機械読解によるWikipediaからの情報抽出

石井 愛
日本ユニシス株式会社 総合技術研究所
ai.ishii@unisys.co.jp

1 はじめに

文章を読み質問に答える機械読解が近年注目を集めており，次々にデータセットが公開され，さまざまな手法が提案されている．提案されている手法は深層学習を用いる手法が中心であり，日本語で作成された機械読解のデータセットにおいても深層学習を用いる手法を適用した報告がある [3][2].

Wikipedia 構造化プロジェクト「森羅 2018」[1] は，Wikipedia に書かれている世界知識を計算機が利用可能な形に構造化することを目的として，タスク参加者によりリソースを作成していくプロジェクトである．タスクは，Wikipediaの各ページに対し、カテゴリごとに定義された属性の値を抽出するものである．例えば、空港名のページであれば、所在地、開港年、滑走路数などの属性値を Wikipediaの本文およびインフォボックスから抽出する。Wikipediaの本文をドキュメント，属性名を質問，属性値を回答と考えると，SQuAD[4] のような文章から回答となる範囲を選択するタイプの機械読解タスクとみなすことができる．

本稿では，Wikipedia 属性値の抽出タスクを解く手法の一つとして，機械読解タスクとみなして解く手法を提案する．評価実験により，機械読解が Wikipedia 属性値の抽出タスクに有効であること，および機械読解形式へのデータセットの変換方法が精度の向上に寄与することを示す．

2 問題

対象とする問題は，人名，企業名，市区町村名，空港名，化合物名の 5 カテゴリの Wikipedia ページから定義された属性の値を抽出するタスクである [1]．システム開発用に，各カテゴリ 600 ページ分のトレーニングデータが与えられる．

3 提案手法

提案手法は Wikipedia 属性値の抽出を，対象ページをドキュメント d，抽出する属性名を質問 q，属性値を回答 a とみなし，機械読解タスクとして解くものである．まず森羅 2018 データセットを機械読解形式に変換し，そのデータセットを用いて機械読解により学習と予測を行い，最後に予測結果をルールで補正する．

機械読解形式へのデータセットの変換

森羅 2018 のデータセットを，SQuAD 形式に変換する．ドキュメント d は Wikipedia 本文の始点にインフォボックスから生成した自然文を挿入したものとする．インフォボックスから項目名と値が抽出[1]できた場合，

[インフォボックス項目名] は [項目値]。

という自然文に変換する．両方を抽出できなかった場合は，”。”で区切る．Wikipedia ページ内では主語が省略されることが多く，上記のような形の文が多く存在する．質問 q は，

[Wikipedia 項目名] の [属性名] は？

という自然文に変換する．ドキュメント d における回答 a の位置情報は，a と文字列が完全一致した箇所すべてを設定する．質問 q，回答 a，ドキュメント d の例を表 1 に示す．

機械読解モデル

機械読解モデルは，SQuAD に対して高い精度を示す機械読解モデルの 一つである DrQA[5] の Document Reader をベースとし，日本語への対応および，複数回答への対応を行ったものである．Document Reader は，学習済みの単語分散表現に加え，品詞などの単語情報や，質問中の単語との一致などの特徴量を入力と

[1] Wikipedia ページの HTML データに対し，Xpath を用いて抽出した

表 1: 機械読解形式への変換例

質問 q	回答 a (テキスト，開始位置)	ドキュメント d
小松飛行場の別名は？	Komatsu Airbase, 12 Komatsu Airbase, 721 (中略) 小松空港, 6 小松空港, 787 (中略) FAC4017 小松補助飛行場, 2045	**インフォボックス:** 小松飛行場（小松空港)、Komatsu Airbase (Komatsu Airport)。ターミナルビル。IATA: KMQ - ICAO: RJNK。概要。国・地域は、日本。所在地は、石川県小松市。母都市は、(中略) **Wikipedia 本文:** 小松飛行場 (こまつひこうじょう) は、石川県小松市にある共用飛行場である。 防衛省が管理しており、航空自衛隊小松基地 (英: JASDF Komatsu Airbase) と民間航空 (民航) が滑走路を共用する飛行場で、特に後者においてはターミナルビルなどの施設の通称として小松空港 (こまつくうこう、英: Komatsu Airport) と呼ばれている。(中略)

して双方向 LSTM にて学習し，回答の start と end の位置を予測する機械読解モデルであり，実装が公開されている[2]．この実装に対し，日本語に対応するため MeCab[6] を用いて単語分割する機能を追加した．また，Wikipedia 属性値は，一つの属性の属性値が複数となる場合や，属性値が対象ページに存在しない場合がある．そのため，鈴木ら [2] を参考に，a の開始位置と終了位置を予測する最終層を，以下の計算を行う層に変更し，d に含まれる各単語 d_t が a の一部である確率 p_t を出力する．

$$p_t = \text{sigmoid}\left(\mathbf{q}\mathbf{W}\mathbf{p}_t\right)$$

ここで，$\mathbf{q}\mathbf{W}\mathbf{p}_t$ は，DrQA の元論文 [5] における，質問 q の分散表現ベクトルと単語 d_t の分散表現と特徴量を結合したベクトル $\tilde{p}_t$ について，それぞれ双方向 LSTM にてエンコードした $\mathbf{q}$ と $\mathbf{p}_t$ の類似性を算出した行列の t 列目である．損失関数としては，2 値の交差エントロピーを用いる．

$$L(\theta) = -\frac{1}{N}\sum_{1}^{N}\sum_{t=1}^{T}\left(y_t \log p_t + (1-y_t)\log\left(1-p_t\right)\right)$$

