学習データ作成の省力化に向けたエンティティ抽出手法

白石 哲也　　山崎 貴宏　　村田 稔樹
沖電気工業株式会社　経営基盤本部　研究開発センター
{shiraishi842, yamasaki635, murata656}@oki.com

1　はじめに

文書からエンティティを抽出することは，自然言語処理の重要な技術の一つである．近年では，質問応答や対話，要約といったアプリケーションの需要の増加に伴い，エンティティ抽出技術の必要性が高まっている．エンティティの抽出手法としては，機械学習を適用した様々な手法が提案されている [1]．それらの手法では，大量のラベル付きデータを準備して学習することにより，高い精度でエンティティの抽出が可能となっている．

しかし，それらの手法は大量の学習データを準備する必要があるため，学習データの作成コストが非常に高いことや，専門用語を取り出しにくい，という問題点が挙げられる．また，ドメインに特化した文書群を対象にする状況を考えると，利用できるデータがそもそも少ないため，従来の手法が適用しづらい，という問題もある．

そこで本論文では，利用可能な学習データが少ない場合においても適用可能なエンティティ抽出手法を提案し，データ量が少ないドメインでの提案手法の適用可能性を検証する．

2　タスク設定

学習データが十分に準備できないデータに対してエンティティ抽出を行うために，文書分類技術を応用することを考える．このとき，必要な学習データは分類カテゴリが付与された文書群である．分類カテゴリは章や条などの文書構造,「スポーツ」や「政治」などのジャンル名を利用できるので，学習データは従来のエンティティ抽出手法の学習データよりも容易に準備できる．また，章や条などの構造を持つ文書は社内の規程集や製品のマニュアルなど一般的に存在するため，様々な文書で学習データを作成できると考えられる．一方，分類カテゴリに関係するエンティティが抽出できるという特徴を用いて，専門用語の抽出など，特定の内容に限ったエンティティのみの抽出に適用することが考えられる．そのため，本論文ではエンティティ抽出に文書分類技術を応用することを考え，分類カテゴリに関係するエンティティの抽出をタスクとする．

文献 [2] の章のタイトル「賃金」を分類カテゴリとし，抽出対象となる，分類カテゴリに関係するエンティティ（出力の下線部）の例を図 1 に示す．

第6章　賃金
第32条　基本給
基本給は、本人の職務内容、技能、
勤務成績、年齢等を考慮して各人別に決定する。

↓

入力：基本給は、本人の職務内容、技能、
勤務成績、年齢等を考慮して各人別に決定する。
分類カテゴリ：賃金
出力：基本給は、本人の職務内容、技能、
勤務成績、年齢等を考慮して各人別に決定する。

図 1: 本論文で抽出するエンティティの例

3　提案手法

本章では，エンティティ抽出に文書分類技術を応用する手法を提案する．提案手法は，アテンション機構付きニューラルネットワーク（以下 NN）を用いて文書分類を行い，分類時のアテンションの重みからエンティティを抽出する．

まず，アテンションの重みからエンティティを抽出する方法を述べる．アテンション機構では，NN が注目した度合いを各単語の重みとして付与する．この重みは，分類タスクにおいては分類精度に寄与する一方で，本タスクにおいてはエンティティ抽出に利用する．アテンションの重みが大きい単語を，分類時に注目した単語であると仮定すると，その単語は分類カテゴリに関係する単語であると考えられる．そこで各単語の

アテンションの重みを抽出し，重みが大きい単語のみを獲得することで，分類カテゴリに関係するエンティティを抽出する．

次に，提案手法の特徴を述べる．本手法は文書分類技術を応用しているため，上述の通り，学習データが従来のエンティティ抽出手法で必要となるラベル付きデータよりも容易に作成できる．

4 評価実験

本章では，提案手法を用いてエンティティ抽出ができる可能性を検証するための実験を行う．本実験では，提案手法に有効な学習データとアテンションの閾値によるエンティティ抽出結果の変化を調査する．

4.1 対象データ

本実験では，社内の総務規程集「従業員旅費支給規程」を対象とする．2 章に記述の通り，文書構造を利用することで，章や条のタイトルを分類カテゴリとしてみなす．表 1 に，本実験で用いる学習データを示す．対象の文書は複数の章から構成され，各章には各規程を条として定めたものが記載されている．各章，各条には分類カテゴリとして利用可能なタイトルが付与されるため，分類カテゴリとしてみなす単位が複数考えられる．また，文書分類における『文書』としてみなすデータ単位も複数考えられる．そのため，表 1 に示す 5 種類の学習データが作成可能である．本実験では，これらの学習データでの抽出結果を比較し，提案手法に有効な学習データを調査する．

