論文の構成要素を考慮した分散表現に基づく類似論文検索

小林 雄太　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{kobayashi.yuta.kp1, matsu}@is.naist.jp

1 はじめに

科学論文の急激な増加に伴う論文検索・推薦技術の需要の高まりを受け，論文の談話構造に基づきテキストや引用文脈を分類する研究が行われている [4, 9, 6]. これらの研究では，分類結果を論文検索や要約等のタスクに応用しており，テキストと引用文脈どちらも論文の基本的な**構成要素 (目的・手法・結果)** に基づいた分類を行っている．本研究では，これらに共通する論文の構成要素 (目的・手法・結果) に着目し，その情報を論文のベクトル分散表現に反映させることで,「目的は異なるが手法は似た論文」といった，構成要素に基づく論文検索の実現を目指す．

上記の検索を実現するためには，**(1) テキストの構成要素の抽出，(2) 引用の構成要素の抽出，(3) 類似論文の評価タスク**の 3 点が問題となる．(1) に関しては，セクション単位の検索の実現のため，各セクションをルールによって論文中の機能 (section functionality) 毎に分類する Section tagger が提案されており，今後の課題として機械学習ベースの手法が望まれている [4]．また (2) に関しては，引用グラフを利用し論文の分散表現学習を行った LINE[8] や，Teufel らが提案した citation function [9] による引用文脈の分類が行われている．(3) については，類似理由が付与された類似論文のデータセットが評価に必要となるが，従来の論文推薦タスク [7] と異なり人手による評価データ構築が困難なため，自動収集可能なデータによるタスクが望ましい．本研究では (1)(2) を考慮し，構成要素（目的・手法・結果）に基づく類似論文検索のための，テキストと引用グラフの両方の情報を用いた新しい論文の分散表現学習の手法を提案した．また，(3) を満たす新評価タスクとして列挙共引用予測タスクを提案し，論文ベクトルの評価を行った．

1.1 貢献

本研究の貢献を以下に示す．

- 論文のテキストと引用グラフの両方の情報を用いた分散表現学習手法を提案した．また，論文の構成要素の情報 (section functionality，citation function) を反映した分散表現が，類似論文検索タスクに有効であることを確認した．
- 構成要素（目的・手法・結果）に基づく類似論文検索のための，NLP 分野の論文を対象とした類似論文データセットを作成し，新評価タスクとして列挙共引用予測タスクを提案した．

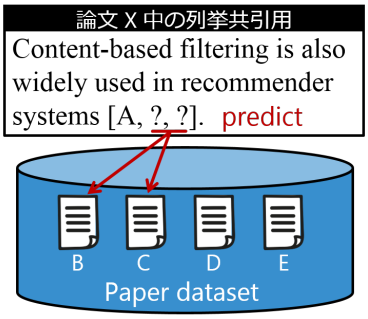

図 1: 列挙共引用予測 (ECCP)

2 提案タスク：列挙共引用予測

本節では，類似論文検索のための列挙共引用予測タスクと評価データセットについて述べる．論文の類似性を扱った先行研究として，Eto の共引用解析の研究が挙げられる [1]．Eto は共引用関係にある論文の中でも，引用箇所が近いほど類似度が高くなり，特に同一引用文脈で引用が隣接した列挙共引用において，最も類似度が高くなることを示した．この研究を踏まえ，我々は類似論文検索のため新評価タスクとして，**列挙共引用予測：Enumerated Co-Citation Prediction (ECCP)** を提案する．

2.1 タスク

以下に ECCP の定義を述べる．まず，システムに論文集合 S，列挙共引用の文脈 c，文脈 c で引用される先頭の論文 p_1 が与えられる．システムは隠された論文 $p_2, \ldots, p_n$（文脈 c において論文 p_1 と列挙共引用される論文）を S から選ぶことを要求される．実際には，システムは隠された論文のトップ k の候補を出力するため，論文の数 n は与えられず，また決定する必要もない．

