複数論文概要の解析による特定分野の技術動向分析

難波 英嗣

広島市立大学大学院情報科学研究科 〒731-3194 広島市安佐南区大塚東 3-4-1

E-mail: nanba@hiroshima-cu.ac.jp

あらまし 本研究では，ある特定分野における複数の学術論文を解析することにより，その技術動向を俯瞰するシステムを提案する．まず，ある分野の論文を収集し，次に，それらの概要から，背景・目的・手法・効果等を抽出し，最後に，論文間で対比的な箇所を自動同定し可視化することにより，その分野の技術動向の把握を容易にすることを目指す．本研究では，3 種類のモジュール：(1)論文分類モジュール，(2)論文概要解析モジュール，(3)対比個所同定モジュールを作成する．これらのモジュールのうち，論文構造解析モジュールの有効性を調べるため，実験を行った．実験の結果，提案手法により日本語論文では再現率 0.862，精度 0.830，英語論文では再現率 0.591，精度 0.627 が得られ，提案手法の有効性が確認できた．

キーワード 技術動向分析, 文書構造解析, 学術論文

1. はじめに

近年，研究者数の増加，学問分野の専門分化と共に学術情報量が爆発的に増加している．また，研究者が入手できる文献の量も増える一方であり，人間の処理能力の限界から，入手した文献全てに目を通し利用することが益々困難になってきている．

このような状況で必要とされるのは，特定の研究分野に関連した情報が整理，統合された文書，すなわちサーベイ論文や専門図書である．サーベイ論文や専門図書を利用することで，特定分野の研究動向を短時間で把握することが可能になる．しかし，論文全体に対するサーベイ論文の占める割合が極端に少ないという指摘がある[1]．そこで本研究では，ある分野における複数の学術論文を解析することにより，その技術動向を俯瞰するシステムを提案する．

本論文の構成は以下の通りである．次節では，論文等の技術文書を対象とした技術動向分析に関する関連研究について述べる．3 節では，複数論文概要の解析により特定分野の技術動向を分析する手法を提案する．提案手法の有効性を調べるために行った実験について 4 節で報告し，5 節で本稿をまとめる．

2. 関連研究

研究動向の解析に対していくつか先行研究があり，文内の語句の意味役割を手掛かりとしたアプローチが多く提案されている．Gupta ら[2]は，研究アイディアの発展過程を調べるために，「FOCUS」「TECHNIQUE」「DOMAIN」という 3 種類のカテゴリに該当する語句を論文概要から自動的に抽出する手法を提案した．彼女らの手法はパターンマッチに基づいており，例えば，動詞「propose」の直後に出現する直接目的語を「FOCUS」を表す語句として抽出している．

Tateishi ら[3]は，論文中に出現するモノとモノの意味関係を同定するための手法を提案した．まず，エンティティを示す語句に対して「TEAM」「OBJECT」「MEASURE」のいずれかのタグを付与し，エンティティ間に「PERFORM (動作主体)」や「CONDITION (実験条件)」など 16 種類の有向関係を付与したタグ付けコーパスを構築した．そして，作成したコーパスを用いて SVM による関係抽出器を作成し，その結果に基づいて論文中の文を解析している．

難波ら[4]は，ある研究分野において，「どのような要素技術がいつ頃から使われているのか」という情報を網羅的に収集するために，論文表題から「RESTRICT」「GOAL」「METHOD」に対応する語句を抽出する手法を提案した．難波らは，各構造タグに対応する手掛かり語表現をいくつか用意し，表題中で一致する手掛かり語を構造タグに置き換えることで解析を行っている．

技術動向分析に関するこのほかの研究として，第 8 回 NTCIR ワークショップ・特許マイニングタスク[5]で実施された技術動向マップ作成サブタスクがある．これは，「要素技術」と「その効果」という観点から，論文と特許を分類した技術動向マップを作成することを目的とした研究プロジェクトである．このようなマップを作成するツールは，先行技術調査や無効資料調査の支援ツールとして利用することができる．そして，技術動向マップを自動的に作成するために，技術動向マップ作成サブタスクでは，論文や特許から要素技術とその効果を表す表現を自動的に抽出するという課題を設定している．

福田ら[6]は，技術動向マップ作成サブタスクにおいて，機械学習に基づいた手法により，日本語論文および日本語特許から要素技術とその効果に関する表現を自動的に抽出している.そして,「論文の表題と概要に，要素技術とその効果を示すタグを付与する」という系列ラベリング問題として捉え，SVM を用いて，以下に示すタグの自動付与を行っている.

- TECHNOLOGY: 要素技術(新しく提案された技術，既存技術を応用した技術，研究課題を解決するための手段など(例: 協調フィルタリング，SVM))
- EFFECT: 効果(新しい機能の追加，新しく得られた物質，精度などの数値または増加・減少など). EFFECT タグには，ATTRIBUTE タグと VALUE タグを含む.
- ATTRIBUTE, VALUE: 「速度(ATTRIBUTE)が向上(VALUE)」のように，要素技術に対する効果は「属性(ATTRIBUTE)」と「属性値(VALUE)」の対で表現する.

