Wikipedia を用いた日本語の固有表現抽出のデータセットの構築

近江崇宏
ストックマーク株式会社
takahiro.omi@stockmark.co.jp

1 はじめに

固有表現抽出（認識）は，人名・組織名といった固有名詞や日付，数値表現を抽出する自然言語処理の基本的な技術である．固有表現抽出は文章の構造化や，人名などのプライバシーに関わる部分のマスキングなどの応用が考えられる。

固有表現抽出器の作成には，コーパスに対して固有表現を付与したデータセットが必要になる．日本語のデータセットに関しては、広く公開されているものとして（無料であり、利用に登録などが必要のないもの），京都大学ウェブ文書リードコーパス[1] や，UD_Japanese-GSD データセット[2]に対して固有表現情報を付与したもの[3]などがある．
我々は最近新たに Wikipedia を用いて日本語の固有表現抽出のためのデータセットを構築して、公開を行った[4]．本論文の主目的は，まずこのデータセットについての詳細な内容を記述することである．また，このデータセットから BERT[5]をファインチューニングした固有表現抽出器を作成し，その性能の評価も行う．

2 データセット

2.1 固有表現の種類

本データセットでは，応用上重要な固有名詞のみを扱った．日付や数値表現については，今後取り入れていくことを検討している．データセットで扱われている固有表現のカテゴリーは「人名」,「法人名」,「政治的組織名」,「その他の組織名」,「地名」,「施設名」,「製品名」,「イベント名」の 8 種類であり，その概要は表 1 にまとめられている．それぞれのカテゴリーに対応する，関根の拡張固有表現階層[6]で定義されている固有表現の種類についても，表 1 にまとめられている．

2.2 データセットの作成

本データセットは日本語版の Wikipedia を用いて作成された．まず，各記事から Wikiextractor[7]を用いて本文を抽出し，本文を文単位に分割し，前処理を行った．前処理の具体的な内容は，文字列の正規化(NFKC)，括弧の削除である．その後，効率的にアノテーションを行うため，ストックマーク社作成の固有表現抽出器（非公開）を用いて，固有表現抽出を行い，何らかの固有表現が含まれていると判定された文を選び出し，人の手によるアノテーションを行った．また固有表現の各カテゴリーで固有表現が 1000 以上含まれるような調整も行った．

最終的には，4859 文に対してアノテーションを行い，その中に含まれる各カテゴリーの固有表現数は表 1 にまとめられている．また，負例として，固有表現が含まれていないデータも 484 文加えたため，データセットに含まれる文の合計は 5343 である．

2.3 データセットの公開

データセットはまず，バージョン 1 として 2020/12/15 に[4]において公開された．その後，データの修正及び追加を行い，バージョン 2 が作成された．今回の論文はこのバージョン 2 について記述する．バージョン 2 も同一のレポジトリで公開される予定である．

3 BERT を用いた固有表現抽出器の性能評価

この章では，本データセットにより，どの程度の性能の固有表現抽出器が作成できるかを定量的に評価する．そのために，BERT[5]を用いて実験を行なった．ここでは試行毎にデータセットからランダムに選び出された 8 割のデータを用いて BERT をファインチューニングし，残りの 2 割のデータをテストデータとして用いて，ファインチューニングされた BERT の性能を評価した．表 2 は適合率，再現率，F

カテゴリー	固有表現数	拡張固有表現との対応
人名	2980	人名
法人名	2485	法人名（大学を含む）
政治的組織名	1180	政治的組織名、国際組織名
その他の組織名	1051	公演組織名、競技組織名、組織名_その他
地名	2157	地名
施設名	1108	施設名
製品名	1009	芸術作品名、出版物名、規則名、乗り物名、キャラクター名、便名、賞名、勲章名、製品名_その他
イベント名	1215	イベント名

表 1：固有表現のカテゴリー

カテゴリー	適合率	再現率	F 値
人名	0.95	0.95	0.95
法人名	0.86	0.90	0.88
政治的組織名	0.76	0.85	0.8
その他の組織名	0.83	0.79	0.81
地名	0.87	0.88	0.87
施設名	0.78	0.84	0.81
製品名	0.71	0.76	0.73
イベント名	0.82	0.87	0.84
全体	0.84	0.88	0.86

表 2：BERT を用いた固有表現抽出の性能評価

値を固有表現のカテゴリー毎と全体で調べた結果である(数値は 10 回の試行の平均値である)．カテゴリー全体での F 値は 86%程度であった．基本的にはデータセットに含まれる固有表現の数が多いカテゴリーほど F 値が高い傾向がある．概ね 2000 以上の固有表現が含まれるカテゴリー（人名，法人名，地名）では F 値が約 90%であった．それ以外の，約 1000 程度の固有表現が含まれるカテゴリーでは概ね，F 値は約 80%程度であったが，製品名のカテゴリーだけは F 値が 73%であった．各カテゴリーの固有表現抽出の難易度により F 値はばらつくが，データセットに固有表現が約 1000 含まれているカテゴリーでは F 値は 80%程度，2000 以上含まれているカテゴリーでは F 値が 90%程度になると，この結果からは類推される．これは目的に応じて，どの程度の規模のデータセットを作れば良いかということに一つの目安を与えると考えられる．

4 まとめ

本論文では Wikipedia を用いた日本語の固有表現抽出のためのデータセットについて詳細な記述を与えるとともに，本データセットを用いた時に，どの程度の性能の固有表現抽出器が作成できるかについても，BERT を用いて評価を行った．

参考文献

1. 萩行正嗣，河原大輔，黒橋禎夫．多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析，自然言語処理，Vol.21，No.2，pp.213-248，2014．http://nlp.ist.i.kyoto-u.ac.jp/?KWDLC

2. 浅原正幸, 金山博, 宮尾祐介, 田中貴秋, 大村舞, 村脇有吾, 松本裕治. Universal Dependencies 日本語コーパス, 自然言語処理, Vol 26, No. 1, pp. 3-36, 2019.
https://universaldependencies.org/treebanks/ja_gsd/index.html
3. 松田寛, 若狭絢, 山下華代, 大村舞, 浅原正幸. UD Japanese GSD の再整備と固有表現情報付与, 言語処理学会第 26 回年次大会発表論文集
https://github.com/megagonlabs/UD_Japanese-GSD/releases/tag/v2.6-NE
4. https://github.com/stockmarkteam/ner-wikipedia-dataset
5. Devlin, Jacob, et al., “Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).
6. Sekine, Satoshi, Kiyoshi Sudo, and Chikashi Nobata. ″Extended Named Entity Hierarchy.″ LREC. 2002.
7. https://github.com/attardi/wikiextractor