6S-01

ALBERT を用いた Web 記事からの社会問題関連事例の自動抽出手法の試作

神谷　晃†　　白松　俊‡

名古屋工業大学大学院　情報工学専攻‡

1 はじめに

近年，日本では少子高齢化等の持続可能性を脅かす様々な社会問題が増加している．これらの問題に対処するためには、市民が積極的に議論を行い、解決に向けて取り組んでいくことが重要である。しかし、議論に参加するための社会問題についての事例についての情報が不足していると考えられるため，積極的に議論に参加することが難しいと考えられる。そこで，本研究では、社会問題や解決策に関する事例を提供することを目的とし，それらの情報を Web 記事上から抽出する手法を提案する．提案手法では，クラウドソーシングを用いてコーパスを構築し，構築したコーパスを学習データとして用い，事前学習済み ALBERT モデル (1) をファインチューニングすることで，Web 記事から社会問題や解決策に関連する事例を抽出する．

2 社会問題関連事例コーパスの作成

2.1 クラウドソーシングを用いたコーパスの作成

本研究の目的は，自動で Web 上の記事等から社会問題に関連する事例を抽出することである．高精度で抽出を行うためには教師有り学習が望ましく，そのためには人手で社会問題に関連する事例の訓練コーパスを作成する必要性がある．まず，我々は 135 の社会問題に関する web 記事を収集した．その次に，クラウドソーシングサービス Lancers を用い，クラウドワーカーに指定した web 記事から社会問題・取り組みそれぞれに関して最大 3 つ抜き出し（図 1），回答にどれくらいの自信があるのかを 5 段階で入力してもらった。1 つの記事に対して 2〜3 人にアノテーション作業をしてもらった．以下はアノテータに抜き出してもらった項目である．

1. 社会問題（社会の構成員が困るネガティブな困りごと）
   （ア） その社会問題について書いた記述
   （イ） どの地域の社会問題か（都道府県名、市町村名、地域名など）
   （ウ） その問題で誰が困っているか（どういう属性の人々か、あるいは個人名や組織名）
   （エ） その社会問題が主題になっている Wikipedia 記事のタイトル
2. 取り組みや活動（社会問題の解決を目指したもの、あるいは地域振興などのためのもの）
   （ア） その取り組みや活動について書いた記述
   （イ） （社会問題の解決や地域振興のための）どの社会問題についての取り組みか（1. 社会問題で抜粋したうちのどれか）
   （ウ） 誰がやっている取り組みか（個人名や組織名など）

表 1: 文の場合のカッパ値

クラス	カッパ値
社会問題	0.494
取り組みや活動	0.316

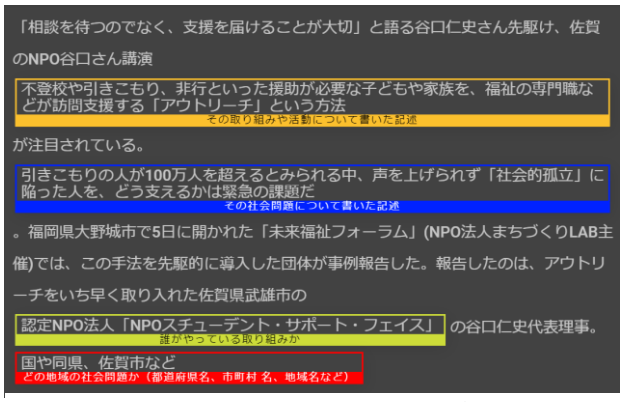

図 1: アノテーションの一例

2.2 社会問題関連事例コーパスの評価

本研究では，文単位での抽出を試みるため，ニュース記事を区切り文字に“。”・“!”・“?”を採用し分解した．アノテーションが行われた部分が 1 文字でも含まれていれば，その文にアノテーションが行われたものとする．ニュース記事を解析した結果，全記事の 1 文の平均文字数は 45.72 文字であった．“社会問題”・

Prototyping a Method for Extracting Cases Relevant to Social Issues from Web Articles Using ALBERT
† Akira Kamiya, Shun Shiramatsu

“取り組みや活動”は，それぞれ 33.92 文字，63.12 文字であり，文単位に近い文字数でアノテーションが行われていることがわかった．

クラウドソーシングの一致度を計るため，Fleiss の Kappa 係数を計算した．アノテータ数が 3 人の場合のコーパスのカッパ値の平均を，表 1 に示す．アノテーションの文字数を考慮して“社会問題”・“取り組みや活動”のみ記載する．“社会問題”については平均が 0.5 弱であり中程度の一致度を確認することができた．“取り組みや活動”についても，一致度の平均が 0.3 強と弱い一致度を確認することができた．

