The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

エンティティリンキングのための言及抽出手法

Mention detection method for Entity Linking

河田　尚孝 *1　　菊井　玄一郎 *2
Naotaka KAWATA　　Genichiro KIKUI

*1岡山県立大学大学院情報系工学研究科システム工学専攻
Graduate School of Computer Science and Systems Engineering, Okayama Prefectural University

*2岡山県立大学情報工学部情報システム工学科
Faculty of Computer Science and Systems Engineering, Okayama Prefectural University

This paper proposes a method for detecting *entity mentions* in the given text. The method consists of two sequential labelling steps. The first sequential labelling is responsible for identifying a chunk of words (morphemes) mentioning an entity. The second sequential labelling checks whether or not each chunk (or un-chunked word) is really an entity. We also investigated the relation between granularity of categories of named entities and mention detection performance. Our experimental results have shown that the second sequential labeling step for checking slightly improved recall and the eleven categories is the best granularity.

1. はじめに

エンティティリンキングとは, テキスト中の固有表現を抽出し, 知識ベースのエントリー（entity）と関連付けるタスクである. 知識ベースには Wikipedia や DBpedia[Auer 07] などが用いられ, 特に Wikipedia のページを用いる場合は wikification と呼ばれる. エンティティリンキングを利用することで, 自然言語処理の様々なタスクにおいて外部の知識ベースから得られる情報を活用することができる. また, マイクロブログなどの文章にリンクを貼ることで内容を豊かにすることができる.

エンティティリンキングは言及抽出と語義曖昧性解消の 2 つのフェーズに分けることが出来る. 前者はテキスト中でエンティティについて言及 (mention) している部分を同定する処理であり，後者は同定された言及と，それが参照している事物（entitiy) とを紐づける処理である． 一般的に言及抽出の部分では固有表現認識（Named Entity Recognition）のツールが利用される．しかしながら，本稿における言及抽出は言及の位置を同定するだけであり，言及のカテゴリラベル（人名，地名など）も同時に推定する固有表現認識の一部と言える．言及抽出を用いた理由は, エンティティリンキングにおいては言及の位置のみが重要であり，カテゴリラベルの推定の重要度は低いからである．その原因は曖昧性解消においてカテゴリへの依存度が低いためである．

英語の場合，分かち書きがなされ，固有名詞が大文字から始まるなどの手がかりがあることから固有表現抽出の精度が８３％であり，言及抽出はそれ以上の精度と考えられるため，語義曖昧性の解消に焦点をあてているものが多い [Mai 18]．一方, 日本語における固有表現認識の精度は 73%であり, 英語の固有表現認識の精度と比べて十分に高いとは言えない.

日本語における言及抽出の性能を改善する上で 2 つのことが考えられる. 1 つ目は, 長い言及に対する固有表現認識の精度を上げることである. 南ら [南和 11] は, 関根の拡張固有表現クラス [Sekine 08] がアノテートされたコーパスで固有表現認識を行った結果, イベント名や施設名を表す固有表現の認識精度が悪いことを指摘している. 我々は, これらのラベルが付与される言及は比較的長い単語で構成されるため, 抽出が難しいと考える. 2 つ目は, 固有表現認識で対象とするラベルとして適切なラベル数を選択することである. 日本語では関根の拡張固有表現クラスという 200 種類のクラスが付与されたコーパスが存在するが, 学習時間とデータスパースネス問題からこの細かい粒度のカテゴリラベルを使うことが最適とは言えない可能性がある．クラスの階層構造を利用して 10~12 ラベルにまとめることが多い [松田 17][Mai 18] が, 適切なカテゴリ粒度を明らかにする必要がある.

以上の 2 つの観点から，本研究ではニューラルネットを用いた固有表現認識手法をベースとして，次の 2 点により言及抽出の精度向上を図る.

1. 固有表現認識を行った結果に対する再ラベリング
2. 固有表現ラベルの粒度の見直し

本稿の構成は次の通りである．2 章では関連研究について，3 章では提案手法, 4 章では実験について説明する．

2. 関連研究

固有表現認識は系列ラベリングとして考えることが出来る. 系列ラベリングには Hidden Markov Model (HMM) [Seymore 99] や Conditional Random Field (CRF) [Lafferty 01] などのモデルが使用される. 最近の研究ではニューラルネットを使用したモデルが用いられている. 特に日本語の場合では, BiLSTM-CRF モデルを利用することで良い結果が得られており [Mai 18], 本研究でも BiLSTM-CRF モデルを使用する.

