検索報告書を教師データとする先行技術文献検索システムの提案

福西 章人　　鶴岡 慶雅
東京大学 大学院情報理工学系研究科
{fukunishi,tsuruoka}@logos.t.u-tokyo.ac.jp

1. はじめに

近年の人工知能技術の発展に伴い、我が国の省庁においても、業務に人工知能技術を活用する取り組みが進められている。特に、特許庁においては、近年の産業財産権を取り巻く環境の多様化に伴う行政事務の複雑化や、新興国等における出願の増加に伴う先行技術調査の対象となる資料の増加など、事務処理量が増大しており、業務の機械化が望まれている。

こうした環境変化に伴う業務量の増大に対応すべく、特許庁は「人工知能技術を活用した特許行政事務の高度化・効率化実証的研究事業」[1] を外部の事業者に委託することで、892 の業務を対象として人工知能技術の活用可能性を検討し、各業務に対する人工知能技術の導入可能性について、その困難度に応じた分類を行うとともに、一部の業務については先行的に実機を用いた実証を行いつつ、当面のアクション・プラン [2] を公表した。

本論文において提案する文献検索システムは、上記の研究事業において機械学習による対応が特に困難とされた業務である先行技術調査業務を支援することを目的としている。特許庁は外部の事業者である登録調査機関に先行技術調査業務の一部を委託しているが、増え続ける外国語特許文献等に対応すべくその規模を拡大しており(平成31年度予算案は約260億円)、機械化が実現すれば恩恵は大きい。

一方、登録調査機関による先行技術調査の結果が記載された検索報告書が、年間約 153,000 件作成されており、先行技術調査の機械化のための学習データとして利用できる可能性がある。

そこで本研究では、特許庁の先行技術調査業務への人工知能技術の活用可能性を探るべく、検索報告書を教師データとする文献検索システムを提案するとともに、当該システムにおいて良い性能を発揮する分散表現を探るための評価実験を行った。

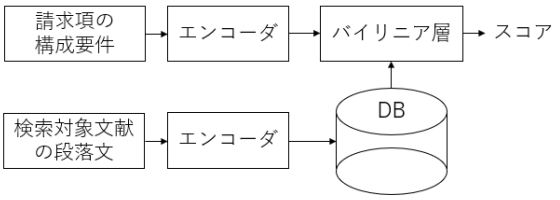
2. 関連研究

先行技術調査の機械化の試みの先行例としては、特徴語の出現頻度に基づく概念検索による類似文献の抽出 [3] がある。この手法は、複雑な検索式や専門性の高い検索インデックスを用いることなく、自然文から容易に類似の文献を検索することができる点で非常に有用である一方、検索対象となる発明概念を正確に捉えられないなどの理由によって、文献の抽出精度が十分でないといった課題も存在している。

また、検索対象となる各文献を文献単位で予めベクトル化し、入力文との類似度をスコア化するものなので、抽出の粒度が文献単位であり、それに伴う課題も存在する。すなわち、上記システムは、抽出された文献のどの部分が特に重要であるかを提示するものではないので、審査官や先行技術調査業務者は、抽出された文献全体に目を通し、引用すべき箇所を特定する必要がある。

さらに、特許審査における先行技術調査においては、進歩性を否定する際の主引用文献と副引用文

図1：検索システムの主要部



献といった、複数の文献の組み合わせを検討する必要があるが、そのような文献の組み合わせを提示する機能は存在しておらず、審査官や検索者は、抽出された文献に請求項に関する構成がどこまで記載されているかを把握し、足りない構成が存在する場合は、他の文献で補えるものがないかを別の方法で探すなど、依然として大きな業務負担を抱えている。

また、検索報告書を学習データとして利用する先行例として、特許第6421963号［4］には、自然文の入力に対して有用な検索インデックスおよび技術用語を提示するシステムを学習させるために、検索者が作成した検索式を教師データとして利用することが記載されている。

