固有表現抽出におけるアノテーション手法の比較

著者	鈴木 雅也, 古宮 嘉那子, 岩倉 友哉, 佐々木 稔, 新納 浩幸
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	385-403
発行年	2017
URL	http://doi.org/10.15084/00001494

固有表現抽出におけるアノテーション手法の比較

鈴木雅也 (茨城大学)*
古宮嘉那子 (茨城大学)†
岩倉友哉 (富士通研究所)‡
佐々木稔 (茨城大学)§
新納浩幸 (茨城大学)¶

Comparison of Annotating Methods in Named Entity Extraction

Masaya Suzuki (Ibaraki University)
Kanako Komiya (Ibaraki University)
Tomoya Iwakura (Fujitsu Laboratories Ltd.)
Minoru Sasaki (Ibaraki University)
Hiroyuki Shinnou (Ibaraki University)

要 旨

本稿では, 非専門家による固有表現抽出のタスクとしてのアノテーションを題材に, ふたつの手法について比較を行った. ひとつは既存の固有表現抽出器によるアノテーション結果に対し, 人手で修正を行う手法であり, もうひとつは人手で一からアノテーションを行う手法である. 実験には現代日本語書き言葉均衡コーパス (BCCWJ) を利用し, 手法ごとに 1 テキストに対し 2 人の非専門家を割り当てて, アノテーションを行った. 評価には, アノテーションにかかる時間, 一致率, Gold Standard との比較による正解率, それぞれの手法で作成されたコーパスを訓練事例とした場合の正解率を利用し, ジャンルごと, 及び, 全ジャンルのマイクロ平均とマクロ平均を算出した. 本実験の結果から, 全ジャンルのマイクロ平均とマクロ平均で比較した場合には既存のアノテーション結果を用いた手法の方が良い結果となるが, 既存の固有表現抽出器の訓練事例から離れたジャンルで同様に比較した場合には人手でアノテーションを行う手法の方が良い結果となることが明らかになった.

1. はじめに

非専門家をアノテータとする, クラウドソーシングによるコーパスへのアノテーションは, 安価で速く仕上がることが Snow ら (Snow et al. 2008) によって明らかとなっている. しかし, アノテーション手法に起因したアノテーションの品質の違いについては, これまで言及さ

* 13t4038a@vc.ibaraki.ac.jp
† kanako.komiya.nlp@vc.ibaraki.ac.jp
‡ iwakura.tomoya@jp.fujitsu.com
§ minoru.sasaki.01@vc.ibaraki.ac.jp
¶ hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

れてこなかった. 固有表現抽出におけるアノテーションはルールが多く複雑なため, 非専門家にとってタグの付け間違いが発生しやすいタスクとなっており, この観点での議論が必要なタスクのひとつであると考えられる. そこで, 本稿では, 固有表現抽出におけるアノテーションを題材として, 非専門家の手で高品質なコーパスを作成するための手法についての考察を行った. なお, 本稿は (Komiya et al. 2016) を元にしている.

固有表現抽出におけるアノテーションでのタグの付け間違いを減らすための手法として, 既存の固有表現抽出器によるアノテーション結果に対し, 人手で修正を行う手法が考えられる. しかし, 訓練事例として特定ジャンルのコーパスのみを用いている固有表現抽出器の場合, 特にそのジャンル以外のコーパスのアノテーションにおいて, タグの付け間違いが発生することがある. そこで, 本研究では, 前述の手法と既存の固有表現抽出器を使用せず, 人手でアノテーションを行う手法のふたつの手法について, アノテーションにかかる時間, タグの一致率, Gold Standard との比較による正解率の各観点から比較することで考察を行った. この際, テキストのジャンルに起因したアノテーションの品質の違いについても考察を行っている.

2. 関連研究

アノテーションに関する先行研究としては, 次のようなものが挙げられる. Snow ら (Snow et al. 2008) は, 非専門家によるコーパスへのアノテーションに関して, アノテーションにかかる時間, アノテーションの品質, コストの観点から, 専門家が行った場合と比較することで考察を行った. Alex ら (Alex et al. 2010) は, 反復的で agile なアノテーション手法を提案し, 既存の線形によるアノテーション手法との比較を行った. van der Plas ら (Plas et al. 2010) は, 英語のテンプレートを用いたフランス語のコーパスへの意味情報の付与を題材に, 言語横断的なアノテーションの信頼性について考察を行った. Marcus ら (Marcus et al. 1993) は, 品詞アノテーションや bracketing といったタスクのための Penn TreeBank を開発するため, 既存のアノテーション結果を用いる手法と人手のみで行う手法について比較を行った. しかし, 我々が知る限り, 非専門家の手で高品質なコーパスを開発するために, 既存のアノテーション結果を用いる手法と人手のみで行う手法を比較したという論文は存在しない.

