

論文概要の分散表現を用いた関連論文検索

高德 卓[†] 檜垣 泰彦[†]

[†]千葉大学大学院工学研究科 〒263-8522 千葉県千葉市稲毛区弥生町 1-33

E-mail: higaki.yasuhiko@faculty.chiba-u.jp

あらまし 本研究の目的は、論文概要の分散表現を用いた提案手法による関連論文検索の有効性を検証することである。本手法では論文概要を入力とし、学習済みの論文概要とのコサイン類似度を計算することで関連論文の検索を行う。検索精度の向上のため、html タグ等の除去と記号や専門用語等の正規化を行った。また、より高い検索精度を持つ環境の検討のために事前実験を行い、得られた結果を踏まえて提案手法に追加実装した。評価実験においては比較手法との比較の際に外的影響を排除するため、実験環境を統一した評価用システムによって評価の信頼性を高めた。評価実験の結果より、提案手法は比較手法と比べ、全ての評価指標において優れた値が得られ、有効性が認められた。

キーワード 分散表現, 論文検索, コサイン類似度, doc2vec, 自然言語処理

Article Retrieval Using Distributed Representations of Abstracts

Suguru TAKATOKU[†] Yasuhiko HIGAKI[†]

[†]Graduate School of Engineering, Chiba University 1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522 Japan

E-mail: higaki.yasuhiko@faculty.chiba-u.jp

Abstract The purpose of this study is to evaluate article retrieval using distributed representations of abstracts. This method outputs articles which have small distances for abstract as input. We removed html tags and normalized special characters and technical words for improvement of precision. We determined environment using my method on the basis of results of preliminary experiments. On evaluation experiments, evaluation system unified experiment environment improved reliability of evaluations. We implemented various preprocess to increase the precision. Therefore, we confirmed that the proposed method has better performance than comparison method.

Keywords Distributed Representations, Articles Retrieval, Cosine Similarity, doc2vec, Natural Language Processing

1. 序論

自然言語分野において文の意味を数値化し、コンピュータ上で扱う試みは盛んに行われている。文をベクトル空間で扱う手法の中で、近年特に注目を浴びているのが、ニューラルネットワークに基づく学習である。その学習過程で得られたベクトルは分散表現と呼ばれ、単語の意味を保持したまま文や文章を扱うことができる。

また、論文検索は研究活動において研究の新規性や既存研究の動向を調査するための重要な過程である。論文検索の手段としては、Google Scholar[1] や CiNii Articles[2] といった論文検索サービスを用いたキーワードによる検索が主流となっている。しかしながら、研究分野に精通した研究者にとっても、適切な専門用語の組み合わせを考慮し、検索漏れを防ぐことは難しい。そのため、ユーザが検索クエリを考案する必要のない関連論文検索システムは存在する[3][4]。しかし、既存システムは入力としては文書を用いるが、検索に用いるのは文書の特徴語であり、文書自体の類似性を考慮したものではない。

そこで、本研究では論文概要の分散表現を用いた関連論文検索の有効性を検証することを目的とした。

2. 文章を入力とした関連論文検索

林らの関連論文検索[3]は論文概要を入力とし、入力から Term Frequency-Residual Inverse Document Frequency (TF-RIDF)を用いて重要単語を抽出し、初期クエリを生成している。次に、単語の意味概念行列を用いて重要単語に対して類似度の高い語を抽出し、サブクエリとしている。その後、初期クエリとサブクエリを用いて検索クエリを発行している。

ふわっと関連検索[4]は任意の文章を入力とし、入力から MeCab の生起コストを用いて特徴語を抽出している。その特徴語を用いてクエリ検索を行っている。本手法との比較対象として、ふわっと関連検索の一手法(比較手法)を用いた。

比較手法:特徴語の抽出方法は MeCab の生起コストを対数化した値を用いる。特徴語の重みで降順に並べた上位 10 単語を AND 条件で連結し、クエリとした発行する。10 件以上のヒットがない場合は、 $n-1$, $n-2$, ..., とクエリを実行する語数をスコアの逆順で減らしていき、10 件以上ヒットするまでこれを行う。その論文をヒット回数の降順で並び替え、出力する。