表 1: 本実験で用いる学習データ

学習データパターン	分類カテゴリ	データ単位	分類カテゴリ数 [件]	学習データ数 [件]
①	章	文ごと	10	128
②	章	条ごと	10	34
③	章	章ごと	10	10
④	条	文ごと	34	128
⑤	条	条ごと	34	34

4.2 マッチングパターン

本実験では，人手で抽出したエンティティを正解とし，提案手法によって抽出したエンティティと正解のエンティティをマッチングすることにより，評価を行う．このときのマッチングパターンを図 2 に示す．図 2 より，マッチングパターンとしては，完全一致するパターン (A)，部分一致するパターン (C, D)，全く一致しないパターン (B, E) が考えられる．本実験では，正解のエンティティが欠けることなく抽出できている図 2 のパターン A, D のみを正解のエンティティが抽出できたとする．

文書： 基本給は、本人の職務内容、技能、勤務成績、年齢等を考慮して各人別に決定する。

分類カテゴリ：賃金

比較パターン	A	B	C	D	E
提案手法	職務内容	-	成績	年齢等	決定する
正解データ	職務内容	技能	勤務成績	年齢	-
抽出できたとみなすか	○	×	×	○	×

図 2: マッチングパターン

4.3 実験設定

本実験では，文献 [3] の方式を用いてアテンション機構を実装した．エンベディング層の次元数は 300, 双方向 LSTM の次元数は 600, アテンション層の次元数は 300 とした．学習エポックは 30, バッチサイズは 20, ドロップアウトは 0.05, 最適化は Adam で行った．また，対象のデータの単語分割は，mecab-ipadic-NEologd[4] を使用した MeCab[5] で行った．

提案手法では，重みが大きい単語を獲得するためにアテンションの重みに閾値を設け，閾値以上の単語をエンティティとして抽出する．そのため，閾値によって抽出されるエンティティの質と量が変化すると考えられる．そこで本実験では，閾値を変化させ，抽出するエンティティの質と量の変化を確認する．0.0〜1.0 に正規化したアテンションのうち，0.1, 0.3, 0.5 の 3 点を設定し，閾値の変化による提案手法の抽出結果を比較する．

4.4 実験結果

表 1 で示した 5 パターンの学習データに対して，4.3 節で示した 3 パターンのアテンションの閾値を設定して実験を行った．表 2 に，Precision, Recall, F 値を評価指標として実験した結果を示す．

表 2: 実験結果

学習データパターン	分類カテゴリデータ単位	アテンションの閾値	Precision [%]	Recall [%]	F 値 [%]
①	章 文ごと	0.1	41.24	75.99	53.46
		0.3	43.76	68.95	53.54
		0.5	39.75	54.97	46.14
②	章 条ごと	0.1	33.29	53.85	41.15
		0.3	32.68	29.93	31.24
		0.5	31.89	20.17	24.71
③	章 章ごと	0.1	39.39	64.73	48.97
		0.3	36.29	28.05	31.64
		0.5	31.36	13.23	18.61
④	条 文ごと	0.1	40.48	**78.24**	53.35
		0.3	**45.05**	70.36	**54.93**
		0.5	43.31	60.04	50.32
⑤	条 条ごと	0.1	36.20	60.51	45.30
		0.3	35.35	36.77	36.05
		0.5	37.29	24.67	29.70

5 考察

5.1 学習データの違いによる変化

本節では，本手法に適した学習データ作成の方針について，分類カテゴリの粒度，および分類カテゴリに含まれるデータ単位の面で考察する．

まず，分類カテゴリの違いによる変化を見る．表 2 より，分類カテゴリが条のときに F 値が高いと分かる．これは，学習データの記載内容が関係していると考えられる．分類カテゴリが章の場合，一つの学習データの記載内容が広範囲にわたる．一方，分類カテゴリが条の場合，それらが別のカテゴリに分かれる．そのため，記載内容が絞られ，分類カテゴリに関係する単語に NN が注目できたため，正解のエンティティをより抽出できたと考えられる．

