過剰に適合する．文字数で制御することによりどのような影響があるかを調査する．

図 2 に，文字数の語彙がいくつ含まれているかを示す．20 文字以上は種類数が少ないため省略する．N 文字以上を対象にした際に，組織名抽出の性能にどのような影響があるかを調査する．辞書には，法人格ありとなしの両方を登録するが，文字数には法人格なしのものを用いる．

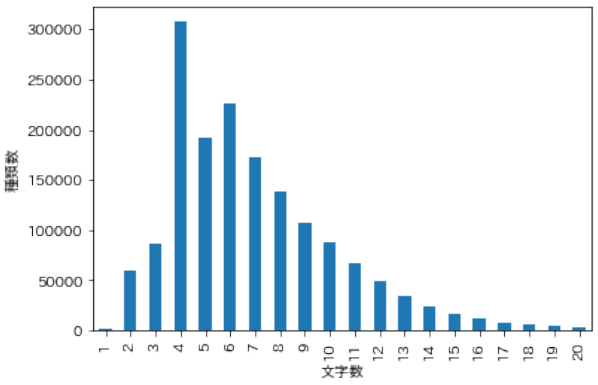

図 2: 法人格を除いた組織名の文字数とその種類数

4.4 辞書中名詞との重なりによる整備

辞書中の名詞 (組織以外の名詞) と同一文字列の組織名は，パターンマッチや形態素解析辞書への追加による方法では，過剰に適合することが考えられる．例えば，一般名詞で頻出するであろう「世界」という名前の組織を辞書に登録した場合，適合率が大きく低下することが考えられる．約 23,000 種類の組織名が辞書中の名詞と重複している．

辞書中の名詞と重複する組織名を加えた場合と加えなかった場合の性能への影響を調査する．

5 評価実験

5.1 実験データ

評価用データとしてニュース記事から抽出した 100 文を用いる．見出しと同じもしくは類似する言い回しが本文冒頭で用いられることがあるため，記事からいずれかを抽出した．対象としたニュース記事はプレスリリースや経済系新聞である．著者が組織名のタグ付けを行った．

教師なし固有表現抽出の開発用データには，評価用データと同様の方法で抽出した 100 文を用いる．評価用データも同様に著者がタグ付けを行った．Distant Supervision 用のテキストには，同系統のニュース記事 10,000 文を用いる．単語ベクトルには、Wikipedia エンティティベクトル[6]を用いる。

評価は，適合率，再現率，F 値を算出する．正解データと同じ区間を抽出していれば正解とする．表 1, 2, 3 中のスコアは，「適合率, 再現率, F 値」の並びである．

[6]https://github.com/singletongue/WikiEntVec

5.2 辞書の語彙数

表 1 に結果を示す。パターンマッチおよび教師なし手法はいずれのスコアも 0 に近い．形態素解析辞書の結果はほとんど変化がない．これは，ベースとなる IPA 辞書に語彙を追加しているため，IPA 辞書中の組織名を取り出したということである．IPA 辞書単体では，適合率:0.20, 再現率:0.23, F 値:0.21 である．

無作為に語彙を選んだ場合は，いずれの手法でも固有表現を抽出できていない．無作為に選んだため，辞書の語彙が評価セットに含まれなかったと考えられる．教師なし手法は Distant Supervision によるデータ作成の際に，辞書中語彙と合致せずアノテーションができていなかった．1 万語彙の場合は，212,137 トークンのうち 345 トークンしか組織名として取り扱われていなかった．

抽出対象に頻出する組織名から辞書に追加するなど，辞書とコーパスの組み合わせを考慮する必要があると考えられる．

5.3 文字数

表 2 に結果を示す。パターンマッチおよび形態素解析辞書への追加では，対象とする文字数を増やすと再現率が低下し，適合率が向上した．教師なし手法も同じ傾向であった．

ここで，文字数 1 文字以上は基となる約 312 万語彙をすべて使用したことを表す．可能な限り網羅した辞書であったが，パターンマッチおよび形態素解析辞書への追加において再現率が 1.0 とはならなかった．アルファベット組織名中の空白が形態素解析によって消失し合致しなくなったことと，Unicode 正規化を行ったが記号の見た目が同一だが別の文字コードだったためである．

パターンマッチおよび形態素解析辞書で 1 文字や 2 文字が含まれる場合は，「業」や「こと」など頻出する形態素と組織名が一致し，誤りとなっていた，これらが適合率を下げている原因だと考えられる．

辞書を利用した抽出器を作る際には，対象となる語彙の文字数が長ければ長いほど正しく抽出できるといえる．

表 1: 語彙数 (表中のスコアは「適合率, 再現率, F 値」)

