日本語 Wikipedia からの属性値抽出タスクにおけるクエリの有効性検証

坂田将樹 [1]　中山功太 [2,3]　竹下昌志 [4]　ジェプカ ラファウ [5]　関根聡 [2]　荒木健治 [5]
[1] 北海道大学工学部　[2] 理化学研究所 AIP　[3] 筑波大学情報生命学術院
[4] 北海道大学大学院情報科学院　[5] 北海道大学大学院情報科学研究院
sakata.masaki.e9@elms.hokudai.ac.jp　{kouta.nakayama, satoshi.sekine}@riken.jp
{takeshita.masashi, rzepka, araki}@ist.hokudai.ac.jp

概要

近年，固有表現抽出や属性値抽出において機械読解に基づく手法が高い精度を達成している．しかし，日本語の属性値抽出タスクにおいては機械読解手法におけるクエリの有効性の調査は未だ無い．そこで本稿では，抽出精度が良いもしくは悪いクエリを探索し，その特徴を調査する．実験では 6 種類のクエリと正答例を含む提案クエリを使用する．結果，属性値のみ，もしくは正答例を含むクエリが最良の抽出精度であった．反対に，抽出対象の記事タイトルを付加すると抽出精度が低下することが明らかになった．また，追加情報によって予測数と誤予測が減少する作用が確認された．ソースコードは https://github.com/Language-Media-Lab/shinra_jp_bert/ にて公開している．

1　はじめに

近年，固有表現抽出タスクと属性値抽出タスクにおいて，機械読解に基づく手法が高い精度を達成している [1, 2, 3, 4]．機械読解とは，システムが文章 (パッセージ) を読解し，与えられた質問 (クエリ) の回答を文章中から抽出する技術である．よって，機械読解モデルへの入力として「クエリ」はタスクを解く上で必要不可欠である．このクエリについて，Li らは [1] 英語を対象とした固有表現抽出タスクにおいて，7 種類のクエリのうち，効果的に固有表現を抽出できるクエリが存在することを明らかにした．

しかしながら，日本語 Wikipedia からの属性値抽出タスクにおいて，効果的に属性値を抽出できるクエリはどのようなものであるのかは不明である．日本語 Wikipedia からの属性値抽出タスクを機械読解として解いた 3 つの既存研究 [2, 3, 4] では，3 つの異なるモデルが提案されているが，それぞれ異なる 1 種類のクエリのみを使用した結果が報告されている．すなわち，既存研究の結果の有効性が提案モデルのアルゴリズムと使用したクエリのいずれに起因するかが不明である．また，英語を対象としたタスクで得られた知見が，日本語を対象としたタスクで同様に有効性を示せるかについては自明ではない．

そのため，本研究では各クエリを用いた際の抽出精度の違いを算出することでクエリの有効性を検証し，抽出精度が良いもしくは悪いクエリの特徴を調査した．実験では，日本語 Wikipedia からの属性値抽出タスクで使用されたクエリと，Li ら [1] が有効性を示したクエリを参考にし，正答例を付加したクエリを使用した (表 1)．

2　関連研究

機械読解モデルとは，与えられた文章とクエリから回答スパンを抽出するモデルである．このモデルは近年，抽出精度の高さが注目され，固有表現抽出と属性値抽出に使用されている．本節では，日本語 Wikipedia からの属性値抽出タスクと英語を対象とした固有表現抽出タスクの概要と，これらのタスクを機械読解に基づく手法で解いた既存研究について概観する．

日本語 Wikipedia からの属性値抽出タスク　このタスクでは，Wikipedia 記事のカテゴリごとに「拡張固有表現」[5] で定義されている属性に対応する値を Wikipedia 記事中から抽出する [6]．例えば，企業名カテゴリに属する記事「シャネル」の記事本文から，属性「本拠地国」に対応する文字列「フランス」を抽出する．このタスクを機械読解の手法

表 1　本研究で使用するクエリ: 1) 属性名: 抽出対象の属性名を付加．2) 疑問詞:「は？」もしくは「ですか？」を付加．3) 5W1H : 属性に対応した「誰・何・どこ・いつ・いくつ」を付加．中山 [3] が作成した 5W1H のリストを使用．5) タイトル: 抽出対象の記事タイトルを付加．6) 正答例: 学習データ内の正答例を付加．

