固有表現抽出におけるエラー分析

株式会社富士通研究所 岩倉 友哉

iwakura.tomoya@jp.fujitsu.com

1 はじめに

本報告書では，固有表現抽出グループの Project Next NLP での活動として行ったエラー分析結果の概要および，固有表現抽出における辞書の利用の効果の調査結果について報告する．2 節にて，2015 年 3 月までの分析の進め方について紹介し，3 節で，分析に用いたデータセットを紹介する．4 節にて，辞書の影響調査結果を報告し，最後に，5 節で，本報告書をまとめる．

2 分析の進め方

固有表現抽出は，テキストに出現する人名や地名などの固有名詞や，日付や時間などの数値表現を抽出する技術であり，日本語の固有表現の種類としては，IREX [1] の 8 種類，約 200 種類が定義された拡張固有表現 [3] が提案されている．

今回の分析対象としては，IREX の定義に基づく固有表現抽出とした．その理由としては，メンバー[1]のほとんどは，固有表現抽出に取り組むのが初めてであり，分析の前に，固有表現抽出とは何か，どのような既存研究があるのか，といったところから始める必要があった．そこで，まずは，少ない種類で定義された固有表現抽出の問題の方が理解が容易と考え，IREX の定義を対象にすることにした．

2015 年 3 月までの進め方としては，次のステップで実施した．

- 事前準備
  - 既存のデータの入手
  - IREX の定義を元に BCCWJ にタグ付け (3 節)．分析対象データの拡張および，固有表現抽出タスクの理解のため．
  - 評価に向けた既存手法の調査：IREX の定義に基づく固有表現抽出機能を有する KNP を中心に調査．[2]
- 結果分析
  - (1) KNP の誤りパタンの分析 [2]:
    固有表現の範囲認識の誤り，固有表現のタイプ判別誤り，抽出漏れといった誤りパタンの観点からの KNP の分析．
  - (2) ドメイン別データの学習による精度評価および KNP との比較 [5]:
    異なるドメインのデータでの精度調査および学習の効果調査．
  - (3) 固有表現抽出における辞書利用に関する調査: 本報告書

本報告書では，(3) について報告する．(1)，(2) については，[2, 5] を参照願いたい．

3 データセット

本節では，本活動の分析で用いたデータセットを紹介する．

3.1 IREX データセット

IREX で配布された 毎日新聞 95 年にタグ付けされた CRL データおよび，IREX のドライランデータを準備した．[3]

[1] 最後にメンバー一覧を載せる．

[2] 今回，IREX の評価データにおいて高い精度を示すフリーソフトであり，論文だけでなく，開発者からも KNP の固有表現抽出手法に関する情報を得ることができ，内部の動作が，ある程度明らかになることから，KNP を評価対象として選定した．

[3] 次のパッケージを利用．http://nlp.cs.nyu.edu/irex/Package/IREXfinalB.tar.gz

3.2 BCCWJ コーパス

今回の活動の中で，以下の理由から BCCWJ コーパスの 136 ファイルに対し，メンバーで手分けして，タグ付を行なった．[4]

- 分析対象の文書の種類を広げるため．BCCWJ には，IREX が対象とした新聞記事だけでなく，白書，書籍，雑誌，Yahoo!知恵袋など，様々な種類のデータが含まれるため，学習データとテストデータで対象文書の違いがある場合の調査など，分析の幅が広がると考えたためである．
- タグ付けを通し対象の固有表現抽出のタスクを理解してもらう．

3.3 京都大学 Web 文書リードコーパス

Web 上の約 2500 文書の冒頭 3 文，合計約 7500 文を対象に，形態素解析，係り受け解析，IREX 定義の固有表現タグなどを付与したコーパス．このデータは，[5] の評価で利用した．

4 分析概要

本報告書では，辞書・学習データに含まれているか否かの観点での誤り分析結果および，形態素解析辞書の追加による精度変化の調査結果を報告する．今回の調査は，Ubuntu 上で，knp-4.12 および juman-7.01 を 用いて行なった．knp-4.12 は，CRL データを基に学習されているので，今回の調査では，Juman の辞書および CRL データを用いて調査を行なう．

4.1 学習データ・辞書の影響調査

DATE，MONEY，PERCENT，TIME は，規則による抽出でも比較的高い精度が得られることから [4]，今回，辞書の影響が大きいと考えられる ARTIFACT，ORGANIZATION，LOCATION，PERSON を対象に調査を行なった．表 1，表 2, 表 3 に調査結果を載せる．

この評価では，学習データに固有表現として出現したか (表 1)，Juman の辞書に登録されていたか (表 2)，その両方 (表 3)，という 3 つの観点で，正解・不正解を調査した結果である．Juman の辞書は，固有名詞は，ARTIFACT，組織名は，ORGANIZATION，地名は，LOCATION，人名は，PERSON とした．