手法	10000 語	20000 語	30000 語	40000 語	50000 語	60000 語	70000 語	80000 語	90000 語
パターンマッチ	0.25,0.01,0.02	0.08,0.01,0.02	0.09,0.04,0.05	0.00,0.00,0.00	0.02,0.01,0.02	0.08,0.04,0.05	0.07,0.05,0.06	0.06,0.04,0.04	0.06,0.06,0.06
形態素解析辞書	0.22,0.25,0.22	0.19,0.24,0.20	0.22,0.26,0.24	0.17,0.24,0.19	0.20,0.24,0.22	0.20,0.26,0.22	0.20,0.27,0.23	0.18,0.25,0.21	0.21,0.31,0.25
教師なし手法	0.00,0.00,0.00	0.00,0.00,0.00	0.09,0.04,0.05	0.00,0.00,0.00	0.02,0.01,0.02	0.00,0.00,0.00	0.02,0.01,0.02	0.00,0.00,0.00	0.01,0.01,0.01

表 2: 文字数 (表中のスコアは「適合率, 再現率, F 値」)

手法	1 文字以上	2 文字以上	3 文字以上	4 文字以上	5 文字以上	6 文字以上	7 文字以上	8 文字以上	9 文字以上
パターンマッチ	0.10,0.98,0.19	0.14,0.98,0.25	0.24,0.93,0.38	0.27,0.81,0.41	0.41,0.72,0.53	0.44,0.48,0.46	0.49,0.33,0.39	0.61,0.26,0.36	0.81,0.20,0.32
形態素解析辞書	0.19,0.94,0.32	0.24,0.94,0.38	0.31,0.92,0.47	0.33,0.88,0.48	0.41,0.82,0.55	0.37,0.62,0.47	0.35,0.52,0.42	0.35,0.46,0.39	0.33,0.41,0.37
教師なし手法	0.09,0.88,0.17	0.13,0.86,0.22	0.20,0.79,0.33	0.26,0.80,0.39	0.39,0.65,0.49	0.40,0.60,0.48	0.46,0.36,0.41	0.47,0.26,0.33	0.50,0.19,0.27

表 3: 辞書中名詞との重なり (表中のスコアは「適合率, 再現率, F 値」)

手法	すべて	辞書中名詞と重複	辞書中名詞にはない
パターンマッチ	0.10,0.98,0.19	0.01,0.07,0.02	0.19,0.91,0.32
形態素解析辞書	0.19,0.94,0.32	0.15,0.25,0.19	0.20,0.93,0.34
教師なし手法	0.09,0.88,0.17	0.01,0.05,0.01	0.18,0.81,0.29

5.4 辞書中の名詞との重なり

表 3 に結果を示す．いずれの手法でも「辞書中名詞にはない組織名」「すべて」「辞書中名詞と重複する組織名」の順に F 値と適合率が高い．「辞書中名詞ではない組織名」「すべて」を辞書に追加した場合，適合率が高い．

形態素解析辞書に追加する場合は，「すべて」と「辞書中名詞にはない組織名」とではあまり差が見られなかった．IPA 辞書単体では，適合率:0.20, 再現率:0.23, F 値:0.21 であるため，再現率を高めるという目的には語彙の形態素解析辞書への追加が有効だと考えられる．

「辞書中名詞と重複する組織名」のスコアが低いのは，該当する語彙の数が少ないからだと考えられる．「すべて」と「辞書中名詞にはない組織名」の辞書は 300 万語彙を超えるが，「辞書中名詞と重複する組織名」は約 5 万語彙である．

6 おわりに

本稿では，辞書としてほぼ網羅可能な組織名を対象に，辞書に基づく組織名抽出の辞書整備の影響について調査した．

ほぼ網羅する辞書を用いた抽出器を作ると，再現率は 1.0 に近づくものの適合率が低くなる．辞書が対象語彙を網羅することを維持できるのであれば，候補選出のために辞書を用い，文脈情報を利用した分類器を作るというのも一つの利用法と考えられる．

また，当たり前の話ではあるが，いくら組織名を集めてもコーパスに存在しない場合は手法が成立しない．辞書を基にした手法が必要な場合，どのように語彙を追加すべきかはわからないため，今後の課題である．

参考文献

[1] Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2054–2064, 2018.

[2] 竹元義美, 福島俊一, 山田洋志. 辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出. 情報処理学会論文誌, Vol. 42, No. 6, pp. 1580–1591, 2001.

[3] 土田正明, 水口弘紀, 久寿居大, 大和田勇人. 辞書とタグ無しコーパスを用いた固有表現抽出器の学習法. 人工知能学会第 23 回全国大会論文集, 2009.

[4] 岩倉友哉. 固有表現抽出におけるエラー分析. 言語処理学会第 21 回年次大会 ワークショップ, 2015.

[5] 落谷亮. 組織名抽出のための知識収集. 言語処理学会第 5 回年次大会, 1999.