新聞記事からの新技術の用途情報の抽出

加藤裕太　　福田悟志　　難波英嗣

中央大学大学院　理工学研究科

1.　はじめに

日々、さまざまな分野で数多くの新技術が開発されており、これらの技術に関する論文も数多く執筆されている。しかし、非専門家にとって、論文を読むだけでは、その技術に関する理解が難しい。また企業などが、新技術を活用して新たな製品やサービスを生み出す際に、新技術の用途や応用先が一目で把握することができれば、新しいアイデアにもつながりやすく、開発が容易になると考えられる。本研究では、新聞記事からの技術の用途や応用先に関する情報を自動抽出する手法を提案する。さらに、抽出された情報との関連付けを行うシステムを構築する。

2. 関連研究

難波ら[1]は、論文用語を特許用語に自動変換する手法を提案している。本研究では、E5[1]を用いることで、新聞記事と論文での用語の違いを考慮せず、新聞記事と論文を対応する手法を提案する。

Iwayama ら[2]は、特許と論文をジャンル横断情報アクセスする手法を提案している。本研究では、新聞記事と論文においてジャンル横断情報アクセスを行う手法を提案する。

3.　技術記事分類器と用途文分類器の構築

本研究では分類器を二つ構築した。まず一つ目は、BERT[2]または T5[3]を用いて新聞記事が新技術に関する記事であるかどうかの二値分類を行う分類器である。二つ目は、新技術に関する記事であると判断された記事の本文を、一行ずつ新技術の応用先や用途などが書かれた文であるかの二値分類を行う分類器を構築した。

4. 新聞記事と論文の対応付け

本研究では、新聞記事と論文ともに E5 を用いて埋め込み表現に変換し、コサイン類似度を計算した。E5 を用いることで、新聞記事と論文での用語の違いを考慮せずに文の類似度を算出することができる。クエリに技術に関する新聞記事を入力し、コサイン類似度が高い論文を上位 10 件出力する。

Extraction of Information on Applications of New Technologies from Newspaper Articles
Yuta Kato , Hidetugu Nanba and Satoshi Hukuda・Chuo University Graduate School of Science and Engineering

5. 実験

構築した分類器の有効性と新聞記事と論文の対応付けの有効性を確認するため、実験をした。

5.1 実験条件

実験に使用するデータは、日本経済新聞と日刊工業新聞に関するデータセットである。新技術に関する記事であるものには、人手で<技術>タグを見出しに付与する。また<技術>タグを付与された記事の本文の新技術の用途となる文には、人手で<用途>タグを付与した。記事データの日本経済新聞と日刊工業新聞の記事数とそれぞれの<技術>、<用途>タグの個数は表 1 に示す。

表 1　データセットの概要

	記事数	〈技術〉	〈用途〉
日本経済新聞	15,000	237	589
日刊工業新聞	3466	340	598

また、新聞記事と論文の対応付けでは、日刊工業新聞の 2020 年の記事を 3 節の技術記事を分類する分類器で分類した記事と J-Stage の 2018 年の論文を無作為に 10 万件抽出したものを用いる。

評価方法は分類器では precision、recall、f1 の値を使用する。新聞記事と論文の対応付けは precision を用いる。

5.2 分類器の実験結果

技術に関する記事の分類結果を表 2、用途文の抽出結果を表 3 に示す。

表 2　技術に関する記事の分類結果

	precision	recall	F1-score
BERT<技術>	0.7104	0.7695	0.7388
T5<技術>	0.7382	0.6636	0.6989

表 3　用途文の抽出結果

	precision	recall	F1-score
BERT<用途>	0.8419	0.8232	0.8324
T5<用途>	0.7993	0.7184	0.7567

5.3 新聞と論文の対応の実験結果

出力された論文が新聞記事に関連しているかを人手で判断し、精度を算出している。E5 を用いてコサイン類似度を出力したものと FAISS を用いてベクトル検索を用いたものを比較したものを表 4 に示す。また、新聞記事から用途に関する文を除いたものと、本文をクエリとした精度の比較もしている。