3. 提案手法

上記した概念検索システムの課題に対応するために、検索対象となる文献を予め段落単位でベクトル化し、抽出の粒度を段落単位とするとともに、検索対象となる請求項を複数の構成要件に分割し、構成要件毎に類似の段落を抽出することで、進歩性を否定しうる文献の組み合わせを提示するシステムを提案する。

3．1　検索システムの概要

提案する検索システムの主要部を図1に示す。検索の母集団となる文の集合は、予めエンコーダで低次元の分散表現に変換して蓄積しておく。請求項の文が入力されたときに、請求項を複数の構成要件に分割し、それぞれをエンコーダで分散表現に変換し、蓄積された検索母集団の各分散表現との関連性をバイリニア層で計算する。こうして計算されたスコアによって、構成要件毎に関連性の高い文献を、段落を特定したうえで提示できるとともに、進歩性を否定しうる複数の文献の組み合わせを提示できる可能性がある。

分散表現間の類似度は、コサイン距離などによっても計算することができるが、本研究ではバイリニア層を使用した。請求項の構成要件と引用対象となる文との関係は、意味的に包含関係にあることが望ましく、順序を交換しても結果が変わらないコサイン距離によるモデルではそういった包含関係を表現できないと考えたからである。

エンコーダ部分およびバイリニア層の学習は、検索報告書から抽出した請求項の構成要件と引用文献の対応箇所との文の対によって行う。

3．2　分散表現

特許文献には、一般的に用いられる専門用語の他に、「デジタル有線テレビジョン放送受信装置」などのように、複数の名詞を連結することによって作られた長い名詞が頻出する。単語の分散表現を生成するにあたり、このような長い名詞については、語彙数削減の観点から分割することも考えられるが、本研究においては1単語として扱った。こういった名詞を1単語として扱う場合、文字列として類似する複数の単語がかなりの割合で語彙に含まれることとなるが、それらは近い分散表現となることが望ましい。そこで、skip-gram［5］のような単語の共起のみに基づく手法ではなく、単語の内部構成も考慮する fasttext［6］を用いることとした。

文の分散表現は、登場する単語の平均によるもの(fasttext-BoW)、および、LSTM の隠れ層によるものについて評価実験を行った。

3．3　学習データ

登録調査機関が作成した検索報告書には、教師データとして有用と考えられる様々なデータが含まれて

表 1:学習データとして作成した文の対の例

	請求項の構成要件	引用文献の段落
正解データ	前記シリンダの内部に保持されて磁場を発生するコイルと、	磁場発生装置をダンパチューブのMR流体収容室に臨む部分に設けるものであり……
不正解データ(段落のみを変更)	前記シリンダの内部に保持されて磁場を発生するコイルと、	MR流体ダンパ10は、ダンパチューブ11のMR流体収容室13に臨む部分……
不正解データ(文献と段落の両方を変更)	前記シリンダの内部に保持されて磁場を発生するコイルと、	そして、上記従来の管継手の場合、継手本体2の受け口21の環状凹部211の奥部側……

表2:分散表現ごとの正解率、F 値

	正解率	F 値
fasttext-BoW	69.7%	0.706
uni-LSTM	53.5%	0.378
LSTM-concat	54.0%	0.383
LSTM-avg	54.6%	0.413

表3:fasttext-BoW について、正解率の内訳

	正解率
全体	69.7%
正解データ	76.8%
段落のみを変更した不正解データ	26.3%
文献と段落の両方を変更した不正解データ	90.3%

いる。本研究では、「スクリーニングサーチの結果(クレーム別形式)」の項目を利用する。当該項目には、請求項と引用文献の関連箇所とが対応付けて記載されている。

検索報告書は一定のフォーマットで記載されているが、完全に統一されているわけではなく、登録調査機関や検索者によって、記載形式や内容の詳細さに違いが存在する。今回は、クレームを複数の構成要件に分割し、それぞれの構成要件について引用文献の関連箇所が個別に対応付けて記載されている形式の検索報告書を抽出して学習に使用した。