ここで，y_t は，単語 d_t が a の一部であれば 1，そうでなければ 0 である教師信号である．T は d の長さ（単語数），N はミニバッチサイズである．出力する答えとしては，$p_t > \alpha$ となる単語 d_t の系列のうち，p_t の平均が最も大きいものから上位 k 件 $(k = 10)$ を採用する．属性値が存在しない場合は，教師信号 y_t を全て 0 にしてモデルを学習する．

予測結果を補正するルール

予測結果に対し，以下のルールを適用し回答の絞り込みを行った．

- 答えが一つになりやすい属性は，二つ目以降 p_tの平均スコア $> \beta_{top1}$ の場合のみ回答に含める
- その他の属性は，p_tの平均スコア $> \beta_{min}$ の場合のみ回答に含める

[2]https://github.com/facebookresearch/DrQA

表 2: データ数

	ドキュメント d	質問 q	正解 a(異なり数)
学習データ	2545	61071	60427
開発データ	150	3600	4179
テストデータ	300	7200	7094

4 実験

データセット

森羅 2018 においてシステム開発用に配布された 5 カテゴリ計 2995 ドキュメントのデータを SQuAD 形式に変換し，表 2 のように学習データ，開発データ，テストデータに分割した．

評価指標

モデルの評価には，出力された回答と回答 a の適合率と再現率から求められる F 値のカテゴリごとのマイクロ平均を用いた．

実験対象モデル

以下の二つのモデルを実験対象とした．

- DrQA:Document Reader と同様のモデル．回答が存在しない場合に対応するため，ドキュメント d の始点に制御文字を挿入し，解答がない場合はその文字を回答として設定して学習した．
- DrQA-Label:Document Reader の最終層を変更し，複数回答に対応したモデル．

モデル詳細設定

単語ベクトルは，fastText[7] を利用し事前に学習した 300 次元の分散表現を用いた．学習には日本語

版 Wikipedia 本文全文を Mecab により単語分割したデータを用いた．

各モデルはエポック数 30，ミニバッチサイズ 4 で各モデルの学習を行った．テストデータに適用するモデルとして，全エポックのうち，最も開発データに対する F 値 が高かったエポックのモデルを採用した．各閾値 α,β_{top1},β_{min} はそれぞれ，0.4，0.8，0.6 に設定した．

5 評価結果

機械読解手法および複数回答対応の有効性

各モデルのシステム開発用データによる実験結果および，[1] より森羅 2018 のテストデータによる結果を表 3 に示す．森羅 2018 のテストデータによる結果は，複数回答対応を行っていない DrQA の結果である[3]．DrQA の結果は，他の参加チームの結果と比較して，僅差ではあるが比較的良い結果を示している．システム開発用データにおける結果においては，DrQA-Label が DrQA を各カテゴリで約 7%から 21%上回った．表 5 に DrQA および DrQA-Label の各カテゴリにおける適合率，再現率，F 値を示す．DrQA-Label は複数回答対応を行ったことにより，再現率が改善し，F 値の改善につながっていることがわかる．

Document Reader 特徴量および実験設定の評価

Document Reader で使用されている特徴量および，いくつかの実験設定の有効性について調査した．f_{token} はドキュメント d の各トークンの品詞情報および単語頻度情報，$f_{aligned}$ は質問 q とドキュメント d の単語ベクトル類似度で重みづけをした単語ベクトル，f_{EM} はドキュメント d の各トークンが質問 q 内のトークンと一致するかどうかという特徴量である．No title は質問 q にタイトルを含めないケース，カテゴリ別学習は，カテゴリごとのデータで別々にモデルの学習を行うケースである．また，No infobox（学習) はインフォボックスから抽出したデータを学習時に使用しないケースであり，No infobox（予測) は予測時に本文のみから回答抽出を行い，本文のみに含まれる属性値のみで評価を行うケースである．

これらの特徴量および実験設定を一つずつ無効化した結果を表 6 に示す．調査結果から，No infobox（予測) と No f_{EM} が特に精度の差分が大きく，有効であることが示された．また，その他の特徴量および実験設定も少しずつ効いているという結果を得た．

No infobox（学習) については，インフォボックスから生成した自然文を学習データに含めないほうが，モデルが Wikipedia 本文から回答を抽出する能力を得られるのではないかという仮説から調査した．しかし仮説に反し，インフォボックスから生成した自然文を含むデータで学習することにより，Wikipedia 本文からの回答抽出の精度も向上するという結果を得た．この結果は，Wikipedia の本文にもインフォボックスから生成した自然文のような形の文が含まれることや，回答データ件数が増えることが影響していると考える．