表 1: ECCP データセットの項目と統計量

項目	統計量
論文数 (ノード数)	20496
引用数（エッジ数）	259743
平均次数	29.6
列挙共引用の文脈	29.6
列挙共引用の平均論文数	2.38

表 2: 列挙共引用 1,000 例の引用された論文数の分布

論文数	統計量
2	722
3	210
4	53
5	7
6+	3
Total	1000
average	2.38

図 1 に ECCP の略図を示す．図 1 では，論文 X 中で論文 A，B，C が列挙共引用され，論文データセット $S = \{B, C, D, E\}$ が 4 本の論文で構成されている．システムは論文 A (= p_1) とその引用文脈 c (=*"Content-based filtering is also widely used in recommender systems"*) を提示され，論文 B，C (p_2，p_3) を論文データセット S から予測する．システムは (1) 全論文の全テキスト (参考文献除く)，(2) 引用グラフ ($X \rightarrow B$，$X \rightarrow C$ は除く [1]) の 2 つの情報源を利用可能である．

2.2 データセット

本研究では，ACL Anthology[2] から 2016 年までの論文 PDF データをクローリングしデータセットを作成した．テキストの抽出には PDF からの XHTML へ変換を伴い，独自にカスタマイズした Poppler[3] を用いた．また，正規表現により論文中の引用と列挙共引用を検出し，引用グラフを構築した．評価には，1 文中に 1 箇所のみ列挙共引用が出現する引用文脈 1,000 例を使用する．データセットの項目と統計量を表 1 に，1,000 例の列挙共引用文脈における，引用されている論文数のヒストグラムを表 2 に示す．表 2 より，先頭の論文を除けば，列挙共引用 1 事例につき平均 1.38 の隠蔽された論文を予測する．

3 提案手法：論文の表現学習

本節では，類似論文検索のための論文ベクトル学習法について述べる．図 2 に提案手法の概略を示す．手順 1 で前処理を行い，手順 2 3 でテキストの表現を学習し，手順 4 で引用グラフに情報を付与し，手順 5 でテキストと引用の情報を統合し論文の表現を得る．

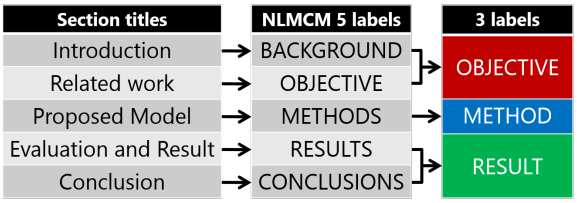

図 3: NLMCM ファイルによるセクションの分類

3.1 手順 1：論文データセットの前処理

まず，英語 Wikipedia コーパスと ACL Anthology から抽出した論文のテキストを組み合わせたコーパスを作成し，このコーパスを用いて分散表現学習ライブラリ fastText[3] により単語ベクトル分散表現の教師なし学習を行う．事前実験において tf-idf や Paragraph Vector[5] に比べ精度の高かった点を考慮し，以後の手順では fastText の分類器を使用する．

3.2 手順 2:構成要素によるセクションの分類

ここでは，論文の各セクションを構成要素 (目的・手法・結果) の 3 ラベルに分類する．そのために，National Library of Medicine Category Mappings (NLMCM)[3] ファイルを使用する．このデータは，医療系論文の構造化アブストラクト (論文の簡潔な要約) に使用された 3,032 のセクションタイトルを，broader NLM Categories の 5 ラベルに対応付けたデータセットである．図 3 に分類の略図を示す．まず，上記の NLMCM により論文の各セクションタイトルが NLMCM 中のものと完全マッチするものについて，5 ラベルへの分類を行い，その後 3 ラベルへとマージすることで分類を行う．NLMCM に存在しないセクションタイトルは，ルールで分類されたセクションの本文テキストを教師データとして教師ありで訓練した fastText によって分類される．この際，入力はセクションの本文テキスト，出力は 3 ラベルのいずれかとなる．