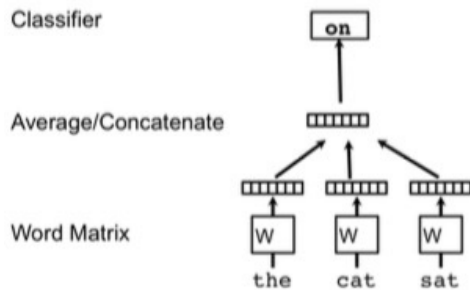


図.1 単語の分散表現のフレームワーク[5]

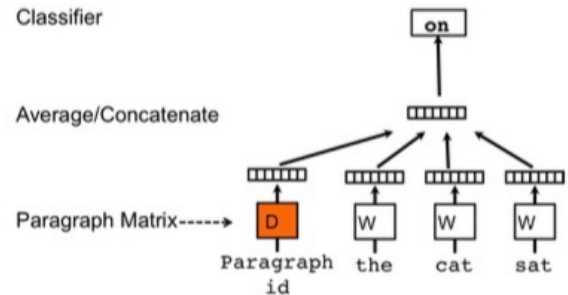


図 2. Paragraph Vector のフレームワーク[5]

3. 分散表現

文を表現する上で広く用いられているのは Bag of Words (BoW)やN-gramである。しかし、それぞれ語順を無視してしまうこと、N の値が大きくなるほど組み合わせの数が増え学習結果の散在に陥ること、等の問題があった。本研究では単語の意味概念を保持したまま文章を扱える Mikolov らの手法、Paragraph Vector[5]を用いる。Paragraph Vector は上記の問題点を解決している。

3.1. 単語の分散表現

まずは、Paragraph Vector の基礎となる単語の分散表現について述べる。訓練語 $w_1, w_2, \dots, w_T$ が与えられた場合に目的関数は、

$$\frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, \dots, w_{t+k}) \quad (1)$$

となり、予測モデルは

$$p(w_t | w_{t-k}, \dots, w_{t+k}) = \frac{e^{y_{wt}}}{\sum_i e^{y_i}} \quad (2)$$

のようにソフトマックスの多クラス分類器として扱える。それぞれの y_i は出力単語 i の非正規化対数確率であり、

$$y = b + Uh(w_{t-k}, \dots, w_{t+k}; W) \quad (3)$$

として計算される。ここで、 b, U はソフトマックスの媒介変数であり、 h は W から抽出された単語ベクトルの平均もしくは連結によって生成される。図 1 に当手法のフレームワークを示す。

3.2. 文章の分散表現(Paragraph Vector)

本研究で用いる Paragraph Vector は 3.1 節の手法に基づいている。Paragraph Vector では全ての文はユニークなベクトルに写像され、行列 D の列に表される。全ての単語もまたユニークなベクトルに写像され、行列 W の列によって表現される。Paragraph Vector と 3.1 節との唯一の違いは、 h は W と D から生成されることである。これを表したものが図 2 である。

当手法により、Paragraph Vector は単語の意味を保持したまま文章を表現することに成功した。英語で書かれた論文の全文を用いて、類似度順に論文を出力したところ、高い精度を残している[6]。しかし、日本語の論文検索には Paragraph Vector は用いられていない。

4. 事前実験

提案手法の評価を行う前段階として、より高い検索精度を持つ環境の検討を行う。論文のデータセットは GSK 加工版「I-Discover データセット」[7]を用いる。データセットの論文のうち、日本語で表記された論文数は 66,563 件である。Python のモジュールである doc2vec [8]を用いて Paragraph Vector の実装を行った。

4.1. 実験前処理

データセットの論文を加工処理せずに、そのまま学習に用いた場合、実用可能な検索精度は得られなかった。そこで、検索精度向上のためにデータセットに対して以下の処理を行った。

(1) 日本語概要が未登録の論文の排除

データセットの中には題目は表記されているが、日本語概要は表記されていない論文も存在する。概要が登録されていない論文は学習に用いないこととした。

(2) html タグの除去

数式等を表す際に用いられる html タグが論文の概要に存在しているため、タグの除去を行った。

(3) 記号や半角カナ等の正規化

データセットの論文には記号や句読点が半角、全角問わず存在するため、正規化を行った。

(4) 専門用語の表記揺れの解消

研究活動の世界では、様々な専門用語が日々生まれ、古いものにとってかわられる。そこで、専門用語辞書ツール[9]を用いて専門用語の正規化を行った。

4.2. システム構成

事前実験に用いるシステム構成を図 3 に示す。本システムは実験を行う事前処理として、4.1 節で示した実験前処理を行い、Paragraph Vecotrを用いてベクトル空間とデータセットの論文の分散表現を構築しておく。システムの稼働後は、入力として用いられる論文の概要をベクトル表現する。その後、入力論文とデータセットの論文全てのコサイン類似度を計算し、その類似度の降順に関連論文を出力する。