2.1 BiLSTM-CRF

BiLSTM-CRF は Lample ら [Lample 16] によって考案されたモデルであり固有表現認識タスクにおいて高い性能を出している. BiLSTM-CRF は 3 層のモデルで構築されており, character-based の BiLSTM Layer, Word-Based の BiLSTM Layer, CRF Layer の 3 層で構成される. モデルの構成を参考文献 [Lample 16] より図 1 に示す. 単語の Word Embedding を入力として BiLSTM に与え双方向の出力を連結した後, 連結

連絡先: 河田　尚孝，岡山県立大学大学院, cd30013b@cse.oka-pu.ac.jp

した出力を CRF Layer に入力して各ラベルを出力する．Word Embedding については単語を文字に分割したものを BiLSTM に入力して作成している．

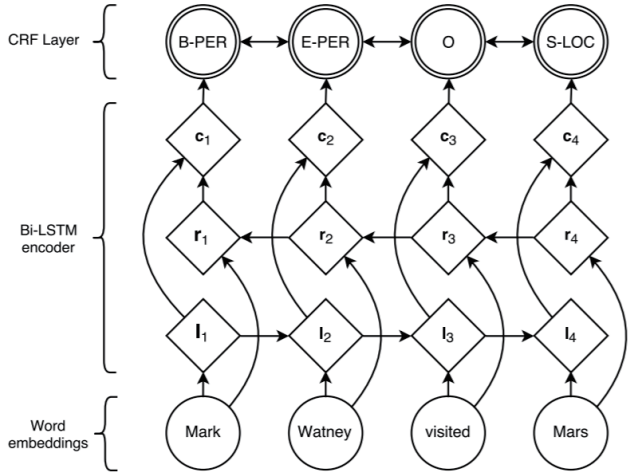

図 1: BiLSTM-CRF のモデル

3. 提案手法

本研究では「再ラベリング」および「固有表現カテゴリの粒度調整」の 2 点により言及抽出の精度向上を図る．以下この 2 点を順に説明する．

3.1 再ラベリング

本研究ではまず入力列（形態素列）に対して通常の固有表現認識を適用したあと，その出力に対して，別の学習データで訓練した固有表現認識を用いて再ラベリングを行う．再ラベリングの目的は最初の固有表現認識の誤りを訂正することである．言及抽出全体の流れを図 2 に示す．

手順 1 では，形態素解析された入力テキストに対して固有表現認識を適用し，得られた BIO-2 ラベルを出力とする．この時に適用する固有表現認識は，固有表現タグ付きの訓練コーパス（開発用コーパスを含む，以下，特記しない限り同様）を形態素解析して BIO-2 タグを付与したデータで訓練したものである．

手順 2 では, 手順 1 の出力に対して一つの言及を構成する形態素列（すなわち，B タグの後に同一カテゴリの I タグが並んだもの）を連結して一つの形態素に変換する処理を行う．

手順 3 では，手順 2 で連結処理を行ったデータに対して次に示す訓練データで学習された固有表現認識を適用して，再ラベリングする．再ラベリング用の訓練データは手順 1 の訓練データをもとに，一つの言及を構成する形態素列を一つの形態素にまとめたものである．図 3 にこの処理を示す．

3.2 ラベルの見直し

固有表現認識において, 言及の位置とラベルの種類の推定が行われるが, 言及抽出においてはラベルの種類は考慮しない．そこでコーパスに与えられたラベルの種類の総数 (細かさ) を変化させて言及抽出を行う．今回はラベルの種類の総数を変更させるために関根の拡張固有表現クラスがアノテートされた拡張固有表現タグ付きコーパス [橋本 08]（以下、ENE コーパスと表記する）の階層構造を利用する．図 4 に固有表現のまとめ方の例を示す．例えば「Organization_Other」や「International_Organization」は「Organization」という上位クラスにまとめることが出来る．「Person」のように細かいラベル分類が存在しないものについてはそのまま「Person」を使用している．ENE コーパスに含まれるラベル数は 200 ラベルであり, この階層構造を利用することで 200 ラベルを 11 ラベルにまとめることが出来る．さらにラベルの種類を考慮せず, 全ラベルを仮に「Mention」と置き換えることでラベルの種類を 1 つにすることが可能になる．