各項目の () の中の値は正解・不正解に占める割合であり，正解 (+) の () の値であれば，

正解 (+) / (正解 (+) + 正解 (-))

となる．また，

C(+) = 正解 (+) /(正解 (+) + 不正解 (+))

C(-) = 正解 (+) /(正解 (-) + 不正解 (-))

である．

全体的に，C(+) の値が高いことから，学習データ，形態素解析の辞書に出現した場合は高い精度が得られていることがわかる．

また， C(-) の値が低いことから，学習データに出現せず，また，辞書に含まれない場合は，精度が低くなる傾向にあることがわかる．PERSON は，他の 3 種類の NE に比べて高い精度となっているが，これは，PERSON が，姓名と二つの単語で表現されることも多く，どちらが学習データか辞書に含まれていることで，抽出できることも少なくなく，C(-) の値が高めになっていると考えられる．

表 2 が示すように，今回の BCCWJ における調査では，学習データ上に出現した ARTIFACT は 1 種類（ノーベル賞が 2 回）だけであり，Juman の辞書に ARTIFACT は一度も出現しておらず，精度も最も低いという結果になった．また，ARTIFACT は，辞書に登録されていても正しく抽出できていない場合があることがわかる．これらの単語は,「ガンダム」(1 回出現) と「ポケモン」(2 回出現) であった．

4.2 辞書の効果調査

続いて，辞書追加による精度改善の可能性および問題点を調査するために，BCCWJ に出現した NE を KNP で使われる形態素解析器 Juman の辞書に追加した場合の精度の調査を行なった．

Juman の辞書として登録する際， 品詞は，ARTIFACT は固有名詞，ORGANIZATION は組織名, LOCATION は地名，PERSON は人名とした．つまり，この実験では，BCCWJ の全ての ARTIFACT，ORGANIZATION，LOCATION および PERSON が，対応する品詞とともに登録されている状態での抽出精度である．表 4 が辞書追加後の結果である．

辞書を追加した ORGANIZATION, LOCATION, PERSON については，10 ポイント以上の Recall の改善が見られ，F-measure は大幅に向上した．この結果から，ORGANIZATION, LOCATION, PERSON

[4]対象は，次のページにある形態素解析グループと同じファイルとした．http://plata.ar.media.kyoto-u.ac.jp/mori/research/NLR/JDC/ClassA-1.list

表 1: 正解・誤り数の調査．”+” は学習に用いられた CRL データに固有表現として出現した場合を意味する．”-” は出現しなかった場合．

NE	正解 (+)	正解 (-)	不正解 (+)	不正解 (-)	C(+)	C(-)
ARTIFACT	2 (2.19)	89 (97.80)	0 (0)	224 (100)	100	28.43
LOCATION	301 (74.87)	101 (25.12)	34 (22.36)	118 (77.63)	89.85	46.11
ORGANIZATION	61 (27.60)	160 (72.39)	26 (17.80)	120 (82.19)	70.11	57.14
PERSON	36 (9.89)	328 (90.10)	5 (2.70)	180 (97.29)	87.80	64.56

表 2: 正解・誤り数の調査．”+” は Juman の辞書に固有名詞・人名・組織名・場所として登録されていた場合を意味する．”-” は出現しなかった場合．() の値は正解・不正解に占める割合．

NE	正解 (+)	正解 (-)	不正解 (+)	不正解 (-)	C(+)	C(-)
ARTIFACT	0 (0)	91 (100)	3 (1.33)	221 (98.66)	0	29.16
LOCATION	290 (72.13)	112 (27.86)	35 (23.02)	117 (76.97)	89.23	48.90
ORGANIZATION	33 (14.93)	188 (85.06)	3 (2.05)	143 (97.94)	91.66	56.79
PERSON	100 (27.47)	264 (72.52)	10 (5.40)	175 (94.59)	90.90	60.13

は，形態素解析の辞書に正しい品詞で登録することで，精度向上に貢献すると期待され，辞書の獲得が，精度向上の一つの要因になりそうである．しかしながら，

- 文脈によって,「クマ」,「タマ」のような普通名詞とも固有名詞とも取れるものの誤り．たとえば，かなり広範囲の文脈を使わないと抽出できないものがある．たとえば，「クマには命を助けられたことがある。」という文は，前後の文を見ることが,「クマ」の意味の区別に必要であった．
- 大川，勝田のように人名・場所と複数の固有表現になる場合の誤り．

場合や，

- 正解：⟨ORGANIZATION⟩ バンプレスト ⟨/ORGANIZATION⟩ さん ••••
- KNP: ⟨PERSON⟩ バンプレスト ⟨/PERSON⟩ さん ••••

のように，形態素解析結果が正しい品詞を与えたとしても，文脈から誤った抽出を行なっている例も見られた．

また，ARTIFACT は，辞書追加により，ARTIFACT と認識される数も減少し，Recall，Precision，F-measure が低下した．これは，次のような原因が考えられる．