また，セクション分類器の有効性を検証するため，手順 2 でラベリングされた 72,721 のセクションタイトルに対し 5 分割交差検定を行った．単語ベクトルの次元数は 100，n-gram は 3，ウィンドウサイズ及びネガティブサンプリング数は 5 とした．NLMCM のルールでラベル付けされたセクションを，その本文テキストから fastText を用いて分類した結果，**96.1** ポイントの高い分類精度を確認した．分類器の精度確認後，ルールでラベル付けされなかった 11,118 セクションについて，セクションタイトルと本文を結合したテキストを入力としてラベル付与を行った．

[1] 過度に容易なタスクにしないための設定である．引用を除外しない場合，システムは隠された論文を参考文献の論文から選択すればよいため，本研究が目指す類似論文検索の設定にそぐわない．

[2] http://aclweb.org/anthology/

[3] https://poppler.freedesktop.org/

[3] https://structuredabstracts.nlm.nih.gov/

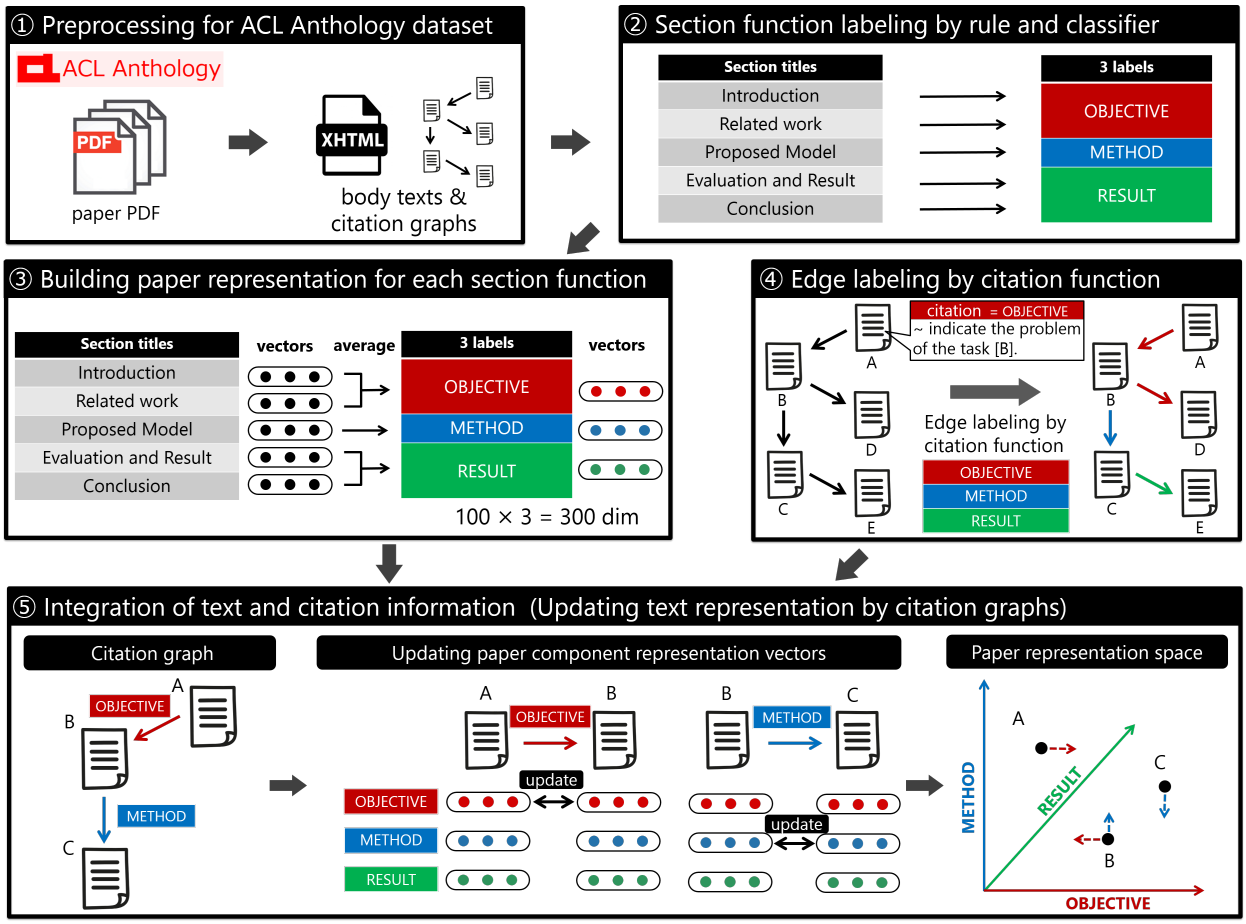

図 2: 類似論文検索のための論文ベクトル学習法

3.3 手順 3：論文のベクトル表現の構築

まず，手順 2 で教師あり学習した fastText の 100 次元の単語ベクトルにより，各セクションのベクトルを各単語ベクトルの平均として計算する．次に，各セクションのベクトルを目的・手法・結果毎に平均し，最終的に 100 次元 × 3 = 300 次元のベクトルとして，これを論文のベクトル表現とする．

3.4 手順 4：グラフへのラベル付与

引用グラフの各エッジに，引用文脈に基づくラベル付与を行う．具体的には，引用理由を表す citation function [9] のラベルを付与する．引用機能のアノテーションスキームには様々なものが存在するが，本研究では Nanba らのスキーム [6] を使用し，引用のエッジを 3 種類 (OBJECTIVE, METHOD, RESULT) に分類する．