以下に，「英語の単名詞句とその他の句の同定問題に SVM を適用し，実際のタグ付けデータを用いて解析を行ったところ，従来手法に比べて高い精度を示した」という論文概要にタグを付与した例を図 1 に示す.

英語の単名詞句とその他の句の同定問題に<TECHNOLOGY>SVM</TECHNOLOGY>を適用し，実際のタグ付けデータを用いて解析を行ったところ，従来手法に比べて <EFFECT> <VALUE> 高い </VALUE><ATTRIBUTE> 精度 </ATTRIBUTE></EFFECT>を示した.

図 1 論文概要へ TECHNOLOGY, EFFECT, ATTRIBUTE, VALUE タグの付与例

この手法は，次節で提案する技術動向分析システムのモジュールのひとつとして利用する.

3. 提案手法

3.1 システム概要

本研究で開発するシステムは以下の 3 つのモジュールから構成される.

1. 論文分類モジュール
2. 論文概要解析モジュール
3. 対比個所同定モジュール

各モジュールについて，3.2 節，3.3 節，3.4 節でそれぞれ述べる.

3.2 論文分類モジュール

論文分類モジュールでは，論文データベース中のすべての論文に対し，文書分類器を用いて分野ごとに分類する．我々の過去の研究で，学術論文を科研費カテゴリ[7]や国際特許分類[8]に自動分類するシステムを構築している．これらのシステムを論文分類モジュールとして利用する.

3.3 論文概要解析モジュール

論文概要解析モジュールでは，2 種類の方法で日本語および英語論文概要の構造を解析する．第一の方法では，論文概要の各文に，background(背景)，purpose(目的)，method(手法)，result(結果)のいずれかのタグを付与する．図 1 は，論文[9]の概要の各文に対し，タグを付与した例である.

<purpose>本研究は，自動車に乗って移動する際の車内会話を記録・提示することで，人と街の間に埋め込まれたタイムリーな知識を流通させることを目的とする.</purpose><background>自動車に乗り合わせた人同士は，走行中の場所やその近辺について様々な会話をする．</background><background>そのような会話は，その場限りの一過性のものではあるが，人がその時・その場にいることによって生起されることが多く，その場所や季節，時間帯と強く結び付いたタイムリーな情報が含まれる．</background><background>これらは運転者に有用な気づきをもたらすだけでなく，街に関する知識ベースの構築を助け，街に関するニーズや機会の理解とそれを応用したシステムへの展開が期待される．</background><method>本論文では，車内会話のようなタイムリーな情報の流通システムを実現するための基礎研究として,およそ 10 カ月にわたり車内会話を収集し，分析・分類したのでその結果を報告する.</method>

図 2 論文概要の構造の例([10]より引用)

論文概要の構造を解析する第二の方法では，2 節で述べた福田ら[6]の手法を用いて，TECHNOLOGY，EFFECT，ATTRIBUTE，VALUE タグを付与する.

3.4 対比個所同定モジュール

対比個所同定モジュールでは，3.2 節で述べた論文分類モジュールを用いて同じカテゴリに分類された論文集合に対し，概要中の文レベルで対比関係にある個所を同定する．この同定処理は，次の 2 つの手順で実現する.

1. 要素技術用語リストの自動作成
2. リスト中の要素技術用語を含んだ 2 文の抽出

手順 1 については，例えば“SVM”などの要素技術用語を入力とし，word2vec[11]を用いて，入力用語と関連する他の要素技術用語を自動収集する．手順 2 では，論文分類モジュールを用いて収集された論文集合の各文に対し，論文概要解析モジュールで “method”が付与された文において，手順 1 で作成された用語を含んだものがあれば，それらをすべて出力する．この結果，例えば，以下のような 2 文の対が対応個所として抽出される．ここで，この 2 文に付与されている TECHNOLOGY タグは，2 節で述べた福田ら[6]の手法で自動的に付与されたものである．また，文頭の 4 桁の数値は，抽出元論文の著作年を示している．

- 2008: また ,<TECHNOLOGY> Support Vector Machines(SVM)</TECHNOLOGY>を用いて,フレーズベース統計的機械翻訳モデルにより生成したフレーズテーブルから得た訳語候補を検証し,信頼度の低いものを排除した.
- 2010: 提案手法では,翻訳知識源が異なる 2 種類の訳語推定手法により推定された訳語候補のうち信頼度の低い訳語候補を絞り込むタスクに対して ,<TECHNOLOGY> Support Vector Machines(SVM)</TECHNOLOGY>を適用する.