本稿では, 固有表現抽出におけるアノテーションを題材にしている. 固有表現抽出とは, 固有名詞に時間や数値といった表現を加えた概念である固有表現を文章中から抽出するタスクであり, 昔から研究が行われてきた. このタスクに関する先行研究としては, 次のようなものが挙げられる. 橋本ら (橋本泰一ほか 2008, 橋本泰一・中村俊一 2010) は CD-毎日新聞'95 データ集(1)や現代日本語書き言葉均衡コーパス (BCCWJ) (Maekawa 2008)(2) を元に, 拡張固有表現タグ付きコーパス(3)を作成した. 徳永ら (徳永健伸ほか 2015) は, 固有表現抽出におけるアノテータの視線分析を行った. Sasada ら (Sasada et al. 2015) は, 部分的なタグ付きテキストを用いて訓練可能な固有表現抽出器を提案した. Sekine ら (Sekine and Isahara 2000)

(1) http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
(2) http://pj.ninjal.ac.jp/corpus_center/bccwj/
(3) http://www.gsk.or.jp/catalog/gsk2014-a/

は Message Understanding Conference-6 (MUC-6)[4] での定義 (Grishman and Sundheim 1996) を元に, Information Retrieval and Extraction Exercise (IREX)[5] で固有表現抽出の共通タスクを行うため, 8 種類の固有表現タグ (組織名, 人名, 地名, 固有物名, 日付表現, 時間表現, 金額表現, 割合表現), 及び, それらと同等に扱われるオプショナルタグからなる 9 種類のタグを定義した. しかし, IREX で用いられたのは, 新聞コーパスのみであった.

2014 年, 6 領域から構成される Project Next NLP (岩倉友哉 2015, 平田亜衣・小町守 2015, Ichihara et al. 2015)[6] において, 前述の拡張固有表現タグ付きコーパスを用いた固有表現抽出のエラー分析が行われた. Ichihara ら (Ichihara et al. 2015) は, 既存の固有表現抽出器の性能について調べ, 固有表現抽出器の訓練事例から離れたジャンルのテキストにおいて, タグの付け間違いが増加することを示した. 本稿では, 訓練事例から離れたジャンルのコーパスにおいて, 既存のアノテーション結果を用いた手法ではタグの付け間違いが発生する場合があるということを示す.

本研究では, 非専門家の手で高品質なコーパスを作成するため, 固有表現抽出のタスクについて, 既存のアノテーション結果を用いた手法と人手のみでアノテーションを行う手法のふたつの手法による, アノテーションにかかる時間, タグの一致率, Gold Standard との比較による正解率の評価を行った.

3. アノテーション手法の比較

本稿では, 次のふたつのアノテーション手法について比較を行った.

- KNP+Manual
  既存の固有表現抽出器 KNP (Sasano and Kurohashi 2008)[7] によるアノテーション結果に対し, 人手で修正を行う.
- Manual
  人手で一から固有表現のアノテーションを行う.

また, 比較を行うにあたり, それぞれのテキストに対するアノテーションにかかる時間, タグの見かけの一致率とカッパ係数, Gold Standard との比較による適合率 (精度), 再現率, F 値を指標として設定した.

ふたつの手法間で一致したタグの個数が表 1 で示されるとき, 見かけの一致率とカッパ係数はそれぞれ式 (1) と 式 (2) で与えられる.

$$d = \frac{\sum_{i=1}^{n} a_{ii}}{a_{00}} \tag{1}$$

[4] http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
[5] http://nlp.cs.nyu.edu/irex/index-j.html
[6] https://sites.google.com/site/projectnextnlp/
[7] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

表 1　ふたつの手法間で一致したタグの個数

		手法 X				
		タグ 1	タグ 2	...	タグ n	合計
手法 Y	タグ 1	a_{11}	a_{21}	...	a_{n1}	a_{01}
	タグ 2	a_{12}	a_{22}	...	a_{n2}	a_{02}
	...	...	...	...	...	...
	タグ n	a_{1n}	a_{2n}	...	a_{nn}	a_{0n}
	合計	a_{10}	a_{20}	...	a_{n0}	a_{00}

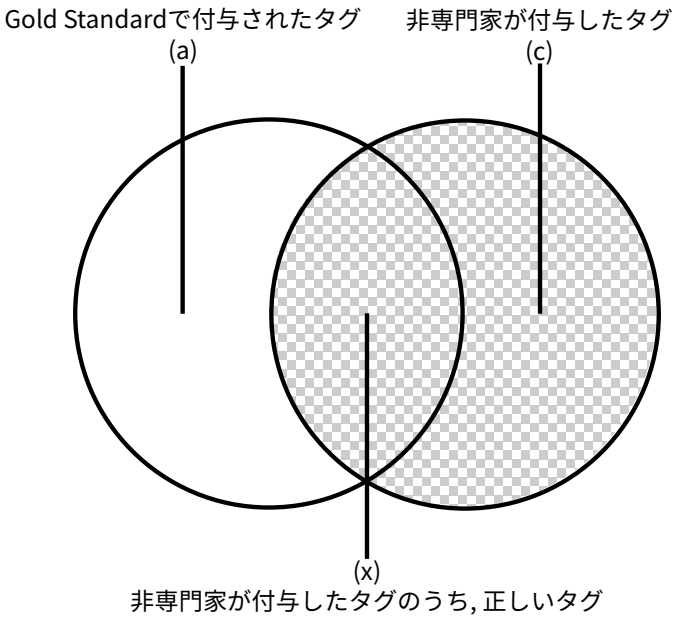

図 1　タグ集合

$$\kappa = \frac{a_{00}\sum_{i=1}^{n} a_{ii} - \sum_{i=1}^{n} a_{i0}a_{0i}}{(a_{00})^2 - \sum_{i=1}^{n} a_{i0}a_{0i}} \tag{2}$$