また，NLP 分野の論文の各引用文脈にラベルがアノテーションされた，Teufel らの Citation Function Corpus[9] 及び，CL-SciSumm-2017 SharedTask [2] のコーパスを組み合わせたデータセットを使用した．citation function の予測には教師ありの fastText 分類器を使用し，引用文脈を入力として 3 種類のラベルを出力する．citation function のアノテーションデータセット合計 1,618 事例のラベル分布は，OBJECTIVE:200, METHOD:1,214，RESULT:204 であった．fastText による引用機能ラベル予測の 5 分割交差検定の結果，**90.1** ポイントの高い精度が得られた．

3.5 手順 5：テキストとグラフの情報の統合

ラベル付与されたグラフに対し，グラフの分散表現手法である LINE [8] を用いて，手順 3 で得られたテキストの論文ベクトルを初期値として各構成要素 (100 次元) 毎に更新を行う．LINE は引用グラフ上でランダムウォークを行い，引用を 2 回辿った経路までのノードを近傍とし，それらのノードのベクトルの内積が大きくなるよう学習を行う．本研究ではグラフに 3 種類のラベルが付与されているため，このラベルに沿って論文のベクトル分散表現を学習させる．そのため，従来の引用グラフ上のランダムウォークと異なり，3 種類の各ラベルのグラフでランダムウォークを行い，各 100 次元ベクトルの内積が大きくなるよう更新を行う．

図 2 手順 5 において，論文 A から論文 B への引用に OBJECTIVE の引用機能ラベルが振られているため，OBJECTIVE のランダムウォークにより論文 A，B は近傍関係となる．したがって，それぞれの OBJECTIVE

の 100 次元ベクトルの内積が大きくなるように更新される．同様に論文 B から C への引用では METHOD の 100 次元ベクトルが更新される．手順 3 のベクトルを初期値としているため，次元数は 100，ネガティブサンプリングのパラメータは 5 とした．以上の手順により，類似論文検索のための構成要素を考慮した最終的な論文のベクトル分散表現が得られる．

