サーベイ支援のための学術論文の内容可視化

桃井 凌* 杉本 徹**
*芝浦工業大学大学院 理工学研究科 **芝浦工業大学 工学部
{ma17118, sugimoto}@shibaura-it.ac.jp

1 はじめに

研究者や学生の研究活動において，既存の論文を読んで先行研究や関連研究の調査を行うことは欠かせない作業である．しかし，読むべき論文を探し，内容を理解する一連の作業には多大な労力を要するため，それを支援する研究が多く行われている．論文を読む際に，その大まかな内容を掴むための作業として，概要や序論，結論の節から読み始めることや，付与されたキーワードリストを見ることが挙げられる．しかし，それらは論文の形式や著者の裁量により情報量が様々であり，統一的な規則に基づいた形式で表現することは，研究の理解や他の論文との比較において有用であると考えられる．

そこで本研究では，論文の提案手法を，キーワードをノードとし，キーワード間の関係をエッジとしたグラフの形で表現することで要約，可視化する手法を提案する． これにより，文章よりも直感的に理解可能で，キーワードのみを提示するよりも情報を含んだ要約を提示し，研究者のサーベイ作業を支援することを目指す．

2 関連研究

学術論文中の用語に対して属性を付与する研究として，Sonal ら [1] の研究がある．この研究では，(1) 最たる貢献を表す FOCUS，(2) 使用した手法やツールを表す TECHNIQUE，(3) 応用分野を表す DOMAIN の 3 種を論文の概要から抽出することを試みている．抽出には依存構造木に基づくパターンマッチングを用いる他，ブートストラップ法によるパターンの学習も行っている．

建石ら [2] の研究では，オントロジーに基づいた用語とその関係の意味クラスを定義したアノテーションスキーマを作成し，それに基づいてアノテーションを行っている．また，アノテートされた用語のクラスや関係を素性に用い，Sonal らの定義した 3 要素を抽出する手法を提案している．

また，サーベイを支援する研究としては，提案手法や研究の位置付けなどの包括的な要約文を生成する Shin ら [3] の研究や，実験結果の要約に力を入れた Eunsol ら [4] の研究など，多くのアプローチが存在する．

3 提案手法

まず，論文の題目および概要からキーワード候補を抽出する．次に，それらの種類を Support Vector Machine (SVM) を用いて分類する．さらに，分類されたキーワードのペアを作成し，同じく SVM を用いて関係の分類を行う．最後に，抽出された関係に基づいてキーワード同士を繋げることで，グラフを生成する．

3.1 使用するデータとアノテーション

言語処理学会論文誌 LaTeX コーパス [1] を用いる．これは会誌「自然言語処理」に掲載された論文の LaTeX のソースファイルである．一般に，Web 上で公開されている論文の多くは pdf 形式であり，プログラムで処理するにあたりテキストファイルに変換する必要がある．この際に，題目や概要，序論といった論文の構造の情報が失われるという問題がある．このコーパスは論文の文書構造を保持した形式で提供されており，本研究で扱うのに適している．

本研究で提案するグラフのノードとエッジに対応する論文中のキーワードの種類 (クラス) とキーワード間の関係の種類の定義をそれぞれ表 1，表 2 に示す．アノテーションはコーパス中の論文のうち，日本語論文の中からランダムに選出した 100 件の論文について行った．

3.2 分析対象文の選出

概要文の多くは，(1) 初めに一般知識や現状などの研究背景について述べ，(2) 次にその研究で行うことや提案する手法を述べ，(3) 最後に結果について述べる構造をとる．このうち本研究で目標とするグラフの

[1] http://www.anlp.jp/resources/journal_latex/

表 1: キーワードのクラスの定義

クラス	説明	例
データ	データ，システムに入力するもの	コーパス/単語ベクトル
手法/処理	データに対して行う動作	機械学習/クラスタリング
目的/結果	出力されるもの，その研究で達成する目標	文書分類/専門用語の抽出

表 2: キーワード間の関係の定義

関係	説明	例
抽出/変換	データから別のデータへの変換，抽出	入力文→名詞
入力/使用	データを入力とした処理	単語ベクトル→ SVM
出力/結果	データや処理に基づいた出力	SVM →専門用語抽出
処理の流れ	ある処理から次の処理への移行	次元削減→機械学習
同義	表記が違うが意味，使われ方が同じもの	重要文抽出ー重要文の抽出

生成を行う上で，(2) にあたる文集合のみを分析対象とするべきだと考えられる．

この判定には手掛かり語の出現をもとにした簡単なパターンマッチを用いる．まず，(2) と (3) について，それぞれ表 3 の手掛かり語のいずれかが初めて出現する文 (以下，手掛かり文とする) を特定する．(2) の手掛かり文から (3) の手掛かり文の前文までを (2) のセクションとし，(3) の手掛かり文以降の全文を (3) のセクションとする．また，(1) は手がかり語が無いため，(2) の手掛かり文の前文までの全文を (1) のセクションとする．また，(2)(3) いずれの手掛かり文も存在しなかった場合は，概要文全文を (2) のセクションとする．次節以降の処理では (2) のセクションのみを対象とする．