また, タグ集合が図 1 のように示されるとき, 適合率, 再現率, F 値はそれぞれ 式 (3), 式 (4), 式 (5) のように与えられる.

$$p = \frac{n(x)}{n(c)} \tag{3}$$

$$r = \frac{n(x)}{n(a)} \tag{4}$$

$$f = \frac{2pr}{p+r} \tag{5}$$

表 2　ジャンルごとのテキストとそこに含まれるタグの数

ジャンル	テキスト	タグ									
		Artifact	Date	Location	Money	Organization	Percent	Person	Time	Optional	合計
OC	74	44	18	65	9	18	0	6	0	8	168
OW	8	86	143	147	9	136	33	15	0	26	595
OY	34	23	61	59	7	64	10	79	3	17	323
PB	5	32	49	100	0	19	5	174	9	20	408
PM	2	9	24	36	5	18	1	216	3	1	313
PN	13	24	166	192	60	123	37	78	22	20	722
合計	136	218	461	599	90	378	86	568	37	92	2, 529

4. 実験

本実験では, ClassA-1[(8)] に分類される 136 テキストを BCCWJ より抽出して用いた. ClassA-1 に分類される BCCWJ のテキストは, Yahoo! 知恵袋 (OC), 白書 (OW), Yahoo! ブログ (OY), 書籍 (PB), 雑誌 (PM), 新聞 (PN) の 6 ジャンルで構成されている. それぞれのジャンルにおけるテキストとそこに含まれるタグの数は表 2 の通りである. なお, 本実験では固有表現抽出器として KNP Ver.4.16 (Linux 版) と JUMAN Ver.7.01 (Linux 版)[(9)] を用いており, 前者は訓練事例として新聞コーパスを用いている (Sasano and Kurohashi 2008)[(10)].

被験者は非専門家 16 人であり, IREX によるアノテーションのルール (Inf 1999) を読み合わせた後, これに従って 9 種類のタグによるアノテーションを行った. この際, 全ての被験者のアノテーション結果を集めたときに, それぞれの手法について, 2 セットのコーパスを構成できるよう, 被験者は割り当てられた 34 テキストに対し, それぞれの手法を半分ずつ適用した. また, 習熟によるバイアスがかかりにくくするため, 被験者をふたつのグループに分け, 最初に適用する手法をグループごとに変えた. なお, アノテーションの際には, テキストごとのアノテーションにかかる時間の記録も行っており, それを元に手法ごとのアノテーションにかかる平均時間を算出した. また, 本実験では Gold standard として BCCWJ NE コーパス (2016 年 2 月 1 日版) (Iwakura et al. 2016)[(11)] を用いている.

Gold standard は IREX によるアノテーションのルールに基づき作成した. Gold standard にオプショナルタグが付与されているときはその範囲を超えてタグが付与されていない場合を, それ以外のときはタグとその範囲が Gold standard と一致している場合を正解としている.

本実験では, 手法ごとに 1 テキストに対し 2 人の非専門家を割り当てて, アノテーションを行ったという条件下で, 2 人のアノテータの平均正解率と, どちらか一方でも正解のタグを付与しているならば正解とみなした際の正解率を算出した. 後者は, 実際にコーパスを作成する際, 2 人のアノテータによるアノテーション結果を統合して作成することが想定されるため, 算出を行った.

これらに加え, 機械学習における訓練事例としての品質を確かめるため, それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いてアノテーションを行った. この際に用い

(8) http://plata.ar.media.kyoto-u.ac.jp/mori/research/NLR/JDC/ClassA-1.list

(9) http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN

(10) 厳密には Web 上の記事も訓練事例として用いられているが, 本稿では訓練事例としてのウエイトが大きい新聞コーパスを KNP の訓練事例として扱っている.

(11) https://sites.google.com/site/projectnextnlpne/

表 3　一致率のマイクロ平均 (全体)

手法	見かけの一致率	カッパ係数
KNP+Manual	**0.79**	**0.75**
Manual	0.57	0.50
Both	0.64	0.58

表 4　一致率のマクロ平均 (全体)

手法	見かけの一致率	カッパ係数
KNP+Manual	**0.66**	**0.48**
Manual	0.52	0.29
Both	0.52	0.31

た素性は, 形態素, 文字種, 品詞タグ, 分類[(12)], キャッシュ素性, 統合素性, 格フレーム素性であり, これはオリジナルの KNP と同様である (Sasano and Kurohashi 2008). なお, それぞれの手法における 2 人分のアノテーション結果を結合したものをその手法の訓練事例としており, また, できる限り多くのジャンルのテキストを含むような形で 5 分割交差検定を行っている.