4 実験

4.1 論文の表現ベクトルの評価

提案手法によって得られた論文の表現ベクトルと従来の表現ベクトル（fastText，LINE）について，ECCP により比較を行う．具体的には，表 1 で示した 20,496 の論文データセットに対して，1,000 事例分の ECCP を行う．各事例において列挙共引用の先頭の論文ベクトルからコサイン類似度を基に全論文をランキングし，上位 100 位に正解の論文が含まれた数とその順位で評価を行った．評価指標として，順位付け問題の精度評価指標である nDCG を用いた．

4.2 ベースライン

ベースラインの表現ベクトルとして 2 種類のベクトルを用意した．3.1 節で得られた教師なしの fastText [3] を用いて，論文を全単語の平均ベクトルとしたものをテキストベースのベースラインとした．また，通常の引用グラフにおいて LINE [8] で学習したベクトルをグラフベースのベースラインとした．論文の各構成要素を持たないため，両ベクトルの次元は 300 とした．

4.3 提案手法

提案手法は 2 種類のベクトルを用意した．1 つは section functionality を予測するよう fastText を用いて教師あり学習を行った，3.3 節のテキストベースの表現ベクトル (提案手法 1 : fastText + secfunc)．2 つ目は提案手法 1 のベクトルを初期値とし，グラフの分散表現手法である LINE を用いて，引用グラフの citation function ラベルに基づいて更新を行った，3.5 節の表現ベクトルである (提案手法 2 : 提案手法 1 + LINE + cifunc)．各論文の表現ベクトルは OBJECTIVE, METHODS, RESULT 各 100 次元のベクトルで表現されており合計 300 次元のベクトルを持つ．最終的には，3.4 節で学習した分類器により入力の引用文脈から citation function ラベルを予測し，そのラベルの 100 次元を用いて論文間のコサイン類似度を測り全論文のランキングを行った．

4.4 実験結果

表 4.4 に論文の表現ベクトルを用いた ECCP の結果を示す．論文のテキストの構成要素の情報を使用した提案手法 1 がベースラインに比べ精度が改善していることが確認できる．さらにテキスト情報に加え引用グラフの情報を付け加えた提案手法 2 はさらに精度の改善が認められた．これらの結果から，構成要素を考慮した，テキストと引用グラフの情報が論文の類似度計算に有用であることが確認できた．

表 3: 列挙共引用予測タスクの結果

論文のベクトル表現	nDCG
fastText [3]	0.44
LINE [8]	0.46
提案手法 1 (fastText + secfunc)	0.51
提案手法 2 (提案手法 1 + LINE + cifunc)	**0.58**

5 おわりに

本研究では構成要素（目的・手法・結果）に基づく類似論文検索のための自動評価タスクとして，類似論文検索のためのデータセットを独自に作成し，列挙共引用予測タスクを提案した．また，論文のテキストと引用グラフの両方の情報を用いた論文の分散表現学習手法を提案し，評価タスクにおいて既存の分散表現に比べ優れていることを確認した．今後の課題として，引用文脈の論文中での位置の利用が挙げられる．

参考文献

[1] Masaki Eto. Spread co-citation relationship as a measure for document retrieval. In *Proc. ACM workshop on Research advances in large digital book repositories and complementary media*, pp. 7–8, 2012.

[2] Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. Insights from cl-scisumm 2016: the faceted scientific document summarization shared task. *International Journal on Digital Libraries*, pp. 1–9, 2017.

[3] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[4] Şenay Kafkas, Xingjun Pi, Nikos Marinos, Andrew Morrison, Johanna R McEntyre, et al. Section level search functionality in europe pmc. *Journal of biomedical semantics*, Vol. 6, No. 1, p. 7, 2015.

[5] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proc. ICML*, pp. 1188–1196, 2014.

[6] Hidetsugu Nanba and Manabu Okumura. Towards multi-paper summarization using reference information. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 926–931, 1999.

[7] Kazunari Sugiyama and Min-Yen Kan. A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries*, Vol. 16, No. 2, pp. 91–109.

[8] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proc. WWW*, pp. 1067–1077, 2015.

[9] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proc. EMNLP*, pp. 103–110, 2006.