4. 実験

提案したモデルで用いる文の分散表現として性能の良いものを探るための実験を行った。検索母集団の巨大さを考慮すると、分散表現の次元数は可能な限り小さくすることが望ましい。

4.1 実験設定

検索報告書から、請求項の構成要件と対応する引用文献番号および段落番号を抽出し、日本語特許公報データから当該段落番号に対応する文を得た。こうして作成した請求項の構成要件と引用文献の該当箇所の文の対 129,685 件を正解データとした。さらに、正解データのうち、段落番号のみを変更することで作成した不正解データ60,878 件および、文献番号と段落番号の両方を変更することで作成した不正解データ 83,997 件も学習データに追加した。作成した学習データの例を表1に示す。

単語の分散表現は、過去25年分の日本語公開特許公報等の本文で学習した fasttext を使用した。語彙数は約 9,000,000 で、次元数は 300 とした。

文の分散表現として、下記の4つについてモデルの学習を行い、学習後に入力された文のペアが正解データであるか不正解データであるかを判別する2値分類問題の正解率によって分散表現の評価を行った。

(a) fasttext-BoW：文に含まれる単語の fasttext によるエンベディングの平均(300 次元)
(b) uni-LSTM：単方向 LSTM の最終隠れ層の出力(512 次元)
(c) LSTM-concat：双方向 LSTM の2つの最終隠れ層の出力を結合したもの(512 次元)
(d) LSTM-avg：双方向 LSTM の2つの最終隠れ層の出力を平均したもの(512 次元)

4.2 結果

100 エポックの学習を行った実験結果を表2，3に示す。LSTM によるものは、正解率および F 値において非常に低い結果となった。これは、入力である引用文献の関連箇所の系列長が長すぎて学習に失敗したためと考えられる。fasttext-BoW は、70%程度の正解率であるが、2値分類問題であることを考慮すると実用に耐える精度であるとは言えない。fasttext-BoW について、個別に正解率を見ると、段落のみを変更して作成した不正解データについては正解率が極めて低い。同じ文献に属する段落間で分散表現の差が小さく、正解、不正解の区別が難しいことが読み取れ、提案するシステムが目指すところである、関連文献の段落を特定して提示するというタスクの困難性が窺がえる。

5. おわりに

本研究において、検索報告書を教師データとする文献検索システムを提案したが、当該システムで使用する文の分散表現については、実験で用いたいずれの手法も十分な性能ではなかった。使用する学習データ数を増やす、或いは、quick-thoughts [7] のような文の分散表現を得る他の手法を試すなど、実用に耐える良い分散表現を得るための研究を継続する必要がある。また、文献レベルで候補を絞り込んでから精査することで提示すべき段落を特定するなど、アプローチを変えながらより良いモデルを模索する必要がある。

参考文献

[1] 株式会社エヌ・ティ・ティ・データ経営研究所，人工知能技術を活用した特許行政事務の高度化・効率化実証的研究事業報告書(エグゼクティブサマリ)， 2017
https://www.jpo.go.jp/torikumi/t_torikumi/pdf/ai_action_plan/02.pdf

[2] 特許庁， 特許庁における人工知能(AI)技術の活用に向けたアクション・プランの平成 30 年度改定版について， 2018
https://www.jpo.go.jp/torikumi/t_torikumi/ai_action_plan-fy30.html

[3] 間瀬 久雄，特許を対象とした概念検索の技術課題 概念検索精度の向上に向けて，Japio YEAR BOOK 2010 pp. 200-207， 2010

[4] 神谷 径， 検索インデックス推定機、コンピュータプログラム及び検索インデックス推定方法， 特許第6421963号， 2018

[5] Mikolov Tomas, Corrado G.s, Chen Kai and Dean Jeffrey. Efficient Estimation of Word Representations in Vector Space. *arXiv:1802.00921,* 2013

[6] Enriching Word Vectors with Subword Information Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Transactions of the Association for Computational Linguistics 2017 Vol. 5, 135-146

[7] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. International Conference on Learning Representations 2018