5. 結果

表 3, 表 4 はそれぞれの手法の見かけの一致率とカッパ係数のマイクロ平均とマクロ平均を示しており, 表 5, 表 6 はそれらをジャンルごとに示したものである. これらにおける **Both** はふたつの手法を用いた計 4 人のアノテータによるアノテーション結果の全てのペアを取ったときの一致率の平均を示している.

表 7, 表 8 はそれぞれの手法の適合率, 再現率, F 値のマイクロ平均とマクロ平均を示しており, 表 9, 表 10 はそれらをジャンルごとに示したものである. これらにおける **KNP** はオリジナルの KNP によるアノテーション結果の正解率を, **Average** は **KNP+Manual** と **Manual** の平均を示している.

なお, ふたつの手法の中でより高い水準を記録した見かけの一致率, カッパ係数, 適合率, 再現率, F 値については太字で示している. また, 表 11 は, それぞれの手法における 1 テキストあたりのアノテーションにかかる平均時間を示している.

次に, 2 人のアノテータのうち, どちらか一方でも正解のタグを付与しているならば正解とみなしたとき (2 人のアノテータによるアノテーション結果を統合したとき) の性能について調べた. 表 12, 表 13 はそれぞれの手法の適合率, 再現率, F 値のマイクロ平均とマクロ平均を示しており, 表 14, 表 15 はそれらをジャンルごとに示したものである. **KNP** と **Average** に関しては, 表 7〜表 10 と同様である.

これらに加え, それぞれの手法で作成されたコーパスを訓練事例とした固有表現抽出器の

(12) データとして存在する場合のみ.

表 5　一致率のマイクロ平均 (ジャンルごと)

ジャンル	手法	見かけの一致率	カッパ係数
OC	KNP+Manual	**0.62**	**0.54**
OC	Manual	0.47	0.34
OC	Both	0.52	0.41
OW	KNP+Manual	**0.78**	**0.73**
OW	Manual	0.41	0.28
OW	Both	0.55	0.46
OY	KNP+Manual	**0.69**	**0.63**
OY	Manual	0.58	0.50
OY	Both	0.57	0.49
PB	KNP+Manual	**0.76**	**0.68**
PB	Manual	0.67	0.56
PB	Both	0.71	0.61
PM	KNP+Manual	**0.87**	**0.84**
PM	Manual	0.61	0.55
PM	Both	0.69	0.64
PN	KNP+Manual	**0.86**	**0.75**
PN	Manual	0.81	0.65
PN	Both	0.80	0.65

性能を調べた. 表 16, 表 17 はそれぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の適合率, 再現率, F 値のマイクロ平均とマクロ平均を示しており, 表 18, 表 19 はそれらをジャンルごとに示したものである.

まず, マイクロ平均について比較する. 表 7, 表 16 における適合率と再現率, 表 14 における適合率について, 有意水準 0.05 のカイ二乗検定で検定を行った場合, **KNP** と **KNP+Manual**, **KNP** と **Manual**, **Manual** と **KNP+Manual** は統計的に有意である. また, 正解率におけるジャンルごとのマイクロ平均 (表 9, 表 14, 表 18) のうち, アスタリスクが付与されている箇所においては, 適合率, または, 再現率について同様に検定を行った場合, **Manual** と **KNP+Manual** は統計的に有意である. しかし, 表 12 において, 再現率について同様に検定を行った場合, **KNP** と **KNP+Manual**, **KNP** と **Manual** は統計的に有意であるが, **Manual** と **KNP+Manual** は有意ではない. また, 正解率のマクロ平均について同様に検定を行った場合, 標本数が少ないという理由から, 統計的に有意ではない.

表 6　一致率のマクロ平均 (ジャンルごと)

ジャンル	手法	見かけの一致率	カッパ係数
OC	KNP+Manual	**0.58**	**0.27**
OC	Manual	0.50	0.15
OC	Both	0.47	0.14
OW	KNP+Manual	**0.80**	**0.73**
OW	Manual	0.45	0.36
OW	Both	0.59	0.50
OY	KNP+Manual	**0.63**	**0.47**
OY	Manual	0.50	0.29
OY	Both	0.47	0.30
PB	KNP+Manual	**0.63**	**0.54**
PB	Manual	0.60	0.43
PB	Both	0.62	0.48
PM	KNP+Manual	**0.87**	**0.83**
PM	Manual	0.62	0.55
PM	Both	0.69	0.63
PN	KNP+Manual	**0.88**	**0.74**
PN	Manual	0.74	0.56
PN	Both	0.77	0.59

表 7　2 人のアノテータの平均正解率のマイクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.78	0.68	0.73
KNP+Manual	**0.84**	**0.81**	**0.83**
Manual	0.75	0.73	0.74
Average	0.80	0.77	0.78