4.3. より高い評価指標を得るための環境の検討

データに実験前処理を行ったところ、ある程度検索精度は向上した。さらに検索精度を向上させるため、事前実験を行い、より高い評価指標を得るための環境の検討を行った。評価指標については次節で述べる。事前実験で影響を評

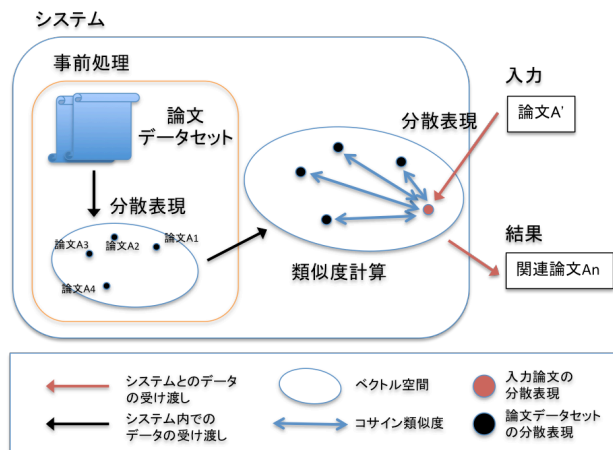


図 3. 事前実験のシステム構成図

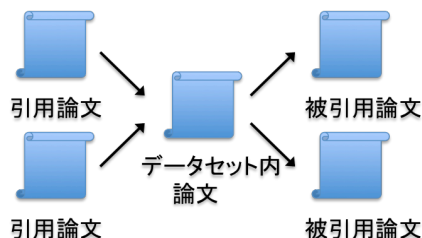


図 4. 論文の引用関係

価する環境の種類は大別して以下の5種類である。

- (1) 形態素解析に用いる辞書を変えた場合
- (2) 引用関係論文の概要を論文概要に付加した場合
- (3) 題目を論文概要に付加した場合
- (4) 学習率を変化させた場合
- (5) 特徴語とその類義語(特徴語群)により論文を限定した場合

4.3.1. 実験条件と評価指標

実験の条件を下記に示す。

評価者：情報工学分野を専攻する学生一名。

入力：データセットに含まれない情報分野の論文 7 件の各概要。なお、評価不可能な状況を排除するために、評価者が知見を持つ分野の論文に限定した。

出力：入力概要に対して類似度の高い論文上位 10 件。

予測モデル：データセットの論文に全て 4.1 節の処理を施した後、学習させる。

評価内容：林らの研究 [3]に基づき、評価内容は適合度を用いて評価する。適合度の基準は表 1 に示す。

ここで、評価指標には、適合度 A のみを適合とみなした評価 $P@10_A$ 及び $MAP@10_A$ と、B 以上を適合とみなした $P@10_{AB}$ 及び $MAP@10_{AB}$ を用いた。ここで、 $P@10$ と $MAP@10$ は以下のように算出する。なお、検索結果数が 10 件に及ばなかった場合は、比較する手法も同じ件数で評価し、不足した部分は不適合とする。

$P@10$ (Precision)：入力した論文に対してそれぞれ得られ

表 1. 適合度の基準[3]

適合度	基準
A (適合)	重要なトピックについて言及しており、参考文献として記載するのにふさわしい。
B(部分適合)	参考文献として記載するのにはふさわしくないが、ある程度参考にはなる。
C(不適合)	内容が異なり、全く参考にならない。

表 2. 辞書を変えた場合の評価結果

	$P@10_{AB}$	$P@10_A$	$MAP@10_{AB}$	$MAP@10_A$
①	0.257	0.0143	0.296	0.00684
②	0.257	0.0429	0.294	0.0447
③	0.214	0.0286	0.248	0.0432