	属性名	疑問詞	5W1H	タイトル	正答例	クエリ例
クエリ 1	✓					本拠地国
クエリ 2	✓	✓				本拠地国は?
クエリ 3	✓	✓	✓			本拠地国はどこですか?
クエリ 4	✓			✓		シャネルの本拠地国
クエリ 5	✓	✓		✓		シャネルの本拠地国は?
クエリ 6	✓	✓	✓	✓		シャネルの本拠地国はどこですか?
正答例を含むクエリ	✓	✓			✓	本拠地国は?例えば日本などです。

で解いた既存研究 [2, 3, 4] では，深層学習モデルとして，DrQA[7]，BERT[8]，RoBERTa[9] が使用された．機械読解手法を用いた既存研究のうち，BERT と中間タスクを用いた機械読解システム [4] は日本語 Wikipedia からの属性値抽出タスクにおいて最も高い抽出精度であった．

英語を対象とした固有表現抽出　Li ら [1] は英語を対象とした固有表現抽出タスク [10, 11, 12, 13, 14] において，BERT を用いた機械読解手法で他の手法よりも良い精度を達成した．この研究で使用されたデータセットのうち，English OntoNotes5.0[14] のみを対象に 7 種類のクエリを使用して，抽出精度の比較を行っている．その結果，使用するクエリによって F 値が変化した．F 値が最も高かったクエリは「Find organizations including companies, agencies and institutions」のようなアノテーションガイドラインで使用されるような文章であった．

3　既存研究の未調査領域

前章の既存研究では以下 2 点が未調査である．

3.1　日本語クエリが抽出精度に与える影響が不明

先行研究 [2, 3, 4] では機械読解としてタスクを解く際にそれぞれ異なる 1 種類のクエリのみを付与しているが，この際与えるクエリの影響については分析していない．Li ら [1] は英語固有表現抽出タスクにおいてクエリの影響を述べている．クエリの種別においては本来のモデル性能を阻害している可能性があるため，本研究では日本語 Wikipedia からの属性値抽出タスクにおいてクエリが与える影響について調査を行った．

使用するクエリの有効性を調査した研究は英語を対象とした研究 [1] のみであり，筆者らの知る限り日本語を対象とした研究は未だない．

3.2　使用クエリの有効性は言語横断的にいえるのかが不明

Li ら [1] は「Find organizations including companies, agencies and institutions」における “companies”, “agencies”, “institutions” の部分が F 値向上の理由だと考察している．その原因として，クエリ内に抽出対象のカテゴリについて記述されていることを挙げている．以上が日本語 Wikipedia からの属性値抽出タスクにも一般化できるかは不明である．

4　実験に用いるクエリの種類

3 章の未調査領域を踏まえ，本稿では，既存研究で使用されたクエリと抽出精度の関係性を統一的に調査する．また「抽出精度が良いクエリは正答例を含む」という仮説立て検証する．

4.1　既存研究のクエリの性能調査

調査する 6 種類のクエリを表 1 に示す (クエリ 1 からクエリ 6 が該当)．表 1 の各列は，クエリの構成要素を表している．例えばクエリ 6 の場合，記事タイトル「シャネル」の属性「本拠地国」を抽出する際「シャネルの本拠地国はどこですか？」となる．

クエリ 1 からクエリ 6 の性能を調べるために，それぞれのクエリで学習と評価を行った 6 種類のモデルの出力を調査する．このうち既存研究 [2, 3, 4] で使用されたものは**クエリ 1**[4]，**クエリ 5**[2]，**クエリ 6**[3] である．クエリ 2, 3, 4 は既存研究で使用されていないが，「記事タイトルの有無」「疑問詞の有無」「5W1H の有無」が抽出精度に与える影響をそれぞれ分離して分析するために追加した．

4.2　提案クエリ：正答例を追加

Li ら [1] は「Find organizations including companies, agencies and institutions」のうち，“companies”, “agencies”, “institutions” の部分が抽出精度が良いクエリの特徴だと主張している．この “companies”, “agen-