表 4　新聞と論文の対応付けの精度の結果

	用途文あり	用途文なし
E5	0.50	0.50
FAISS	0.50	0.60

E5 と FAISS の出力した論文の一致度を、用途文ありと用途文なしで比較したものを表 5 に示す。

表 5　E5 と FAISS の一致度

用途文あり	用途文なし
0.60	0.70

5.4 考察

技術に関する記事の検索結果と用途文の抽出結果は、どちらも T5 に比べ BERT の方が良い結果が出た。〈技術〉タグに関しては、データの割合的にタグが振られているものが少ないため、学習時に〈技術〉タグのついた記事の割合を増やせば評価は上がると考える。〈用途〉タグは、学習時にタグが付与された割合も多いため、全ての評価指標において８割近くの値が出た。残りの２割は〈用途〉タグが付与されるような文末になっているが、〈用途〉タグが付与されないものにタグを付与してしまったことで予測が外れていると考える。

新聞と論文の対応においては、精度が E5、FAISS のともに約 0.50 を出せている。今回は、論文を無作為に 10 万件抽出したため、全ての論文を用いることで精度は上がると考える。また、精度を出す際に、上位 10 件を出力したが、出力する論文を増やすことで精度が上がると考えられる。

6. システムの動作例

新聞記事からの技術の用途や応用先に関する情報を自動抽出するシステムの動作例を図 1 に示す。検索をしたい技術名をクエリとして入力すると、技術名が含まれる新聞記事が出力される。

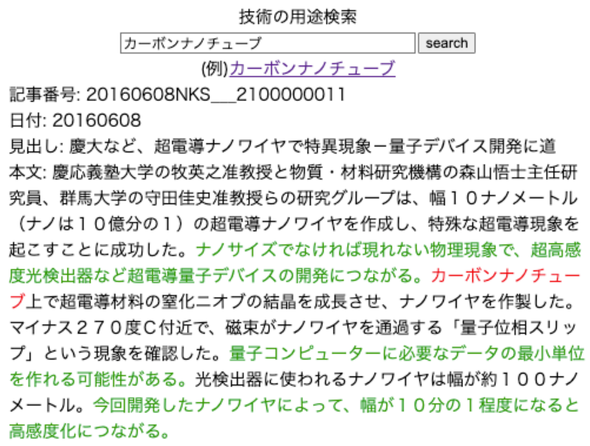

図 1　用途情報抽出システムの一例

出力された新聞記事は、図 1 のように技術の用途情報の文が色付けされて出力される。

図 1 の例では、カーボンナノチューブが、「超高感度光検出器など超電導量子デバイスの開発につながる。」などということが一目でわかる。

次に新聞記事と論文の対応付けの例を図 2 に示す。

対応付けの例
〈新聞記事〉 慶応義塾大学の牧英之准教授と物質・材料研究機構の森山悟士主任研究員、群馬大学の守田佳史准教授らの研究グループは、幅１０ナノメートル（ナノは１０億分の１）の超電導ナノワイヤを作成し、特殊な超電導現象を起こすことに成功した。ナノサイズでなければ現れない物理現象で、超高感度光検出器など超電導量子デバイスの開発につながる。〈略〉
〈論文〉 THz からの UV に広がる広いスペクトル領域の分光偏光解析法(SE)測定は, 単層カーボンナノチューブ薄膜の非ドープおよび硝酸ドープ状態の両方でである。 <中略> 硝酸ドーピングは～100 倍～4 と ρ_⊥の因子による ρ_を減少させるという事実は, ドーピングはチューブ - チューブ接合で輸送するエネルギー障壁を減少させるのに重要な役割を果たすことを示唆する。

図 2　新聞と論文の対応付けの一例

7. おわりに

本研究では、約 18,000 件の新聞記事データを用いて、新技術に関する記事であるか、また新技術の用途に関する文であるかを分類する分類器を構築した。実験結果としてはどちらの分類器も BERT が最も良い結果となった。

また、新聞記事と論文の対応付けも行った。E5 を用いて類似度を算出したものと、FAISS を用いたものともに精度は約 0.50 であった。

謝辞

本研究の遂行にあたり、データを提供してくださった日刊工業新聞様、また研究に関して議論をしていただいた、株式会社ジー・サーチ様に厚くお礼申し上げます。

参考文献

[1] 難波 英嗣 他, “論文用語の特許用語への自動変換”, 『情報処理学会論文誌データベース』, Vol.2, No.1, 81-92 (2009).

[2]Iwayama,M. et al, “Overview of Patent Retrieval Task at NTCIR-3, Working Notes of the 3rd NTCIR Workshop Metting, Part Ⅲ”, Patent Retrieval Task, 1-10 (2002)

[1] https://huggingface.co/intfloat/multilingual-e5-base
[2] https://huggingface.co/docs/transformers/model_doc/bert
[3] https://huggingface.co/docs/transformers/model_doc/t5