6. 考察

6.1 一致率とアノテーションにかかる時間

表 3, 表 4 より, **KNP+Manual** の一致率は, **Manual** の一致率よりもマイクロ平均, マクロ平均ともに高い数値となっていることがわかる．また, 表 5, 表 6 より, 全てのジャンルについて同様の傾向が見られることがわかる．これは, **KNP+Manual** の人手により修正される前のコーパスが共に同じ固有表現抽出器によってアノテーションされたものであることが影

表 8　2 人のアノテータの平均正解率のマクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.47	0.40	0.43
KNP+Manual	**0.55**	**0.55**	**0.55**
Manual	0.53	0.51	0.52
Average	0.54	0.53	0.53

響していると考えられる．さらに, 表 11 より, **KNP+Manual** における 1 テキストあたりのアノテーションにかかる時間は, **Manual** よりも平均約 2 分程度短いということがわかる．これは有意水準 0.01 の F 検定で検定を行った場合, 統計的に有意である．これらのことから, **KNP+Manual** は **Manual** よりもアノテーションにかかる時間が短く, 一致率が高いということがいえる．

また, 表 5, 表 6 より, **Both** の一致率は多くの場合, **Manual** と同等以上の数値となっているが, OC における一致率のマクロ平均は, **Both** が 0.01 ポイント以上 **Manual** を下回っていることがわかる．このことから, OC には新聞コーパスから生成したルールだけでは抽出できないような固有表現が多く含まれているということがわかる．

さらに, 表 3, 表 4 より, **Manual** のカッパ係数に関して, マイクロ平均では適度な値だったのに対し, マクロ平均では低い値となっていることがわかる．マイクロ平均は固有表現ごとの平均であり, マクロ平均はテキストごとの平均であるということから, **Manual** ではテキストごとに見たときに, 一致率の偏りが大きいということがいえる．

6.2　正解率

表 7, 表 8 より, **KNP+Manual** の正解率は, **Manual** の正解率よりもマイクロ平均, マクロ平均ともに高い数値となっていることがわかる．しかし, 表 9 より, OC における再現率と PM における適合率のマイクロ平均についてはこの傾向が見られず, また, **KNP** におけるこれらの指標は, 他のジャンルよりもかなり低い値となっていることがわかる．このことから, **KNP+Manual** の正解率は **KNP** の正解率に依存しているということがいえる．

また, 表 10 より, **KNP+Manual** の正解率は, OY, OW, PN については **Manual** の正解率よりも高い値となっている一方, OC, PB, PM については, PM の再現率を除き **Manual** の正解率よりも低い値となっていることがわかる．さらに, KNP の訓練事例である新聞コーパスに近く, **KNP** による正解率が高くなることが示されている (Ichihara et al. 2015) OW と PN において, **KNP** の正解率は **Manual** の正解率よりも高い値となっている．これらのことから, 非専門家がアノテーションを行う場合, KNP の訓練事例に近いジャンルのテキストについては **KNP+Manual** の方が良い結果を得られ, KNP の訓練事例から離れたジャンルのテキストについては **Manual** の方が良い結果を得られるということがいえる．

表 9　2 人のアノテータの平均正解率のマイクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.72	0.48	0.57
OC	KNP+Manual	***0.78**	0.75	**0.77**
OC	Manual	0.67	**0.80**	0.73
OC	Average	0.72	0.78	0.75
OW	KNP	0.79	0.79	0.79
OW	KNP+Manual	***0.82**	***0.85**	**0.83**
OW	Manual	0.65	0.67	0.66
OW	Average	0.73	0.76	0.74
OY	KNP	0.73	0.57	0.64
OY	KNP+Manual	***0.85**	***0.75**	**0.80**
OY	Manual	0.80	0.68	0.74
OY	Average	0.83	0.72	0.77
PB	KNP	0.75	0.60	0.66
PB	KNP+Manual	**0.79**	**0.74**	**0.76**
PB	Manual	0.78	0.73	0.75
PB	Average	0.78	0.73	0.76
PM	KNP	0.61	0.58	0.59
PM	KNP+Manual	0.89	**0.86**	**0.87**
PM	Manual	**0.90**	0.85	0.87
PM	Average	0.89	0.86	0.87
PN	KNP	0.88	0.78	0.83
PN	KNP+Manual	***0.88**	***0.85**	**0.86**
PN	Manual	0.77	0.72	0.75
PN	Average	0.83	0.79	0.81

6.3　2 人のアノテータによるアノテーション結果を統合したときの正解率

表 12, 表 13 より, どちらか一方でも正解のタグを付与しているならば正解とみなした場合, **KNP+Manual** の正解率は **Manual** の正解率よりも高い値となっているが, 2 人のアノテータの平均正解率 (表 7, 表 8) に比べると, その差はかなり小さいということがわかる. これは, 2 人のアノテータのうち, 少なくともどちらか一方は正しいタグを付与していることが多いためであると考えられる.