表 2 実験結果：各クエリの F 値．太字は各カテゴリの中で F 値の最大値であり，下線は F 値の最小値を表している．

カテゴリ	クエリ 1	クエリ 2	クエリ 3	クエリ 4	クエリ 5	クエリ 6	正答例× 1	正答例× 5	正答例× 10
市区町村名	63.56	63.7	63.23	-	62.54	61.44	63.0	**64.0**	62.69
企業名	57.23	57.38	57.19	55.84	54.6	54.63	57.72	**58.03**	**58.03**
人名	**75.12**	73.24	73.43	-	70.24	69.09	74.69	75.01	74.59

cies”, “institutions” はいわば抽出クラスと抽出対象の中間概念を示しているため，本タスクでも同様に，この中間概念をクエリに含ませることが考えられる．しかし，拡張固有表現 [15] には全ての属性値と属性名に対する中間概念は定義されていないため，適用できない．そこで本稿では，中間概念の代わりに正答例をクエリに付加する．具体的には「創業国は？例えば、アメリカ合衆国、日本、シンガポール、イングランド、イギリス、などです。」のように正答例を結合する．付加する正答例は，学習データからランダムサンプル[1)]により取得する．取得数は 1, 5, 10 とした．[2)]

5 実験

本稿の実験設定は石井 [4] が行った実験に準ずる．

5.1 学習用データセット

森羅 2019-JP[6] において配布されたアノテーションデータセット[3)]を使用する．本データセットは HTML 形式の Wikipedia 記事と属性値のアノテーションからなる．HTML タグを除去した記事も用意されているが，本実験では使用しない．

機械読解データセット構築 HTML 形式の Wikipedia 記事を<p>タグを元にパラグラフ単位へ分割し，$P = \{p_1, p_2, \ldots, p_l\}$ を得る．この際，パラグラフ i に含まれる属性値を $A_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,t}\}$ とする．また，機械読解として扱うため，クエリ q を生成する．

使用クエリ 4 章で定義した合計 9 種類のクエリを使用する．

データ分割 データセットは学習データ，検証データ，テストデータをそれぞれ 85%，5%，10%の割合で分割する．

5.2 BERT を用いた機械読解モデル

機械読解モデルへの入力は (p, q) であり，出力は各単語の BIO タグである．クエリ q の単語トークン列を $q = \{q_1, q_2, \ldots, q_m\}$, パラグラフ p_i の単語トークン列を $x_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$ とすると BERT への入力系列は以下の通りである．

$$\{[\mathrm{CLS}], q_1, q_2, \ldots, q_m, [\mathrm{SEP}], x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$$

[CLS], [SEP] は BERT の特殊トークンである．BERT の最終出力のうち，x_i に対応する出力を $\boldsymbol{H}_i = \{\boldsymbol{h}_{i,1}, \boldsymbol{h}_{i,2}, \ldots, \boldsymbol{h}_{i,n}\}$ とする．$\boldsymbol{H}_i$ を出力層に入力し各 BIO タグの予測スコア $S_i = \{\boldsymbol{s}_{i,1}, \boldsymbol{s}_{i,2}, \ldots, \boldsymbol{s}_{i,n}\}$ を得る．i 番目のトークンに対する最終的な予測 $\hat{y}_{i,j} \in \{\mathrm{B}, \mathrm{I}, \mathrm{O}\}$ は $\hat{y}_{i,j} = \mathrm{argmax}_k\, \boldsymbol{s}_{i,j,k}$ により求まる．学習は正答ラベル $Y_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,n}\}$ と予測結果 $\hat{Y}_i = \{\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,n}\}$ との交差エントロピー誤差を最小化することで行う．

5.3 実験設定

モデルは「NICT BERT 日本語 Pretrained モデル BPE あり」[4)]を使用する．ハイパーパラメータは石井 [4] の研究と同様にバッチサイズは 32，学習率は 2e-05，トークンの最大長は 384，ストライドは 128，エポック数は 10 回とする．テスト時には，検証データで F 値が最も高いモデルを用いた．入力文字の前処理では，MeCab-Juman 辞書 [16] を用いて形態素に分割した後，subword-nmt[17] で生成された語彙を用いてサブワード化した．

5.4 評価指標

森羅 2019-JP により提供される評価用スクリプト[5)]によって算出される属性ごとの F 値のマイクロ平均を，各カテゴリの評価指標とする．

6 結果と考察

各クエリの実験結果は表 2 の通りである．まず，各クエリによる F 値の違いを概観する．

クエリ 4，5，6 の結果より，クエリに記事タイトルを追加すると一貫して抽出精度が悪化した．よって，クエリに記事タイトルを使用している 2 つの既存研究 [2, 3] は，使用クエリが原因で抽出精度が減少していた可能性がある．記事タイトルを追加する