次に，データ単位の違いについて考察する．表 2 より，データ単位が文ごとのときに最も F 値が高いと分かる．これは，カテゴリ当たりの学習データ数に関係していると考えられる．データ単位が章ごとや条ごとの場合，分類カテゴリ数と学習データ数が同一となる．つまり，一つのカテゴリに対して一つの学習データを作成することになる，その結果，分類時に分類カテゴリに関係する単語に必ずしも注目するとは限らないと考えられる．一方，データ単位が文ごとの場合，分類カテゴリ当たりの学習データ数が多い．そのため，同一の分類カテゴリに共通して含まれる，カテゴリの特徴を表す表現に NN が注目できたため，正解のエンティティをより抽出できたと考えられる．

以上より，文書構造を利用して本手法の学習データを作成するときは，記載内容が絞られる分類カテゴリを付与し，カテゴリ当たりの学習データ数が多くなるデータ単位とする方が良いと考えられる．

5.2 アテンションの閾値による変化

本節では，アテンションの閾値による，抽出したエンティティの質と量を評価し，考察する．表 2 より，アテンションの閾値が小さいほど，Precision と Recall が高い傾向があると分かるが，誤って抽出するエンティティの量が増える．そのため，閾値が小さいほど抽出するエンティティの質が良いとは言えない．現状の提案手法では，抽出する目的に応じて閾値を設定することが重要であると考えられる．例えば，とにかく多くのエンティティの抽出を目的とするような，抽出するエンティティの量を優先する場合では，閾値を低く設定した方が良いと考えられる．また，正解のエンティティのみの抽出を目的とするような，抽出するエンティティの質を優先する場合では，閾値を高く設定した方が良いと考えられる．

5.3 提案手法の適用可能性

提案手法は文書分類技術を応用しているため，学習データの準備が従来よりも容易であると考えられる．また，実験結果より，正解のエンティティのうち約 7 割を抽出できたと分かった．そのため，提案手法を用いてエンティティを抽出できる可能性を見い出せたと考えられる．

5.4 今後の課題

本節では，本手法における，精度面での課題を述べる．

まず，単語区切りに課題があると考える．本実験では文書を単語に区切るために MeCab を用いた．しかし，MeCab の単語区切りと正解のエンティティの区切りが異なることがある．その場合，アテンションの重みが付与される単語と正解のエンティティにズレが生じ，正解のエンティティを正しく抽出できない．これを解決するためには，単語の区切り方を正解のエンティティに合わせる必要があるため，単語の区切り方を変更できる方法が必要となると考えられる．

次に，データ量に課題があると考える．対象であるデータ量が少ないドメインでは，作成できる学習データ量が限られるため，局所的に頻出する一般的な表現を，分類カテゴリ特有の表現と誤って抽出する可能性が高い．そこで，ドメインに依存しない膨大なテキストを用いて事前学習したモデルを活用することで，この誤りを低減し，抽出精度を高めることができると考える．

また，学習データ作成コスト面の評価を実施し，本手法により学習データ作成の省力化の効果を確認することも必要だと考える．

6 おわりに

本論文では，大量の学習データを準備できないデータに対してエンティティ抽出を行うことを目的に，文書分類技術を応用する手法を提案した．また，データ量が少ないドメインでの提案手法の適用可能性を検証し，正解のエンティティを約 70%抽出できたことから，提案手法によってエンティティ抽出ができる可能性を見い出せた．一方で，抽出したエンティティのうち，半分以上が不正解のエンティティを含むという問題点も存在する．今後は，提案手法の精度向上に向け，課題の改善を検討する．

提案手法の応用先としては，プレゼン資料のような，タイトルと本文で構成されている資料に対し，タイトルを分類カテゴリとして本手法を適用することにより，本文からタイトルに関係するエンティティの抽出が考えられる．さらに，抽出するエンティティが分類カテゴリとして利用する文書のタイトルと関係がある，という特性を用い，分類カテゴリとエンティティの関係を特定することでオントロジーとして利用する可能性も考えられる．

参考文献

[1] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[2] 厚生労働省ホームページ. モデル就業規則 第 6 章 賃金. `https://www.mhlw.go.jp/content/000496456.pdf`.

[3] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[4] 奥村学 佐藤敏紀, 橋本泰一. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. In **言語処理学会第** *23* **回年次大会** *(NLP2017)*, pages NLP2017–B6–1. 言語処理学会, 2017.

[5] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 230–237, 2004.