7Q-03

CoCoA：情報科学分野における日本語の学会発表タイトルの分野推定データセット

宮田 莉奈† 眞鍋 光汰† 福島 啓太† 花房 健太郎† 高田 一慶† 梶原 智之† 桂井 麻里衣‡ 二宮 崇†
愛媛大学† 同志社大学‡

1 はじめに

学会運営の中で生じるプログラム編成や査読者割り当てなどの業務を支援するために、本研究では、学術ドメインに特化したテキスト分類に取り組む。近年、学術ドメインに特化した日本語の事前訓練モデル[1,2]の開発が進んでいるため、それらの評価用データセットの公開によって、研究開発の更なる進展が期待できる。

日本語における既存の学術ドメインコーパスには、論文抄録と投稿先学会名の組からなるCiA25 コーパス[1]や研究課題名と審査区分の組からなる KAKEN コーパス[2]があるが、どちらも非公開である。これらは、研究内容に関する短いテキストから研究分野を推定するという評価のために使用されているが、前者の投稿先学会名は、例えば自然言語処理の分野の研究発表が情報処理学会と言語処理学会の両方で扱われるなど、研究分野の単位としては小さすぎ、後者の審査区分は、例えば自然言語処理の分野と画像処理の分野がどちらも知能情報学として扱われるなど、研究分野の単位としては大きすぎる。

本研究では、これらの課題を解決するために、学会発表プログラムから発表題目とセッション名の組を抽出した上で、人手でセッション名を統廃合して研究分野をアノテーションする。さらに、著者名や発表年の情報も収集し、発表題目・研究分野・著者名・発表年の 4 つ組データセットを構築する。本稿では、情報処理学会全国大会・人工知能学会全国大会・言語処理学会年次大会の過去 10 年分の発表プログラムを対象に構築した CoCoA データセット[1] について述べる。

Japanese Dataset Annotated with Research Fields for Presentation Titles in the Area of Computer Science
Rina Miyata† (miyata@ai.cs.ehime-u.ac.jp)
Kouta Manabe† (manabe@ai.cs.ehime-u.ac.jp)
Keita Fukushima† (fukushima@ai.cs.ehime-u.ac.jp)
Kentaro Hanafusa† (hanafusa@ai.cs.ehime-u.ac.jp)
Kazuyoshi Takata† (takata@ai.cs.ehime-u.ac.jp)
Tomoyuki Kajiwara† (kajiwara@cs.ehime-u.ac.jp)
Marie Katsurai‡ (katsurai@mm.doshisha.ac.jp)
Takashi Ninomiya† (ninomiya.takashi.mk@ehime-u.ac.jp)
†Ehime University ‡Doshisha University

[1] https://github.com/EhimeNLP/CoCoA

2 CoCoA データセット

情報科学分野の 3 つの国内学会を対象に、2014 年度から 2023 年度までの 10 年分の発表プログラムから、発表題目・セッション名・著者名・発表年の 4 つ組を収集した。ただし、日本語が含まれない発表題目は対象外とした。そして、以下のとおり、セッション名を人手で統廃合して研究分野をアノテーションした。

情報処理学会 全国大会の一般・学生セッションからデータを収集した。研究分野として、講演募集分野の一覧[2]に掲載の 41 分野を採用し、キーワードの一覧[3]を参考にアノテーションした。

人工知能学会 全国大会の一般セッションからデータを収集した。研究分野として、AI マップβ2.0[4]の「E：AI 研究の現在」に掲載の 11 分野を採用し、各分野に対応するキーワードを参考にアノテーションした。

言語処理学会 年次大会の一般発表からデータを収集した。ただし、セッション名と研究分野の関連が明らかではない一部のポスターセッションは対象外とした。研究分野は、本稿の著者らがセッション名を統廃合して 23 種類を定義した。統合は「マルチモーダル」と「マルチモーダル処理」など同一分野を指すセッション名をまとめ、廃合は「機械翻訳／質問応答」など異なる分野が（おそらく発表件数の都合で）同一セッションとしてまとまっている場合にセッション単位で除外した。

表 1 に、本データセットの統計情報を示す。語彙サイズは、発表題目を Academic RoBERTa[2] のトークナイザで分割した際のトークンの種類数である。研究分野のアノテーションは、学会ごとに本稿の著者の 1 人が担当した。アノテーションの信頼度を評価するために、別の著者 1 人が 50 件ずつをアノテーションしたところ、$0.72 < \kappa < 0.91$ の充分高い一致を確認できた。

本データセットは、学術ドメインにおける自然言語処理のベンチマークとして複数の方法で利用できる。例えば、発表題目からの研究分野の推定や著者の推定というタスクが考えられる。