Document Reader の特徴量については，$f_{aligned}$ と f_{EM}，および両方を無効化する実験において，表 4 に著しく精度が低下した属性を示す．各特徴量が互いに影響し，特徴量の組み合わせにより影響を受ける属性が異なることがわかった．表内の属性名の肩に，記事冒頭に出現しやすい属性は (a)，インフォボックスに出現しやすい属性は (i)，両方に出現しやすい属性は (ia) を示した．記事冒頭に出現しやすい属性については，ドキュメント d 内で Wikipedia 記事が始まる位置の特定に，質問 q 内の Wikipedia 項目名との一致をみる特徴量が効いている可能性がある．また，企業名は創業者の名字が含まれていることや，人名の両親は名字が同じことがあり，Wikipedia 項目名内の単語が含まれる可能性が高く，そのため f_{EM} が効いていると考えられる．

6 おわりに

本研究では，Wikipedia 属性値の抽出タスクにおいて，機械読解とみなして解く手法は有効であり，複数回答対応によりさらに精度が改善されることを示した．また，インフォボックスから生成した自然文をドキュメントに追加することで，Wikipedia 本文からの回答抽出の精度も向上することを示した．この結果から，インフォボックスを自然文に変換する部分にバリエーションを持たせることにより，少し複雑な文からも属性値を抽出できるようになる可能性があると考える．そのような性能評価は今後の課題である．

参考文献

[1] 関根聡, 小林暁雄, 安藤まや. Wikipedia 構造化プロジェクト「森羅 2018」. 言語処理学会第 25 回

[3] DrQA-Label の複数回答対応は森羅 2018 の結果提出日の後に行ったため．また，森羅 2018 に提出した人名および市区町村名のデータは期日後に提出したものであり，参考値として公開されている．

表 3: システム開発データおよびテストデータによる実験結果

	システム開発用データ					森羅 2018 テストデータ				
手法	人名	企業名	市区町村名	空港名	化合物名	人名	企業名	市区町村名	空港名	化合物名
人手作成パターン	-	-	-	-	-	20	41	28	72	-
深層学習	-	-	-	-	-	36	38	46	71	46
DrQA	46.29	57.91	50.71	72.48	42.26	44	53	42	67	47
DrQA-Label	61.89	64.89	68.55	84.59	63.32	-	-	-	-	-

表 4: 精度が低下した属性

	NO $f_{aligned}$	No $f_{aligned}$ and f_{EM}	No f_{EM}
人名	時代$^{(a)}$	両親，師匠，時代$^{(a)}$	両親，時代$^{(a)}$，死因，没年月日$^{(ia)}$
企業名	従業員数（連結）データの年$^{(i)}$	従業員数（連結）データの年$^{(i)}$， 業界内地位・規模， 社名使用開始年	業界内地位・規模，創業者， 社名使用開始年
化合物名	示性式$^{(a)}$	種類$^{(a)}$，用途	種類$^{(a)}$，原材料

表 5: DrQA および DrQA-Label の比較

	DrQA			DrQA-Label		
カテゴリ	適合率	再現率	F 値	適合率	再現率	F 値
人名	81.85	32.27	46.29	69.12	56.04	61.89
企業名	82.02	44.76	57.91	74.41	57.53	64.89
市区町村名	81.50	36.80	50.71	77.72	61.31	68.55
空港名	92.23	59.70	72.48	87.17	82.17	84.59
化合物名	73.13	29.72	42.26	71.18	57.02	63.32

表 6: 特徴量・実験設定評価（F 値平均）

	テストデータ（回答 7094 件）		No infobox（予測）（回答 3105 件）	
Full	68.23		46.39	
No f_{token}	67.7	-0.53		
No title	66.88	-1.35		
No $f_{aligned}$	66.03	-2.20		
No $f_{aligned}$ and f_{EM}	65.27	-2.96		
No f_{EM}	64.99	-3.24		
カテゴリ別学習	67.62	-0.61		
No infobox（学習）			43.00	-3.39

年次大会, March 2019.

[2] 鈴木正敏, 松田耕史, 岡崎直観, 乾健太郎. 読解による解答可能性を付与した質問応答データセットの構築. 言語処理学会第 24 回年次大会 (NLP2018), March 2018.

[3] 西田京介, 斉藤いつみ, 大塚淳史, 浅野久子, 富田準二. 情報検索とのマルチタスク学習による大規模機械読解. 言語処理学会第 24 回年次大会 (NLP2018), March 2018.

[4] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad:100,000+ questions for machine comprehension of text. In EMNLP, pp. 23832392, 2016.

[5] D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading wikipedia to answer open-domain questions. In ACL, pp. 18701879, 2017.

[6] 工藤拓. MeCab : Yet Another Part-of-Speech and Morphological Analyzer.

[7] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. TACL, 5:135146, 2017.