表 3: セクションごとの手掛かり語

セクション	手掛かり語
(1) 背景	手掛かり語なし
(2) 提案手法	本研究で/本論文で/本稿で/提案
(3) 結果	実験/結果/評価/精度が/精度に

3.3 キーワード候補の抽出

グラフを構成するキーワードは簡潔なものが相応しいと考えられる．そのため，本研究では「クラスタリング」や「機械学習」のような名詞，複合名詞や,「重要文の抽出」のような表現を抽出するための規則を定める．対象文書を形態素解析器 JUMAN++[2] で形態素解析し，以下のような品詞の組み合わせをキーワード候補として抽出する．

A). 名詞または未定義語が 1 つ以上連続する語
例：「文書ベクトル」,「SVM」

B). 接頭辞，副詞，形容詞かつ活用形が語幹，接尾辞かつ活用形が語幹のいずれかが A と連続する語
例：「重要文」,「自然言語処理」

C). (A または B) + (「の」(助詞) または「を」(助詞)) + (A または B かつ末尾がサ変接続名詞)
例：「専門用語の抽出」

3.4 SVM によるキーワード候補の分類

抽出したキーワード候補を，表 1 のキーワードのクラスのいずれか，あるいは非キーワードという計 4 つのクラスに分類する．分類には SVM を使用し．素性は以下のものを用いる．

- 単語ベクトル： Bag of Words (BoW) と単語の分散表現の 2 種類を用いる．
  - キーワード自身の単語ベクトル．複合語の場合は各構成語の和．
  - 前 3 形態素の和
  - 後 3 形態素 (係り先文節から数える) の和
- 品詞： JUMAN 品詞体系の大分類，小分類
  - キーワード自身の品詞．複合語の場合は各構成語の和．
  - 前 3 形態素の和
  - 後 3 形態素 (係り先文節から数える) の和
- tf-idf 値： 文書集合をコーパス [1] の全論文としてキーワードの tf-idf を算出
- 文字列中のアルファベットの割合
- 文字列中の数字の割合

[2]http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN++/

- 題目に出現するか否か
- 概要に出現するか否か
- 論文のキーワードリストに出現するか否か

3.5 Self-Training によるキーワードの水増し

論文へのアノテーションは人手で行う必要があり，大量の教師データを用意するのは困難である．そこで，半教師あり学習である Self-Training を用いることで，教師データの増加，および抽出精度の向上を試みる．Self-Training は以下のプロセスで行う．

1. アノテーション済みの論文を訓練データとテストデータに分割する．
2. 未アノテーションの論文を N 件ずつ分割する．
3. 訓練データを用いて分類器を作成する．
4. 分類器を用いて，分割された論文集合 1 グループを分類する．
5. 分類の結果，確信度が高いものを訓練データに追加する．(本研究では SVM の分離超平面からの距離を確信度として用いる)
6. 3~5 の処理を，分割された全ての論文集合に対して行う．
7. 作成された分類器を用いてテストデータを分類する．

3.6 SVM による関係の分類

2 つの非キーワード以外に分類されたキーワード候補の組み合わせについて，その関係を SVM を用いて分類する．ただし，キーワードの組は同文中，または前後の文中に出現するもの同士で作る．素性は 3.4 節で説明した単語ベクトルと品詞に加え，以下のものを用いる．係り受け解析，格解析には KNP[3] を用いる．

- 2 語間の係り受け階層数
- 2 語の 1 文中での共起頻度 (全論文から算出する) の逆数
- 直前と直後の助詞
- 格の種類
- キーワードのクラス

3.7 グラフの生成

上記の方法により抽出したキーワードをノード，関係をエッジとしたグラフを生成する．本研究では Python のライブラリである NetworkX と PyGraphviz を用いてグラフを生成し，出力する．生成したグラフの例を図 1 に示す．