る、検索結果 10 件の適合論文数の平均値である。

$MAP@10$ (Mean Average Precision)：入力した論文に対してそれぞれ得られる、検索結果 10 件の平均適合率 (P)の平均値である。

4.3.2. 環境の検討

上記の評価指標を用いて、環境の検討を行う。各条件下で、高い評価指標を示した項目数が多い環境を本手法に採用する。

(1) 形態素解析に用いる辞書を変えた場合

論文概要に適した形態素解析結果が得られる辞書の選定を行った。

- ① IPA 辞書(今後の環境の雛形とする)
- ② IPA 辞書と Web 上の言語資源を基に作成された辞書(mecab-ipadic-NEologd)の併用
- ③ くだけた表現にも対応可能な JUMAN の辞書

の三種類の評価指標に対する影響を調べた。形態素解析器には MeCab を用いる。なお、事前実験で使用する環境の雛形を①とし、環境を変更しない場合を考慮し、全ての条件と①とを比較する。表 2 に結果を示す。

各評価指標で高い値を得ていることから、②を用いることとした。例えば、「形態素解析」をそれぞれの辞書を用いて分ち書きした場合、

- (ア) 形態素解析
- (イ) 形態素解析
- (ウ) 形態素解析

となるように、②は複合語も高い精度で形態素解析を行うことができる。よって、この評価結果の差は形態素解析の精度の差に起因すると考えられる。

(2) 引用関係論文の概要を論文概要に付加した場合

関連する情報量が多くなり、より精度の高い学習結果が得られる可能性がある。図 4 に示すような

- ④ データセットの論文を引用する論文(引用論文)
- ⑤ データセットの論文から引用される論文(被引用論文)

表 3. 引用関係論文を利用した場合の評価結果

	P@10 _{AB}	P@10 _A	MAP@10 _{AB}	MAP@10 _A
①	0.257	0.0143	0.296	0.00684
④	0.243	0.0143	0.285	0.00302
⑤	0.229	0	0.353	0

表 4. 題目を論文概要に付加した場合の評価結果

	P@10 _{AB}	P@10 _A	MAP@10 _{AB}	MAP@10 _A
①	0.257	0.0143	0.296	0.00684
⑥	0.343	0.0571	0.471	0.00452

を利用した場合の検索精度に関して評価を行った。

以下、引用関係論文と総称する。引用論文は CiNii Articles API を用いて、被引用論文は CiNii Articles の html をパースし、それぞれ取得した。現状では、doc2vec を用いて関連要素を考慮しベクトルで表現する方法は実装されていないため、引用関係論文の概要をデータセットの論文概要に付加することで、学習させた。表 3 に結果を示す。

以上の結果より引用関係論文は用いないこととする。精度の低下は、以下に示すようにデータセットの論文毎の引用関係論文数に大きなばらつきが出たためと考えられる。

被引用論文を持つデータセットの論文の件数：

20,344 件

被引用論文数:83,627 件

引用論文を持つデータセットの論文の件数:9,330 件

引用論文数:20,043 件

(3) 題目を論文概要に付加した場合

題目は良質な特徴語を含んでいるため、より精度の高い学習結果が得られる可能性がある。そこで、

⑥ 論文の題目を概要に付加

についての影響を調べた。結果は表 4 に示す。

論文題目の利用によって、三つの評価指標に良い影響を与えた。よって、論文の分散表現に題目を利用することとする。(2),(3)では論文の概要に付加するという同じ方法で実装したが評価指標への影響が異なった。これは、(2)では引用関係論文数にばらつきがあり量にも大きな差異が見られたが、(3)では題目の量には大きなばらつきが出なかったことに関係すると考えられる。

(4) 学習率を変化させた場合

doc2vec では学習率を減衰させる過程で、学習率が固定され、望ましい結果が得られないという報告がある[8]。そこで、

⑦ 学習率の変化

についての影響を調べた。表 5 にその結果を示す。全ての評価指標に良い影響を与えたことから、学習率に変化を与えることとした。

(5) 特徴語とその類義語(特徴語群)により論文

表 5.学習率を変化させた場合の評価結果

	P@10 _{AB}	P@10 _A	MAP@10 _{AB}	MAP@10 _A
①	0.257	0.0143	0.296	0.00684
⑦	0.386	0.0857	0.381	0.107

表 6. 特徴語群により論文を限定した場合の評価結果

	P@10 _{AB}	P@10 _A	MAP@10 _{AB}	MAP@10 _A
①	0.257	0.0143	0.296	0.00684
⑧	0.443	0.114	0.383	0.0899
⑨	0.386	0.0714	0.362	0.0807