1) シード値は 42 とした．
2) 実際に付加した正答例は全て https://github.com/Language-Media-Lab/shinra_jp_bert/ にて公開している．
3) http://shinra-project.info/download/
4) https://alaginrc.nict.go.jp/nict-bert/index.html
5) https://github.com/k141303/shinra_jp_scorer

表 3　正答例の有無による予測数・TP 数・FP 数の違い．TP 差は増加すれば好ましい．FP 差は減少すれば好ましい．値の太字は改善した結果を表しており，下線は悪化を表している．

カテゴリ	正答例× 1 – クエリ 2			正答例× 5 – クエリ 2			正答例× 10 – クエリ 2		
	予測数の差	TP 差↑	FP 差↓	予測数の差	TP 差↑	FP 差↓	予測数の差	TP 差↑	FP 差↓
市区町村名	-212	-130	**-82**	-545	-147	**-398**	-156	-140	**-16**
企業名	-403	-88	**-315**	-253	**26**	**-279**	-41	**42**	**-83**
人名	-986	**11**	**-997**	-607	**237**	**-844**	-1520	-214	**-1306**

表 4　記事タイトルの有無による予測数・TP 数・FP 数の違い．TP 差は増加すれば好ましい．FP 差は減少すれば好ましい．値の太字は改善した結果を表しており，下線は悪化を表している．

カテゴリ	クエリ 5 – クエリ 2			クエリ 6 – クエリ 3		
	予測数の差	TP 差↑	FP の差↓	予測数の差	TP 差↑	FP 差↓
市区町村名	-1139	-461	**-678**	-376	-270	**-106**
企業名	-1458	-627	**-831**	-2342	-858	**-1484**
人名	-2069	-1512	**-557**	-992	-1435	443

表 5　正答例× 5 を付加した際に F 値が 5%以上変化する属性．

カテゴリ	5%以上増加	5%以上低下
市区町村名	人口 読み 温泉・鉱泉 人口データの年	友好市区町村
企業名	創業地 創業者	売上高（連結） 売上高（連結）データの年
人名	国籍 没年月日	没地 死因 師匠

と抽出精度が悪化する原因として，各タイトルと属性値のペアに対してモデルが過学習していると考えられる．また，Wikipedia 本文中には記事タイトルは出てこない場合が多いため，タイトルがタスクを解く上でヒントになっていないと考えられる．

そして，クエリ 1 と 2，クエリ 2 と 3 の結果より，5W1H と疑問詞の有無によって精度はほとんど変化しないことが明らかになった．

クエリに正答例を追加した場合，わずかながら最良の抽出精度となっているカテゴリもあるが，3%以上の増加は認められなかった．

予測数・TP 数・FP 数　次に，予測数・TP 数[6]・FP 数[7]について正答例の有無，記事タイトルの有無によって生じる差をカウントした (表 3，表 4)．結果，記事タイトルや正答例をクエリに含ませた場合において，予測数と FP 数が一貫して減少した．クエリの変化によってこのような作用が一貫して生じることは筆者らの知る限り報告されていない．記事タイトルの有無と正答例の有無を比較すると，TP 数の減少が大きく異なり，正答例有りの方が TP の減少を抑えることができている．正答例を入れることで起こる TP 数の減少を抑えた FP 数の減少は，今後適合率の高いモデルを開発する上で活用可能であると考える．付加する正答例の数に着目すると，予測数・TP 数・FP 数の増減とは比例していない．逆に正答例を 1 つだけ追加した場合でも，予測数と FP 数の減少が確認される．よって，予測数と FP 数の減少には，正答例の数より，追加情報の有無もしくは質によって決定づけられていることが考察される．したがって，予測数と FP 数が減少する理由は，クエリへの「追加情報」が関係していると考えられる．

正答例を入れることで抽出精度が変化する属性　表 5 に正答例の有無によって，F 値が 5%以上変化した属性を示した．適合率と再現率の変化については表は付録 A を参照されたい．一見カテゴリを横断して強く共通している属性は無いように見えるが，「〇〇年」「〇〇人」「〇〇温泉」等の日付表現，人数，温泉などのスパンの最後の文字列がおよそ共通なものが向上している．より詳細な共通点や，精度が増加した理由の分析は今後の課題とする．