また, 表 7, 表 8 は 2 人のアノテータの正解率の平均であることから 1 人のアノテータの正解率, 表 12, 表 13 は 2 人のアノテータの正解率とみなすことができる. すると, 表 7, 表 8, 表

表 10　2 人のアノテータの平均正解率のマクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.31	0.26	0.28
OC	KNP+Manual	0.39	0.41	0.40
OC	Manual	**0.42**	**0.44**	**0.43**
OC	Average	0.40	0.42	0.41
OW	KNP	0.77	0.80	0.79
OW	KNP+Manual	**0.83**	**0.86**	**0.84**
OW	Manual	0.70	0.73	0.71
OW	Average	0.76	0.79	0.78
OY	KNP	0.58	0.44	0.50
OY	KNP+Manual	**0.68**	**0.63**	**0.66**
OY	Manual	0.56	0.49	0.52
OY	Average	0.62	0.56	0.59
PB	KNP	0.66	0.46	0.54
PB	KNP+Manual	0.71	0.65	0.68
PB	Manual	**0.81**	**0.67**	**0.74**
PB	Average	0.76	0.66	0.71
PM	KNP	0.60	0.66	0.63
PM	KNP+Manual	0.82	**0.87**	0.85
PM	Manual	**0.86**	0.84	**0.85**
PM	Average	0.84	0.85	0.85
PN	KNP	0.88	0.78	0.82
PN	KNP+Manual	**0.88**	**0.85**	**0.86**
PN	Manual	0.78	0.72	0.75
PN	Average	0.83	0.78	0.81

12, 表 13 より, 2 人のアノテータによる正解率は常に 1 人のアノテータによる正解率よりも高い値となっていることがわかる. さらに, 2 人のアノテータによる **Manual** の正解率は, 常に 1 人のアノテータによる **KNP+Manual** の正解率よりも高い値となっていることがわかる. このことから, 非専門家をアノテータとする場合, 既存の固有表現抽出器を使用すること以上に, アノテータの人数を増やすことが良い結果を得る上で重要であるといえる.

6.4　訓練事例としてのアノテーション結果

表 16, 表 17 より, それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合, **KNP+Manual** を訓練事例とした場合の正解率は **Manual** を訓練事例とした場合の

表 11　アノテーションにかかる平均時間 (手法ごと)

手法	時間
KNP+Manual	3:19
Manual	5:23

表 12　2 人のアノテータによるアノテーション結果を統合したときの正解率のマイクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.78	0.68	0.73
KNP+Manual	**0.91**	**0.89**	**0.90**
Manual	0.87	0.88	0.88
Average	0.89	0.89	0.89

表 13　2 人のアノテータによるアノテーション結果を統合したときの正解率のマクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.47	0.40	0.43
KNP+Manual	**0.63**	**0.62**	**0.63**
Manual	0.62	0.62	0.62
Average	0.63	0.62	0.63

正解率よりも高い値となっていることがわかる．しかし, 表 18, 表 19 より, PB と PN における適合率のマイクロ平均とマクロ平均, 及び, PB における F 値のマクロ平均についてはこの傾向が見られないことがわかる．このことから, **KNP+Manual** よりも **Manual** を訓練事例とした方が良いアノテーション結果となる場合もあるということがわかる．

また, 表 16, 表 17 より, オリジナルの **KNP** の正解率は **KNP+Manual** や **Manual** を訓練事例とした場合の正解率よりも高い値となっていることがわかる．これは, **KNP+Manual** や **Manual** で作成されたコーパスが, オリジナルの **KNP** の訓練事例に比べ, とても少ないためであると考えられる．一方で, 表 16, 表 17 より, **KNP+Manual** や **Manual** を訓練事例とした場合, 及び, オリジナルの **KNP** において, 適合率は再現率に比べて大きな差がなく, また, 表 18 より, OC と OY の適合率のマイクロ平均において, オリジナルの **KNP** よりも **KNP+Manual** や **Manual** を訓練事例とした場合の方が高い値となっていることがわかる．このことから, 訓練事例が少ないとしても, 適合率はオリジナルの KNP と同等以上になるということがいえる．

表 14　2 人のアノテータによるアノテーション結果を統合したときの正解率のマイクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.72	0.48	0.57
OC	KNP+Manual	**0.87**	0.86	0.87
OC	Manual	0.86	**0.91**	**0.88**
OC	Average	0.87	0.89	0.88
OW	KNP	0.79	0.79	0.79
OW	KNP+Manual	***0.91**	**0.91**	**0.91**
OW	Manual	0.76	0.89	0.82
OW	Average	0.84	0.90	0.87
OY	KNP	0.73	0.57	0.64
OY	KNP+Manual	**0.94**	**0.87**	**0.90**
OY	Manual	0.93	0.86	0.89
OY	Average	0.94	0.87	0.90
PB	KNP	0.75	0.60	0.66
PB	KNP+Manual	0.87	0.82	0.84
PB	Manual	**0.90**	**0.86**	**0.88**
PB	Average	0.89	0.84	0.86
PM	KNP	0.61	58	0.59
PM	KNP+Manual	0.93	**0.94**	0.93
PM	Manual	***0.97**	0.93	**0.95**
PM	Average	0.95	0.94	0.94
PN	KNP	0.88	0.78	0.83
PN	KNP+Manual	***0.93**	**0.90**	**0.92**
PN	Manual	0.89	0.87	0.88
PN	Average	0.91	0.89	0.90