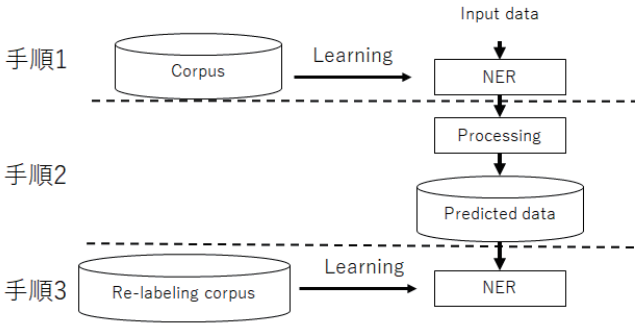

図 2: 言及抽出全体の流れ

Mention	Label
安倍	B-PER
晋三	I-PER
は	O
昨日	O
の	O
夜	O

Mention	Label
安倍晋三	B-PER
は	O
昨日	O
の	O
夜	O

図 3: 固有表現のまとめ方 (左:加工前, 右：加工後)

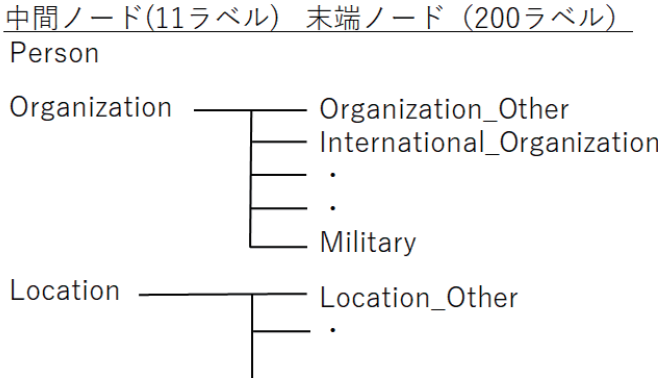

図 4: 関根の拡張固有表現ラベルの階層構造

4. 実験

本研究では再ラベリングによる固有表現認識の誤り訂正 (実験 1) とラベルの種類の総数を変化させた場合の言及抽出 (実験 2) の 2 つの実験を行った．まず, 本研究で使用した系列ラベリングツールとコーパスについて説明する．

4.1 系列ラベリングツール

言及抽出の精度を測定するために系列ラベリングツールとして anaGo[*1] と CRFsuite[Okazaki 07] を用いて実験を行った．

anaGo とは BiLSTM-CRF モデルで作成された系列ラベリングツールである．anaGo のパラメータについては参考文献 [Lample 16] をもとに設定した．使用した最適化アルゴリズムは Adam であり, パラメータは Adam の初期値を使用している．また, anaGo に対して単語ベクトルとして日本語 wikipedia エンティティベクトル [鈴木 16] を追加したものを「+word2vec」と表記している．

続いて CRFsuite を用いた実装を説明する．CRFsuite については素性として, 単語の表層形, 品詞細分類, 文字種を対象単語およびその前後 2 単語から抽出している．

4.2 使用したコーパス

本研究では ENE コーパスと日本語 wikification コーパス [Davaajav 16] を使用した．日本語 wikification コーパスとは ENE コーパスの一部に対してエンティティ情報としてエンティティの ID(Wikipedia の記事にユニーク付与されている ID) を付与したコーパスであり，日本語エンティティリンキングのコーパスとして使用することができる [松田 17]．次に使用するコーパスの内訳について説明する．トレーニングデータとして ENE コーパスに含まれる約 1600 記事, ENE コーパスと包含関係にある日本語 wikification コーパスの 340 記事を半分に分割したものをバリデーションデータ, テストデータとしてそれぞれ使用した．ENE コーパスのラベルの種類について, ENE コーパスにはおよそ 200 ラベルが付与されている．本研究では 200 ラベルから抽出が容易でかつ比較的高い F 値を達成できる時間表現・数値表現を除いた 154 ラベルと,154 ラベルを上位クラスにまとめた 11 ラベルと, ラベルの種類を無視して 1 ラベルにまとめた合計 3 種類の ENE コーパスを準備した．