を限定した場合

特徴語による検索手法を併用した場合の影響について調べた。入力論文の概要から MeCab の生起コストを対数化した値を足し合わせ、その値が大きい n 語を特徴語とする。専門用語辞書ツール[9]を用いて特徴語の類義語を収集する。上記の処理を行うことで得られた、特徴語と類義語からなる集合を特徴語群とする。その特徴語群のいずれかが、概要あるいは題目にヒットしたデータセットの論文を類似度順に関連論文として出力する。また、特徴語の表記揺れによって、検索漏れを防ぐために専門用語辞書ツール[9]を用いて、特徴語の類義語もヒットするように実装した。ただし、出力にかかる時間と、特徴語の質を考慮し、

⑧ 重みの大きい上位3単語とその類義語

⑨ 重みの大きい上位4単語とその類義語

を特徴語群とした場合の影響について調査した。結果を表 6 に示す。

この結果より、全ての評価指標において優れた影響を与えた⑧を提案手法に導入する。

これらの結果から、評価指標が最も高い項目数が多い環境の、②、⑥、⑦、⑧を本手法に採用した。

5. 評価実験

4章の結果から、MeCab と mecab-ipadic-NEologd を併用した辞書を形態素解析に用いること、題目の利用、学習率に変化を与えること、特徴語群の利用が評価指標に良い影響を与えたため、本手法に採用した。

5.1. 実験条件

評価実験を行う条件について述べる。下記に記す条件以外は 4.3 節と同様である。

評価者:情報工学分野を専門とする学生 4 名

入力:I-Discover データセットに含まれない論文 10 件の各概要。なお、評価不可能な論文を排除するため、評価者の研究分野に関する論文を抽出した。

出力: 結果 1 は入力概要に対して⑧を用いた場合の類似度の高い論文上位 10 件。結果 2 は入力概要に対して比較手法を用いた場合の類似度の高い論文上位 10 件。

予測モデル: データセットの論文に全て 4.1 節の処理を施した後、②、⑥、⑦を用いて学習させる。

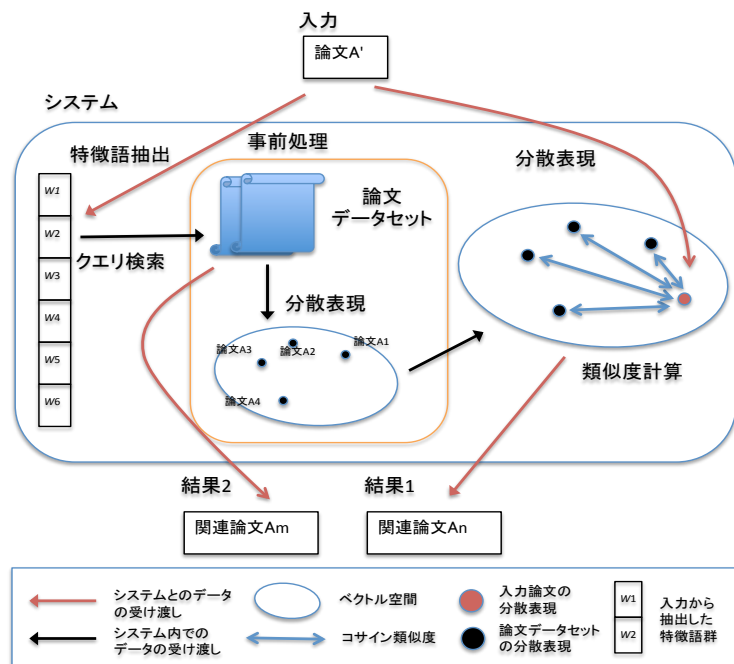


図 5. 評価実験のシステム構成図

5.2. 環境を統一したシステム構成

システムの構成は図 5 に示す。

4 章の結果を踏まえ、提案手法を再実装した。提案手法の有効性を検証するために提案手法と、2 章で示した比較手法を実装した。実験用システムとは異なり、一件の入力で提案手法と比較手法の二種類の検索結果が出力される。これによって、検索に用いるデータセットが全く同じものとなり、データセットの論文がある分野に偏っていたとしても、信頼性の高い評価を行うことができる。また、提示項目やデザインによって検索結果に対するユーザが持つ印象に変化を及ぼさないよう、ユーザインタフェースを統一した。このように統一した一つの環境を用いることで、外的影響を排除し、信頼性の高い評価実験を行うシステムを構築した。