[2] https://www.ipsj.or.jp/event/taikai/86/keyword.html
[3] https://www.ipsj.or.jp/kenkyukai/bunya2024.html
[4] https://www.ai-gakkai.or.jp/aimap/

表 1：CoCoA データセットの統計情報

学会名	論文数	分野数	著者数	語彙サイズ
情報処理学会（IPSJ）	12,018	1,085 → 41	17,346	10,361
人工知能学会（JSAI）	3,505	113 → 11	6,015	6,183
言語処理学会（ANLP）	1,818	139 → 23	2,412	3,894

表 2：分野推定の実験結果（Accuracy）

	IPSJ	JSAI	ANLP
ランダム	0.03	0.08	0.03
最頻値	0.09	0.21	0.14
早稲田 RoBERTa	0.36	0.45	0.40
Academic RoBERTa	**0.44**	**0.52**	**0.51**

表 3：発表年推定の実験結果（RMSE）

	IPSJ	JSAI	ANLP
ランダム	4.06	3.88	4.14
最頻値	3.82	4.99	3.85
早稲田 RoBERTa	**3.01**	3.48	2.81
Academic RoBERTa	3.52	**3.41**	**2.71**

3 評価実験

3.1 実験設定

本研究で構築した CoCoA データセットを用いて、発表題目からの研究分野の推定および発表年の推定に関する評価実験を行った。前者は分類タスクとして設計して Accuracy で評価し、後者は回帰タスクとして設計して RMSE で評価した。データセットは、研究分野および発表年の偏りがないように、8:1:1 の割合で訓練用・検証用・評価用に分割して使用した。

比較手法　ベースラインとして、無作為に選択肢を選ぶ「ランダム」および最も頻出する正解選択肢を常に選ぶ「最頻値」の 2 手法を用いた。また、汎用的な事前訓練モデルとして Wikipedia および Web テキストで訓練された早稲田 RoBERTa[5][3]、学術ドメインに特化した事前訓練モデルとして論文抄録で訓練された Academic RoBERTa[6][2]の 2 モデルを用いた。

ハイパーパラメータ　ファインチューニングでは、バッチサイズを 64 文、学習率を 5×10^{-5} に設定し、最適化手法には AdamW を用いた。そして、検証用データにおける各タスクの評価指標が 3 エポック連続で改善しない場合に訓練を終了する early stopping を適用した。

3.2 実験結果

分野推定　発表題目から研究分野を推定する実験の結果を表 2 に示す。機械学習モデルがベースラインよりも顕著に高い性能を示したことから、発表題目には研究分野ごとの特徴的な表現が含まれていることが示唆される。また、全ての学会において、論文抄録を用いて事前訓練された Academic RoBERTa が最高性能を示したことから、学術ドメインに特化した事前訓練によって学術ドメインの自然言語処理の性能を改善できることが示唆される。

発表年推定　発表題目から発表年を推定する実験の結果を表 3 に示す。分野推定と同様に、機械学習モデルがベースラインよりも高い性能を示した。本タスクにおいても、3 学会のうち 2 学会に対しては、学術ドメインに特化した Academic RoBERTa が高性能を達成した。

考察　分野推定タスクにおいて機械学習モデルがベースラインよりも顕著に高い性能を示したことから、研究分野ごとに特徴的な表現が発表題目に含まれていると考えられる。これを確認するために、発表題目中の単語と研究分野の間の関連度(Pointwise Mutual Information; PMI)を算出した。その結果、例えば言語処理学会においては、PMI の上位に「Web 応用-SNS」や「Web 応用-コロナ」などの組が見られた。SNS やコロナは、自然言語処理の Web 応用という研究分野においてキーワードであると考えられる。

4 おわりに

本研究では、情報科学分野における日本語の学会発表タイトルを著者名や発表年とともに収集し、研究分野を付与した。本データセットを用いることで、学術ドメインに特化した自然言語処理モデルのベンチマークが可能となる。

謝辞　本研究は、JSPS 科研費（基盤研究 B，課題番号：JP20H04484）の助成を受けて実施した。

参考文献

[1] 壹岐 太一, 金沢 輝一, 相澤 彰子. 学術分野に特化した事前学習済み日本語言語モデルの構築. 情報処理学会第 139 回情報基礎とアクセス技術研究会, 2020.

[2] Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai, Takashi Ninomiya. A Japanese Masked Language Model for Academic Domain. In Proc. of SDP, pp. 152–157, 2022.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, 2019.

[5] https://huggingface.co/nlp-waseda/roberta-base-japanese
[6] https://huggingface.co/EhimeNLP/AcademicRoBERTa