- 形態素解析結果の複数の単語から構成される ARTIFACT を，構成する一部の形態素を手書かりに抽出していたが，辞書登録されてしまったため，その手掛りが使えなくなった．たとえば,「文化多様性条約」辞書追加により一単語として認識されるようになったため,「条約」という ARTIFACT を判別するための手掛りが使えなくなったと思われる．
- Juman における固有名詞という品詞は，ARTIFACT 以外に相当する場合にも用いられているため，判別の手掛りとしては十分でないと考えられる．たとえば,「ワールドカップ」といったイベント名，曹洞宗といった宗教名に用いられている．

これらの結果は，形態素解析辞書への固有表現の追加が，悪影響となる場合があることを示唆している．

これらの結果から，固有表現の形態素解析辞書への追加は，ある一定の効果が得られるが，それだけで，全ての項目の抽出に対応するのは難しいと言えそうである．特に，今後の精度改善のためには，次のような事が課題となると考える．

- 獲得した辞書と文脈情報を考慮した学習
  - たとえば,「クマ」,「タマ」のように普通名詞としても，固有表現としても使われる場合，大川，勝田のように人名・場所と複数の固有表現になる場合など，文脈によって異なる意味を持つ単語への対処．
- ARTIFACT のような，パタンの学習が難しい固有表現への対処，

表 3: 正解・誤り数の調査．”+” は学習に用いられた CRL データに固有表現として出現したか，Juman の辞書に固有名詞・人名・地名・組織名として登録されていた場合を意味する．”-” はどちらにも出現しなかった場合.

NE	正解 (+)	正解 (-)	不正解 (+)	不正解 (-)	C(+)	C(-)
ARTIFACT	2 (2.19)	89 (97.80)	3 (1.33)	221 (98.66)	40	28.70
LOCATION	328 (81.59)	74 (18.40)	39 (25.65)	113 (74.34)	89.37	39.57
ORGANIZATION	71 (32.12)	150 (67.87)	27 (18.49)	119 (81.50)	72.44	55.76
PERSON	108 (29.67)	256 (70.32)	14 (7.56)	171 (92.43)	88.52	59.95

表 4: BCCWJ での評価結果．KNP は knp-4.12 + juman-7.01 による結果．”KNP+辞書” は BCCWJ に出現した NE を Juman の辞書に追加して，実行した結果.

	KNP			KNP+ 辞書		
NE	Recall	Precision	F-measure	Recall	Precision	F-measure
ARTIFACT	**28.89**	**64.54**	**39.91**	14.29	47.37	21.95
LOCATION	72.56	73.76	73.16	**86.82**	**87.61**	**87.22**
ORGANIZATION	60.22	63.69	61.90	**83.65**	**74.70**	**78.92**
PERSON	66.30	73.68	69.80	**81.60**	**80.14**	**80.87**

5 まとめ

本活動では，KNP を対象に，KNP の誤りパタンの分析，ドメイン別データの学習による精度評価および KNP との比較，異なるドメインのデータでの精度調査および学習の効果調査．辞書の影響調査の，3 つの分析を行ない，本報告書では，特に，辞書の影響調査の報告を行なった．BCCWJ 中の NE を形態素解析の辞書として登録した場合の実験では，ARTFACT では低下したものの， ORGANIZATION，LOCATION，PERSON においては，大きな精度改善が見られ，結果の分析から，解決すべき課題が明らかになった．

今後は，今回対象としなかった，DATE，MONEY，PERCENT，TIME に対する分析および，今回得られた知見を基にしたさらなるエラー分析を進め，辞書や新しい素性を追加により，誤りを解決できるかの調査に進みたい．

メンバー (2015 年 3 月時点)

- 岩倉友哉 （株式会社富士通研究所）
- 古宮嘉那子 （茨城大学，講師）
- 市原正陽 （茨城大学，B4）
- 立花竜一 （首都大学東京，M2）
- 平田亜衣 （首都大学東京，M1）
- 山崎舞子 （東京工業大学，M1）

謝辞

本活動にあたり，京都大学 Web 文書リードコーパスを，京都大学の河原准教授からご提供いただきました．また，東工大の笹野助教からは，knp の実装の詳細につきまして，お教えいただきました．ここに感謝の意を表します.

参考文献

[1] IREX Committee. *Proc. of the IREX workshop*. 1999.

[2] Masaaki Ichihara, Kanako Komiya, Tomoya Iwakura, and Maiko Yamazaki. Error analysis of named entity recognition in bccwj. In **エラー分析ワークショップ (言語処理学会年次大会** *2015*) .

[3] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proc. of LREC'02*, 2002.

[4] 竹本義美 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**, 42(6):1580–1591, 2001.

[5] 平田亜衣 小町守. 様々なジャンルのテキストに対する固有表現認識の分析. In **エラー分析ワークショップ** *(***言語処理学会年次大会** *2015*) .