このような情報を，論文の著作年順に時系列に並べれば，ある分野の技術動向が概観できると考えられる．

4. 実験

3.3 節で述べた論文構造解析モジュールの有効性を調べるため，実験を行った．

4.1 実験条件

実験データ

日本語論文に関しては，Yamamoto ら[9]が構築した論文概要における役割推定コーパス[1]を用いた．このコーパスは，1,000 論文の概要の各文に，3 人の被験者が background(背景)，purpose(目的)，method(手法)，result(結果)のいずれかのタグを付与している．本研究では，3 人全員が同じタグを付与した 5,553 文を実験に用いた．

英語論文については，Willot ら[12]が構築したものを用いた．このコーパスは，自然言語処理分野 100 論文の概要に含まれる計 662 文に，上述の Yamamoto らのものと同じタグを付与したものである．

分類手法

各文への役割の付与には Joulin ら[13]が提案している fastText を用いた．fastText は入力層，隠れ層，出力層の 3 層からなる DNN(Deep Neural Network)である．概要の各文を，MeCab[2]を用いて分かち書きにし，入力を(1)単語, (2)単語バイグラム, (3)単語トライグラム, (4)単語 4 グラムとした場合で学習および評価を行った．なお，単語の分散表現は 100 次元とした．

評価尺度

再現率および精度で評価した．

4.2 実験結果

日本語論文概要を対象に実験結果を表 1 に，英語論文概要を対象にした結果を表 2 に，それぞれ示す．表より，日本語論文では単語トライグラムの時が，英語論文では単語バイグラムの時が，再現率，精度共に最も良い結果が得られた．

表 1　日本語論文概要の各文への役割の付与結果

手法	再現率	精度
単語	0.804	0.774
単語バイグラム	0.850	0.818
単語トライグラム	**0.862**	**0.830**
単語 4 グラム	0.856	0.823

表 2　英語論文概要の各文への役割の付与結果

手法	再現率	精度
単語	0.584	0.620
単語バイグラム	**0.591**	**0.627**
単語トライグラム	0.590	0.625
単語 4 グラム	0.584	0.619

英語論文概要の解析に関して，Willot らは，論文概要中の各文に付与されるラベルには，例えば「目的」→「手法」→「結果」のようにパターンがあると考え，このタスクを系列ラベリング問題と捉え，リカレントニューラルネットワークの一種である Long short-term memory (LSTM)を用いてラベルの自動付与を実現している．実験の結果，Willot らは，再現率 0.650，精度 0.619 を得ている．本研究では，LSTM を用いない単純な手法で，精度は若干上回っているが，再現率には約 0.06 の差がある．この結果から，日本語論文概要についても，LSTM を用いることでタグ付与精度がさらに向上する可能性があると考えられる．

5. おわりに

本研究では，論文概要の構造を，2 種類の方法を用

[1] http://nlp.inf.kyushu-u.ac.jp/resource.html

[2] http://taku910.github.io/mecab/

いて解析し，同一分野の対比個所を自動的に対応付けることにより，技術動向分析を可能にする手法を提案した．

謝辞

本研究の一部は科学研究費補助金(基盤研究(A))(研究課題番号:15H017214)の支援を受けて行われた．論文データは，国立情報学研究所から提供いただいた．ここに記して感謝の意を表する．

参 考 文 献

[1] William D. Garvey 著,/ 津田良成監訳, コミュニケーション –科学の本質と図書館員の役割-, 敬文堂, 1979.

[2] Sonal Gupta and Christopher D. Manning, Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers, *In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2011.

[3] Yuka Tateishi, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa, Annotation of Computer Science Papers for Semantic Relation Extraction, *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp.26-31, 2014.

[4] 難波英嗣，谷口裕子，学術論文データベースからの研究動向情報の抽出と可視化，言語処理学会第12回年次大会 併設ワークショップ－言語処理と情報可視化の接点－，pp.35-38, 2006.

[5] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto, Overview of the Patent Mining Task at the NTCIR-8 Workshop. *In Proceedings of the 8th NTCIR Workshop Meeting*, pp. 293-302, 2010.

[6] 福田悟志，難波英嗣，竹澤寿幸，論文と特許からの技術動向情報の抽出と可視化，情報処理学会論文誌データベース，Vol. 6, No. 2, pp.16-29, 2013.

[7] 福田悟志，難波英嗣，竹澤寿幸，要素技術とその効果を用いた学術論文の自動分類，日本図書館情報学会誌，Vol.63，No.3, pp.145-162，2016.

[8] 難波 英嗣，竹澤 寿幸，2 種類の翻訳システムを用いた学術論文の特許分類体系への自動分類，情報処理学会論文誌データベース，Vol.2，No.3，pp.76-86，2009.

[9] Takafumi Yamamoto and Yoichi Tomiura, Constructing Corpus of Scientific Abstracts Annotated with Sentence Roles, *In Proceedings of the 5th International Congress on Advanced Applied Informatics*, 2016.

[10] 松村耕平，角康之，自動車内における会話と場所の関連性の分析：タイムリーな情報の流通に向けて，情報処理学会論文誌，Vol.56，No.4，pp.1258-1268，2015.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, Distributed Representations of Words and Phrases and their Compositionality, *In Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013.

[12] Paul Willot, Kazuhiro Hattori, and Akiko Aizawa, Extracting Structure from Scientific Abstracts using Neural Networks, Digital Libraries: Providing Quality Information (17th Asian Digital Library Conference, ICADL 2015), pp.329-330, 2015.

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, Bag of Tricks for Efficient Text Classification. *arXiv Preprint arXiv*:1607.01759, 2016.