[3]http://nlp.ist.i.kyoto-u.ac.jp/?KNP/

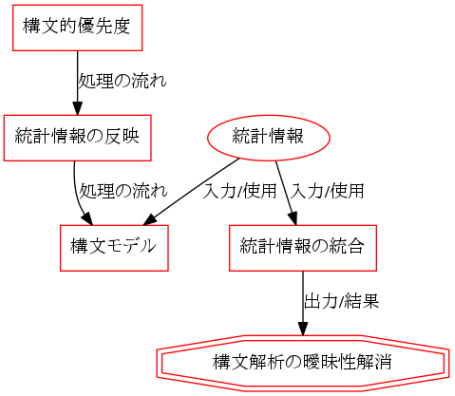

図 1: 生成したグラフの例

4 評価実験

評価実験では，提案手法によるグラフ生成の精度，およびグラフの有用性の 2 つの観点で評価を行う．

4.1 抽出精度の評価

単語ベクトルの作成には，論文誌コーパス [1] のうち日本語論文の題目，概要および各節のタイトルと本文において，単語の活用形を基本形に直したものを用いた．BoW は文書頻度が 0.3 以上の単語から作成し，分散表現は skip-gram により 200 次元のベクトルを作成した．アノテーションを行った論文 100 件中，訓練データを 80 件，テストデータを 20 件とし，アノテーションを行っていない論文は Self-Training に使用した．SVM は Python の scikit-learn パッケージのものを用い，カーネルはグリッドサーチにより最適化された RBF カーネル (C=10，γ=0.01) を用い，5 分割交差検証を行う．キーワード候補は非キーワードの数が多い不均衡データであるため，学習の際は対象のクラスのデータ数の割合に応じた重み付けを行った．また，Self-Training における論文の分割数は N=10，確信度の閾値は距離のノルムが 2.5 以上のものとした．キーワード抽出の結果，およびそれに基づく関係抽出の結果を表 4 に示す．

表 4: 抽出の精度

	適合率	再現率	F 値
キーワード抽出	0.65	0.71	0.68
関係抽出	0.41	0.39	0.40

キーワードの抽出誤りの原因としては，学習データの不足が考えられる．「A を用いる」「B を出力する」のような文脈で出現するキーワードについては正しく抽出されるが，「A に着目して」「B をまとめる」など，

周辺に現れる単語のバリエーションに対応できていないと思われるものが見られた．これには単語ベクトルだけでなく，手掛かり語を収集して用いる対策が考えられる．また，データクラスと手法/処理クラスの分類誤りが多く見られたが，これにはキーワードの上位概念や意味クラスを考慮するといった対応が考えられる．

4.2 有用性の評価

生成したグラフがサーベイの役に立つか，アンケートによる評価実験を行った．自動生成したグラフだけでなく，人手による理想的なグラフについても実験を行い，その違いの比較を行った．提案手法により自動生成したグラフ 3 件 (abc)，教師データとして人手で作成したグラフ 3 件 (def) の計 6 件の論文について，概要文とグラフを提示し，アンケートに回答してもらった．被験者は学部生 6 人，大学院生 5 人の計 11 人で，提示する論文の順序はランダムとした．アンケート項目は以下の通りである．

1. 題目とグラフだけを見て研究の概略を掴むことができたか．
2. グラフが論文の内容に適しているか．
3. グラフを提示されることで論文の理解の助けになったか．

回答は 1(そう思わない) から 5(そう思う) までの 5 段階評価とし，どれに当てはまるかを回答してもらった．実験の結果を表 5 に示す．数値は 5 段階評価の回答の平均値であり，5 に近いほど高評価，1 に近いほど低評価となる．

表 5: アンケートの結果

論文		アンケート項目		
		1	2	3
自動生成	a	3.3	2.8	3.0
	b	3.0	2.9	2.8
	c	2.0	1.7	1.9
手動生成	d	2.9	4.4	3.6
	e	3.7	3.6	3.8
	f	3.3	3.6	3.7
自動生成平均		2.8	2.4	2.6
手動生成平均		3.3	3.9	3.7

自動生成したグラフ a と b は 3 前後，c は 2 以下の評価であった．被験者からは，その論文固有の用語や曖昧な用語がノードになっているとき，意味や役割を理解し辛いという意見があった．今回の手法では，論文中に出現した単語をそのまま使用していたが，上位語や下位語，同義語へ言い換えることが望ましい場合もあると考えられる．また，完全なグラフでなくても，文章を読みながら頭の中で随所を補完することができるという意見も見られた．そのため，サーベイ支援の観点からは，再現率よりも適合率を上げることを重視した抽出を行うことが重要であると考えられる．

5 おわりに

論文の提案手法を要約したグラフを生成するため，論文中のキーワードと関係のアノテーションおよび抽出を行った．実験の結果，キーワード抽出の精度は (適合率，再現率，F 値)=(0.65,0.71,0.68)，関係抽出は (0.41,0.39,0.40) であった．また，グラフの効用を評価するための被験者実験も行った．今後の課題は抽出の精度を向上することに加え，サーベイに役立つ情報の抽出や提示方法の考案が挙げられる．また，概要だけでなく本文の情報を反映した抽出を行うことも必要である．

参考文献

[1] Sonal Gupta and Christopher D. Manning.“Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers” in 5th International Joint Conference on Natural Language Processing(IJCNLP), 2011.

[2] Yuka Tateisi, Tomoko Ohta, Yusuke Miyao, Sampo Pyysalo, Akiko Aizawa. “Typed Entity and Relation Annotation on Computer Science Papers” in 10th International Conference on Language Resources and Evaluation (LREC2016), 2016.

[3] Shin Wonha, 白井 清昭. “セグメント構造を考慮した学術論文の包括的要約の自動生成の提案” 言語処理学会 第 23 回年次大会 発表論文集 pp.230-233, 2017.

[4] Eunsol Choi, Matic Horvat, Jonathan May, Kevin Knight, Daniel Marcu. “Extracting Structured Scholarly Information from the Machine Translation Literature” in 10th International Conference on Language Resources and Evaluation (LREC 2016), 2016.