4.3 実験方法

まず再ラベリングでの実験 (実験 1) を説明する．実験 1 では 11 ラベルに分類された ENE コーパスを使用した．使用した系列ラベリングツールは anaGo であり, 単語ベクトルとして日本語 wikipedia エンティティベクトルを加えた．

続いてラベルの種類を変化させる実験 (実験 2) について説明する．実験 2 では ENE コーパスの階層構造を用いてラベル数を変更して 154, 11, 1 ラベルの分類に変更した計 3 種類の ENE コーパスを使用した．使用した系列ラベリングツールは anaGo と CRFsuite である．

4.4 評価尺度

最後に言及抽出の性能の評価方法について説明する．学習に用いるラベルデータはそのまま使用し, 出力されるカテゴリラベルを無視することで, 言及抽出の結果を測定することとする．

5. 実験結果と考察

まず固有表現を 1 単語にまとめて学習したモデルを用いて再ラベリングを行った場合の言及抽出の結果について表 1 に示す．再ラベリングを行ったことにより再現率が 0.0526 向上, 適合率は 0.0327 減少している．そして F 値は 0.011 向上している．次に再ラベリングによって新たに抽出された言及と削除された言及のうち, 1 形態素のみで構成されている言及それぞれについて, ラベルの変化によって正解になった数および不正解になった数の内訳を表 2 に示す．再ラベリングによって抽出できるようになった言及は 2199 個あった．抽出できるようになった言及は「セ大阪」や「イラク戦争」など言及の 1 部に単独で固有表現となり得るものを含んだものが多かった．一方, 再ラベリングによって削除された言及は 530 個あった．その中には「コメ」や「猫」などコーパス中に出現する頻度が少ない文字種であるカタカナや 1 文字で構成される言及が多かった．

この結果から固有表現を 1 単語にまとめて学習したモデルで再ラベリングを行うことで, 再現率を優先した出力が得られたと言える．言及がより抽出できた原因の一つとして, 固有表現を 1 単語にまとめることにより言及の前後の単語が必ず別の言及か言及以外の単語 (対象とした言及の前後の単語のラベルに B タグか O タグが付与されている) になることで, 言及の前後に出現しやすい単語に注目できるようになったことが挙げられる．また 言及を 1 単語にまとめることで言及の前後の単語との距離が短くなる．LSTM を用いた学習では前後の単語がより重視されるため言及の周りの単語の特徴をより学習できた可能性がある．エンティティリンキングタスクにおいては, 語義曖昧性解消のフェーズで言及にリンクを付与しないことが出来るため, エンティティリンキングの性能向上につながると考えることが出来る．

続いてラベル数を変化させて言及抽出を行った結果を図 5 に示す．図 5 より, anaGo を用いた言及抽出ではラベル数が 1 の場合はラベル数 11 と 154 の場合に比べて F 値は下がっている．ラベル数 11 と 154 の場合について, F 値はほとんど変わらなかった．一方, CRFsuite を用いた場合はラベル数が多くなるほど結果が悪くなった．また word2vec の言及抽出の結果を表 3 に示す．表 3 より, anaGo に word2vec を加えることで言及抽出の性能は向上した．またラベル数が及ぼす F 値への影響は少なくなっている．これは word2vec によって追加される単語ベクトルにより言及抽出の性能が向上し, ラベル数の差が言及抽出の結果に及ぼす影響が小さくなったからだと考える．

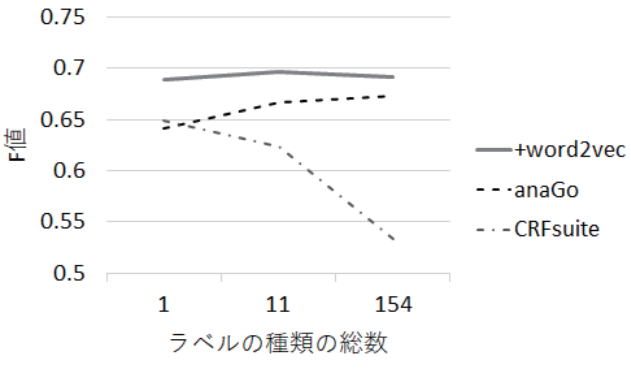

図 5: 日本語に対する言及抽出の結果

6. おわりに

本研究では言及抽出の性能を上げるために 2 つの方法を提案した．1 つ目の方法では, コーパスで分かち書きされている

*1 https://github.com/Hironsan/anago

表 1: 再ラベリングによる言及抽出の性能の変化

	再ラベリングなし	再ラベリングあり
適合率	0.7274	0.6947
再現率	0.6674	0.7200
$F_{\beta=1}$	0.6961	0.7071