3 社会問題に関連する事例の抽出手法

3.1 提案手法

本研究では，節で構築されたコーパスから，“社会問題”や“取り組みや活動”を文単位で抽出することを目的とする．この問題は，マルチラベル分類問題として扱うことができる．提案したタスクを扱うために，事前学習済みの ALBERT モデルに対し，sigmoid 関数を用いたマルチラベルレイヤーを追加し，ファインチューニングすることによりモデルを構築する．そして，以下のことについて検証を行った．

1. 事前学習に使われたコーパスによる影響
2. 入力形式が与える影響

3.1.1 事前学習済みモデルによる影響

社会問題に関する記事内の文を分類するタスクであることから，転移先のドメインに近いドメインで事前学習を行ったモデルの方が良い精度が出ると考えられる．そのため事前学習済みモデルとして，**日本語ビジネスニュースコーパスで事前学習を行ったモデル**（2）と**日本語 Wikipedia の生テキストをコーパスとして事前学習を行ったモデル**（3）の 2 種類を用意した．

3.1.2 入力形式の仕方が与える影響

Web 記事内の文は周りのコンテキストの情報も重要であると考えられる．そのため ALBERT への入力形式として，以下の 3 つの入力形式を採用する．

1. **当該文のみを入力する形式** 分類したい文をそのまま配置する方法．
2. **前の文脈を考慮した入力形式** 分類したい文を中心として，前の文を出現順通りに配置する方法．分類したい文の segmentID を 1，それ以外を 0 とする．
3. **前後の文脈を考慮した入力形式** 分類したい文を中心として，前後の文を出現順通りに配置する方法（図 2）．分類したい文の segmentID を 1，それ以外を 0 とする．前後の文を入力として用いることで，文脈を考慮した分類を行うことができると考えられる．

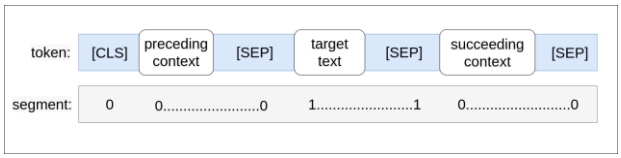

図 2: 前後の文脈を考慮した入力形式

4 評価実験

我々は提案したタスクの評価を行うために，ラベルベースの評価指標のマイクロ平均を用いた．表 3 に，それぞれの分類の適合率・再現率・F1 値のマイクロ平均をまとめる．実験の結果，ニュースコーパスで事前学習を行った分類モデルの F 値が最も高いという結果が出た．ニュースコーパスで事前学習を行ったモデルが全ての入力形式において，Wikipedia で事前学習を行ったモデルと比較して，F 値を上回る結果となった．

表 3: 社会問題に関連する事例の抽出の実験結果

事前学習	入力形式	P	R	F
News コーパス	当該文のみ	0.45	0.57	0.50
	当該文+先行文脈	0.45	0.62	0.52
	当該文+前後の文脈	0.47	0.62	0.53
Wikipedia コーパス	当該文のみ	0.35	0.65	0.46
	当該文+先行文脈	0.44	0.53	0.48
	当該文+前後の文脈	0.49	0.48	0.48

5 おわりに

実験の結果，事前学習モデルのコーパスにニュース記事を用い，前後の文脈も入力して分類を行った場合に，最大の F 値 0.53 を得た．今後は，自動抽出結果を人手で精査し，真の精度を明らかにしたい．その他にも，本研究では分類に使用できなかったコーパス内のアノテータによる確信度も分類を行うための情報として用い，抽出精度の向上を行っていきたいと考えている．

謝辞 本研究の一部は，NEDO（JPNP20006），JST CREST（JPMJCR15E1，JPMJCR20D1），および科研費（17K00461）の支援を受けた．

参考文献

(1) Lan, Z et al.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv:1909.11942 (2020)
(2) Stockmark.: 大規模日本語ビジネスニュースコーパスを学習した ALBERT モデル. https://qiita.com/mkt3/items/b41dcf0185e5873f5f75 (2020)
(3) alinear-corp.: 日本語 Wikipedia コーパスを学習した ALBERT モデル. https://github.com/alinear-corp/albert-japanese (2020)