5.3. 実験結果

図 6 に評価者 4 名それぞれの評価結果の平均とそれら全体の平均を示す。

図より、提案手法は全ての評価指標において比較手法よりも高い結果を得たことがわかる。また、 $P@10_A$ と $MAP@10_A$ においては片側 t 検定を行ったところ、有意水準 0.1 で提案手法が有意であることが認められた。

5.4. 考察

評価者 1 と評価者 3 において提案手法が比較手法の精度を下回ることが確認された。彼らの入力はいずれもツイッターと防災に関する論文であった。データセット内にはこれらに関する論文が少なかったことが、学習の精度が上がらなかった理由として考えられる。これはデータセットの論文が様々な分野を網羅するように拡大することで解決できる。ま

表 7. 平均応答時間

従来の応答時間(s)	実装後の応答時間(s)
10.3	0.0351

た、関連論文検索に関する論文を入力とすると、「論文」が特徴語として抽出されてしまう。このように様々な分野の論文で頻繁に用いられる単語が特徴語として抽出されると、論文の限定が効果的でなくなる。これは IDF 等を用いて特定の論文にしか出現しない単語の重要度を上げることによって解決できる。

5.5. 応答時間の短縮

提案手法では、学習率を変動させることによって、予測モデルの読み込みに大きな時間がかかるようになった。類義語の取得も http リクエストによって行うため、時間がかかる。特徴語群により論文の限定においても、類似度の降順に並べ替えた論文の題目や概要を一件ごとにデータベースから取得するため、応答時間の低下が見られた。つまり、提案手法の応答にかかる時間の大部分はコサイン類似度の計算ではなく、

- (1) 予測モデルの読み込み
- (2) 類義語の取得
- (3) データベースへのクエリの発行

の 3 点であることがわかった。そのため、システム稼働前にはあらかじめ、予測モデルの読み込み、類義語辞書の構築、論文データの読み込みを行うことで改善を図り、実装した。従来の応答時間と実装後の応答時間を調べるために、システム内における入力を受け取ってからそれぞれの関連論文 10 件を出力するまでの時間を計測した。この試行を 100 回繰り返す、その平均を算出した。表 7 にその結果を示す。