表 2: 再ラベリングにより変化した 1 形態素のみで構成される言及の内訳

	言及	非言及	合計
新たに抽出された言及	1030	1169	2199
削除された言及	272	258	530

固有表現を 1 単語にまとめて学習したモデルを用いて予測ラベルの再ラベリングを行った. 2 つ目の方法では, ラベルの種類を変化させて固有表現認識を行った.

1 つ目の方法を行った結果, 再現率を優先した出力が得られた. 適合率は下がっているが, 今後エンティティリンキングタスクに応用する場合, 言及抽出の次に語義曖昧性解消を行う段階で棄却することが可能であるため, F 値を下げることなく, より多くの言及を抽出できたことはエンティティリンキングの性能向上につながると考える.

2 つ目の方法を行った結果, コーパスのラベルを変化させたとき, anaGo に word2vec による単語ベクトルを加えた場合はラベル数を増やしても言及抽出の結果は変わらなかった. 関根の拡張固有表現タグ付きコーパスを使う場合について 154 ラベルよりも 11 ラベルを使用することで言及抽出の性能を落とすことなく計算時間を短くすることができる.

今後の方針について, 今回は言及抽出の性能に着目したが, エンティティリンキングタスクに活用することで言及抽出とエンティティリンキングの性能の関係を調べることが出来る. 日本語 wikification コーパスを用いて日本語でのエンティティリンキングの性能を調査したい.

参考文献

[Auer 07] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z.: Dbpedia: A nucleus for a web of open data, in *The semantic web*, pp. 722–735, Springer (2007)

[Davaajav 16] Jargalsaikhan, D., 岡崎直観, 松田耕史, 乾健太郎：日本語 Wikification コーパスの構築に向けて, 言語処理学会第 22 回年次大会, pp. 793–796 (2016)

[Lafferty 01] Lafferty, J., McCallum, A., and Pereira, F. C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)

表 3: +word2vec の言及抽出

		適合率	再現率	$F_{\beta=1}$
ラベル数	1	0.7249	0.6567	0.6891
	11	0.7274	0.6674	0.6961
	154	0.7418	0.6483	0.6919

[Lample 16] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C.: Neural Architectures for Named Entity Recognition, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, Association for Computational Linguistics (2016)

[Mai 18] Mai, K., Pham, T.-H., Nguyen, M. T., Tuan Duc, N., Bollegala, D., Sasano, R., and Sekine, S.: An Empirical Study on Fine-Grained Named Entity Recognition, in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 711–722, Association for Computational Linguistics (2018)

[Okazaki 07] Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007)

[Sekine 08] Sekine, S.: Extended Named Entity Ontology with Attribute Information, in *LREC 2008* (2008)

[Seymore 99] Seymore, K., McCallum, A., and Rosenfeld, R.: Learning hidden Markov model structure for information extraction, in *AAAI-99 workshop on machine learning for information extraction*, pp. 37–42 (1999)

[橋本 08] 橋本泰一, 乾孝司, 村上浩司　他：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 (NL), Vol. 2008, No. 113 (2008-NL-188), pp. 113–120 (2008)

[松田 17] 松田耕史, 岡崎直観, 乾健太郎：日本語 wikification ツールキット: jawikify, 言語処理学会 第 23 回年次大会 発表論文集, pp. 250–253 (2017)

[南和 11] 南和江, 藤井康寿, 土屋雅稔, 中川聖一：大規模コーパスを用いた固有表現抽出手法の検討, 言語処理学会 第 17 回年次大会 発表論文集, pp. 328–331 (2011)

[鈴木 16] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎：Wikipedia 記事に対する拡張固有表現ラベルの多重付与 (2016)