7　おわりに

本研究では日本語 Wikipedia からの属性値抽出タスクで使用されたクエリの調査と正答例を含むクエリの有効性検証を行った．その結果，記事タイトルを含むクエリは抽出精度が悪化し，正答例を含むクエリはわずかながら最良の抽出精度となった．そして，記事タイトルや正答例を含むことで，予測数と誤予測が減少する作用が確認された．この作用はクエリに情報を追加したことが原因と考えられるが，より詳しい分析が必要である．また，正答例を付加することで抽出精度が大きく増加・減少した属性があることを明らかにしたが，それらの共通点や変化した理由についても詳しい分析が必要である．

6）　True Positive : ラベルとスパンが完全一致している予測
7）　False Positive : 誤予測

謝辞

本研究は JSPS 科研費 JP20269633 の助成を受けたものです.

参考文献

[1] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. **CoRR**, Vol. abs/1910.11476, , 2019.

[2] 石井愛. 機械読解による wikipedia からの情報抽出. 言語処理学会第 25 回年次大会 (NLP2019), 2019.

[3] 中山功太. RoBERTa を用いた QA 式属性値抽出システム. 森羅 2019-JP 最終報告会資料, 2019.

[4] 石井愛. Stilts を適用した機械読解による wikipedia 記事の構造化. 言語処理学会第 27 回年次大会 (NLP2021), 2021.

[5] Satoshi Sekine. Extended named entity ontology with attribute information. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[6] 小林暁雄, 中山功太, 安藤まや, 関根聡. Wikipedia 構造化プロジェクト「森羅 2019-JP」. 言語処理学会第 26 回年次大会 (NLP2020), 2020.

[7] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. **CoRR**, Vol. abs/1704.00051, , 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. **CoRR**, Vol. abs/1810.04805, , 2018.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. **CoRR**, Vol. abs/1907.11692, , 2019.

[10] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In **Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)**, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[11] Stephanie Strassel Julie Medero Christopher Walker and Kazuaki Maeda. The automatic content extraction (ACE) program – tasks, data, and evaluation. Philadelphia 57, 2006. Linguistic Data Consortium.

[12] Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. The genia corpus: An annotated research abstract corpus in molecular biology domain. In **Proceedings of the Second International Conference on Human Language Technology Research**, HLT '02, p. 82–86, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[13] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In **Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003**, pp. 142–147, 2003.

[14] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. In **Proceedings of the Seventeenth Conference on Computational Natural Language Learning**, pp. 143–152, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[15] 拡張固有表現階層定義 version 8.0.0. http://liat-aip.sakura.ne.jp/ene/ene8/definition_jp/.

[16] T KUDO. Yet another part-of-speech and morphological analyzer. https://taku910.github.io/mecab/.

[17] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

A　正答例を入れることで適合率・再現率が変化する属性

正答例を 5 つ付加することで適合率・再現率が変化する属性をカテゴリごとに示した (表 6，表 7，表 8)．適合率と再現率の両方の変化が 1%未満だった属性は省いている．F 値が 5%以上増加した属性は青字，5%以上減少した属性は赤字で示した．

表 6　【企業名】正答例× 5 を入れることで適合率・再現率が変化する属性．

		再現率	
		増加 or 0	減少
適合率	増加 or 0	**創業地** **創業者** 設立年 過去の社名 事業内容 取扱商品 子会社・合弁会社 従業員数（単体）データの年 正式名称 売上高データの年 従業員数（単体） 業界内地位・規模 資本金 コーポレートスローガン	別名
	減少	本拠地国 種類 起源 売上高（単体） 創業時の事業 読み 資本金データの年	代表者 本拠地 業界 **売上高（連結）** **売上高（連結）データの年**

表 7　【市区町村名】正答例× 5 を入れることで適合率・再現率が変化する属性．

		再現率	
		増加 or 0	減少
適合率	増加 or 0	**人口** **読み** **温泉・鉱泉** 成立年 人口密度 所在地	**人口データの年** 別名 国 国内位置 座標・緯度 座標・経度 鉄道会社 面積 首長 観光地 恒例行事
	減少		産業 合併市区町村 旧称 **友好市区町村**

表 8　【人名】正答例× 5 を入れることで適合率・再現率が変化する属性．

		再現率	
		増加 or 0	減少
適合率	増加 or 0	**国籍** **没年月日** 両親 作品 所属組織 読み 参加イベント	学歴 家族 居住地 別名 受賞歴 地位職業 生年月日
	減少	異表記 時代 本名	生誕地 **没地** **死因** **師匠**