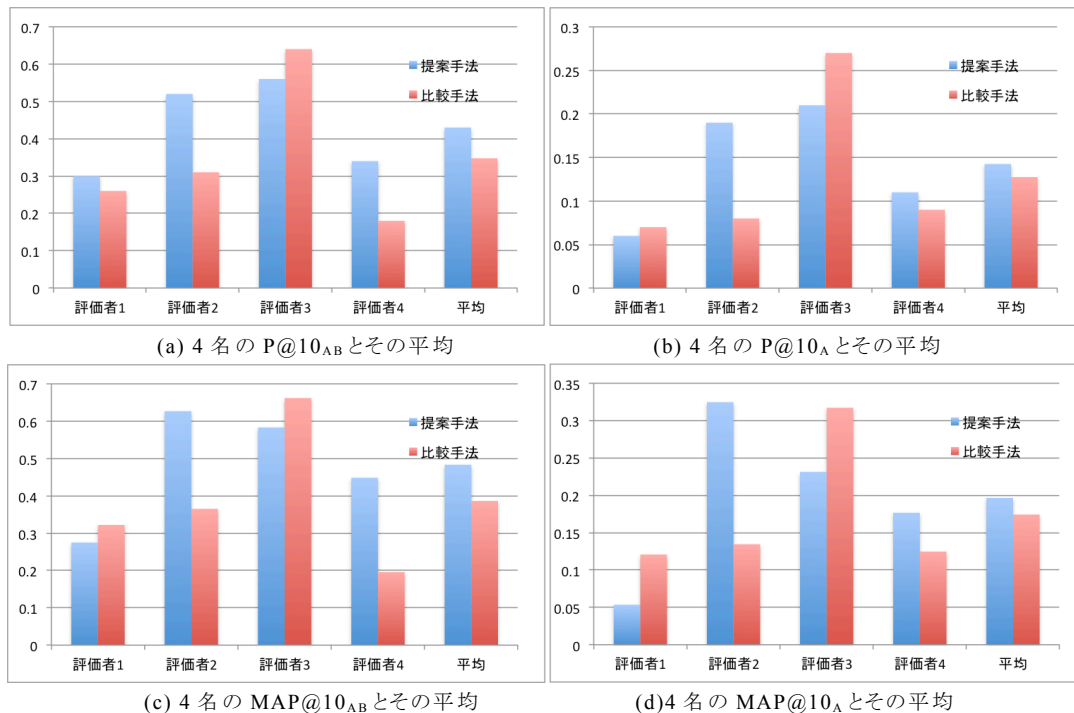


図 6. 評価結果と平均

表7より大幅に応答時間の改善が見られ、速度の面からも十分に実利用可能な手法となった。

6. 結論

本研究では、論文概要の分散表現を用いた関連論文検索の有効性を検証することを目的とした。

分散表現には、感情分析やテキスト分類の実験においてそれぞれ、エラー率を16%と30%程度改善することに成功しており、有効性が認められている、Mikolov らの Paragraph Vector を用いることとした。Paragraph Vector の実装には Python モジュールである doc2vec を用いた。関連論文検索の精度を上げ、高い評価指標を得る環境の選定を行うために事前実験を行うこととした。論文のデータセットには「I-Scover データセット」を用い、日本語概要が未登録の論文は排除した。学習結果が散在してしまうことを防ぐため、事前処理として html タグの除去、記号や半角カナの正規化、専門用語の表記揺れの解消を行った。事前実験の中では、評価指標が最も高い項目数が多い環境を用いることとし、MeCab と mecab-ipadic-NEologd を併用した辞書を形態素解析に用いること、題目の利用、学習率に変化を与えること、特徴語群の利用が評価指標に良い影響を与えたため、本手法に採用した。

評価実験においては比較手法との比較の際に外的影響を排除するため、ユーザインタフェースと検索データセットを統一することによって、評価の信頼性を高めた。その上で、比較手法との比較を行った結果、 $P@10_{AB}$ 、 $P@10_A$ 、 $MAP@10_{AB}$ 、 $MAP@10_A$ においてそれぞれ精度の向上が見られ、特に $P@10_A$ と $MAP@10_A$ においては片側 t 検定を行

ったところ、有意水準 0.1 で提案手法が有意であり、提案手法の有効性を確認した。また、特徴語の類義語と分散表現における予測モデル、データセットをあらかじめメモリ上に保持することで、実利用に十分に耐えうる応答時間を実現した。データセットの拡大や特徴語群の精度を高めることによって、今後も改良の余地がある。

文 献

- [1] Google: "Google Scholar", <https://scholar.google.co.jp/>, 2016.1.25 訪問.
- [2] CiNii: "CiNii Articles", <http://ci.nii.ac.jp/>, 2016.1.25 訪問.
- [3] 林佑磨: "単語の意味概念行列を用いたキーワード生成による関連論文検索システム", 研究報告データベースシステム (DBS), 2014-DBS-159(10), pp.1-6, 2014.07.25.
- [4] 高久雅生, 江草由佳: "簡易類似文書検索手法「ふわっと関連検索」の予備的評価と分析", 情報処理学会研究報告, Vol. 2010, No. 14, pp.1-6, 2010.
- [5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean: "Distributed Representations of Sentences and Document", Computation and Language, 2013.10.16
- [6] Andrew M. dai: "Document Embedding with Paragraph Vector", Computation and Language, 2015.6.29
- [7] GSK 加工版「I-Scover データセット」, <http://www.gsk.or.jp/catalog/gsk2014-b/>, 2016.1.25 訪問.
- [8] gensim, doc2vec, <https://radimrehurek.com/gensim/models/doc2vec.html>, 2016.1.25 訪問.
- [9] 槇俊考, 若原俊彦: "I-Scover および Wikipedia を用いた専門用語辞書ツール構築", 電子情報通信学会技術研究報告 115(138), pp.13-17, 2015.07.13