7. まとめ

本稿では, 非専門家の手で高品質なコーパスを作成する手法について調べるため, 固有表現抽出におけるアノテーションを題材として, ふたつの手法について比較を行った. ひとつは既存の固有表現抽出器によるアノテーション結果に対し, 人手で修正を行う手法 (**KNP+Manual**) であり, もうひとつは人手で一からアノテーションを行う手法 (**Manual**) である. この際, アノテーションにかかる時間, タグの一致率, Gold Standard との比較による正解率の各観点から比較を行っている. また, これに加え, 機械学習における訓練事例としての品質を確かめる

表 15　2 人のアノテータによるアノテーション結果を統合したときの正解率のマクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.31	0.26	0.28
OC	KNP+Manual	0.46	0.47	0.47
OC	Manual	**0.49**	**0.51**	**0.50**
OC	Average	0.48	0.49	0.49
OW	KNP	0.77	0.80	0.79
OW	KNP+Manual	**0.91**	0.91	**0.91**
OW	Manual	0.83	**0.91**	0.87
OW	Average	0.87	0.91	0.89
OY	KNP	0.58	0.44	0.50
OY	KNP+Manual	**0.79**	**0.74**	**0.76**
OY	Manual	0.68	0.65	0.67
OY	Average	0.74	0.70	0.72
PB	KNP	0.66	0.46	0.54
PB	KNP+Manual	0.84	0.78	0.81
PB	Manual	**0.94**	**0.86**	**0.90**
PB	Average	0.89	0.82	0.86
PM	KNP	0.60	0.66	0.63
PM	KNP+Manual	0.86	0.93	0.89
PM	Manual	**0.98**	**0.93**	**0.95**
PM	Average	0.92	0.80	0.92
PN	KNP	0.88	0.78	0.82
PN	KNP+Manual	**0.93**	**0.90**	**0.92**
PN	Manual	0.89	0.86	0.88
PN	Average	0.91	0.88	0.90

ため, それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いたアノテーションも行った. これらの実験の結果から, 全ジャンルのマイクロ平均とマクロ平均で比較した場合, **KNP+Manual** は **Manual** よりもアノテーションにかかる時間が少なく, 一致率や正解率についても高い値になることが明らかになった. 一方で, 新聞から離れたジャンルで同様に比較した場合, **Manual** の方が良い結果となることが明らかになった. これらのことから, 新聞に近いジャンルのテキストについては **KNP+Manual** を, そうでないテキストについては **Manual** を採用するのが良いといえる.

表 16　それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の正解率のマイクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.78	0.68	0.73
KNP+Manual	**0.74**	**0.38**	**0.50**
Manual	0.67	0.29	0.40
Average	0.71	0.33	0.45

表 17　それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の正解率のマクロ平均 (全体)

手法	適合率 (精度)	再現率	F 値
KNP	0.47	0.40	0.43
KNP+Manual	**0.40**	**0.24**	**0.30**
Manual	0.31	0.16	0.21
Average	0.36	0.20	0.26

表 18　それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の正解率のマイクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.72	0.48	0.57
OC	KNP+Manual	**0.88**	**0.29**	**0.43**
OC	Manual	0.84	0.20	0.32
OC	Average	0.87	0.24	0.38
OW	KNP	0.79	0.79	0.79
OW	KNP+Manual	***0.74**	***0.53**	**0.62**
OW	Manual	0.55	0.36	0.43
OW	Average	0.65	0.45	0.53
OY	KNP	0.73	0.57	0.64
OY	KNP+Manual	**0.84**	***0.32**	**0.46**
OY	Manual	0.80	0.18	0.30
OY	Average	0.82	0.25	0.38
PB	KNP	0.75	0.60	0.66
PB	KNP+Manual	0.70	**0.31**	**0.43**
PB	Manual	**0.73**	0.28	0.40
PB	Average	0.72	0.29	0.41
PM	KNP	0.61	0.58	0.59
PM	KNP+Manual	**0.55**	**0.19**	**0.29**
PM	Manual	0.52	0.14	0.22
PM	Average	0.54	0.17	0.25
PN	KNP	0.88	0.78	0.83
PN	KNP+Manual	0.76	***0.43**	**0.55**
PN	Manual	**0.78**	0.36	0.49
PN	Average	0.77	0.40	0.52

表 19　それぞれの手法で作成されたコーパスを訓練事例とした KNP を用いた場合の正解率のマクロ平均 (ジャンルごと)

ジャンル	手法	適合率 (精度)	再現率	F 値
OC	KNP	0.31	0.26	0.28
OC	KNP+Manual	**0.24**	**0.16**	**0.19**
OC	Manual	0.17	0.12	0.14
OC	Average	0.21	0.14	0.17
OW	KNP	0.77	0.80	0.79
OW	KNP+Manual	**0.72**	**0.57**	**0.63**
OW	Manual	0.63	0.43	0.51
OW	Average	0.67	0.50	0.57
OY	KNP	0.58	0.44	0.50
OY	KNP+Manual	**0.52**	**0.24**	**0.33**
OY	Manual	0.31	0.09	0.14
OY	Average	0.42	0.17	0.24
PB	KNP	0.66	0.46	0.54
PB	KNP+Manual	0.51	**0.24**	0.32
PB	Manual	**0.65**	0.22	**0.32**
PB	Average	0.58	0.23	0.33
PM	KNP	0.60	0.66	0.63
PM	KNP+Manual	**0.55**	**0.29**	**0.38**
PM	Manual	0.53	0.25	0.34
PM	Average	0.54	0.27	0.36
PN	KNP	0.88	0.78	0.82
PN	KNP+Manual	0.75	**0.44**	**0.55**
PN	Manual	**0.78**	0.37	0.50
PN	Average	0.77	0.40	0.53

謝 辞

本研究は文部科学省科学研究費補助金 [若手 B (No.15K16046)] と富士通研究所の助成により行われました. ここに謹んで御礼申し上げます.

また, KNP についての有益な情報を提供してくださった東京工業大学の笹野遼平先生に, この場を借りて御礼申し上げます.

文 献

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng (2008). "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks." *Proceedings of the conference on empirical methods in natural language processing*, pp. 254–263., Association for Computational Linguistics.

Kanako Komiya, Masaya Suzuki, Tomoya Iwakura, Minoru Sasaki, and Hiroyuki Shinnou (2016). "Comparison of Annotating Methods for Named Entity Corpora." *LAW X*, pp. 59–67.

Bea Alex, Claire Grover, Rongzhou Shen, and Mijail Kabadjov (2010). "Agile corpus annotation in practice: An overview of manual and automatic annotation of CVs." *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 29–37., Association for Computational Linguistics.

Lonneke van der Plas, Tanja Samardžić, and Paola Merlo (2010). "Cross-lingual validity of PropBank in the manual annotation of French." *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 113–117., Association for Computational Linguistics.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). "Building a large annotated corpus of English: The Penn Treebank." *Computational linguistics*, 19:2, pp. 313–330.

橋本泰一・乾孝司・村上浩司 (2008).「拡張固有表現タグ付きコーパスの構築」 情報処理学会研究報告, 自然言語処理研究会報告 (NL-188-17), pp. 113–120.

橋本泰一・中村俊一 (2010).「拡張固有表現タグ付きコーパスの構築-白書, 書籍, Yahoo! 知恵袋コアデータ」 言語処理学会第 16 回年次大会発表論文集, 2010, pp. 916–919.

Kikuo Maekawa (2008). "Balanced Corpus of Contemporary Written Japanese." *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pp. 101–102.

徳永健伸・西川仁・岩倉友哉・湯上伸弘 (2015).「固有表現認識課題におけるアノテータの視線分析」 情報処理学会研究報告自然言語処理, 2015:8, pp. 1–8.

Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, and Yoko Yamakata (2015). "Named Entity Recognizer Trainable from Partially Annotated Data." *Proceedings of the PACLING 2015*, pp. 10–17.

Satoshi Sekine, and Hitoshi Isahara (2000). "IREX: IR and IE Evaluation project in Japanese." *Proceedings of the 2nd International Conference on Language Resources &*

Evaluation.

Ralph Grishman, and Beth Sundheim (1996). “Message Understanding Conference-6: A Brief History..” *COLING* Vol. 96., pp. 466–471.

岩倉友哉 (2015).「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 (NLP2015) ワークショップ：自然言語処理におけるエラー分析

平田亜衣・小町守 (2015).「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」 言語処理学会第 21 回年次大会 (NLP2015) ワークショップ：自然言語処理におけるエラー分析

Masaaki Ichihara, Kanako Komiya, Tomoya Iwakura, and Maiko Yamazaki (2015). “Error Analysis of Named Entity Recognition in BCCWJ.” 言語処理学会第 21 回年次大会 (NLP2015) ワークショップ：自然言語処理におけるエラー分析

Ryohei Sasano, and Sadao Kurohashi (2008). “Japanese Named Entity Recognition Using Structural Natural Language Processing..” *IJCNLP*, pp. 607–612.

Information Retrieval and Extraction Exercise`http://nlp.cs.nyu.edu/irex/NE/df990214.txt` (1999).『ルール、定義』.

Tomoya Iwakura, Ryuichi Tachibana, and Kanako Komiya (2016). “Constructing a Japanese Basic Named Entity Corpus of Various Genres.” *ACL 2016*, pp. 41–46.