手掛り語を用いた論文概要から課題の自動抽出

倉本, 佑太
九州大学

中藤, 哲也
九州大学

廣川, 佐千男
九州大学

出版情報：情報処理学会研究報告. 2011, pp.1-4, 2011-03. 情報処理学会九州支部
バージョン：

情報処理学会研究報告
IPSJ SIG Technical Report

手掛り語を用いた論文概要から課題の自動抽出

倉本佑太[†1]　中藤哲也[†1]　廣川佐千男[†1]

関連研究調査を効率良く行うには、各論文が対象とする課題とその解決法を捉えなければならない。本稿では、手掛り語に着目し、課題を表す文を自動的に抽出する方法を提案する。

Automatic Extraction of Problems from Theses Abstracts Using Clue Words

Yuta Kuramoto,[†1]　Tetsuya Nakatoh[†1]
and Sachio Hirokawa[†1]

When we reed academic articles, we have to grasp problems what is intended for the article and solutions to them. This paper provides a method of extracting sentences describing problems automatically the abstracts using clue words.

1. はじめに

あらゆる分野において研究を進めていく際に，関連した先行研究の調査は欠かせない．しかし，世の中には極めて多くの論文が存在しており，自分が必要とするものを探し出すのは時間がかかり労力を要する．関連研究調査を効率良く行うには，目的や課題, 解決法などを各論文から捉えなければならない. 本稿では, 論文概要において, 目的や課題などを記している文によく使われている語（句）を手掛りに，それらを自動的に抽出する方法を提案する．

簡単な例を図 1 に示す．論文概要に手掛り語として「目的」と「困難」を与える．するとそれらを含む,「本論文は，証券市場における…相場操縦発見を目的とする.」と「不正取引の中でも…特に困難である.」という 2 つの文が抽出できる．これらは論文が対象としている目的や課題を表している文である．このように文を抽出することで, 論文が書かれた背景や目的などを知るために概要をすべて読む必要がなくなれば, 関連研究調査の効率化につながる．

関連研究として, 読みづらい文書の読解支援として手掛り句や定型表現に着目した研究が多数行われている．たとえば，特許文書の特徴的で難解な言い回しを逆に利用し内容を可視化する研究 [1),2),3),4)]や，大量に読む必要のある技術文書に関して，記述されている特長表現を利用し必要な情報を抽出する研究 [5)]がある．本研究では定型句や表現ではなく「単語」あるいは「語句」を用いる手法を提案する．論文に関しても，タイトルや概要から重要なキーワードを抽出する研究 [6),7),8)]が行われている. 本研究での提案手法は重要な情報をキーワードではなく文ごと抽出するため，情報が断片的になるのを防ぐ効果もある．

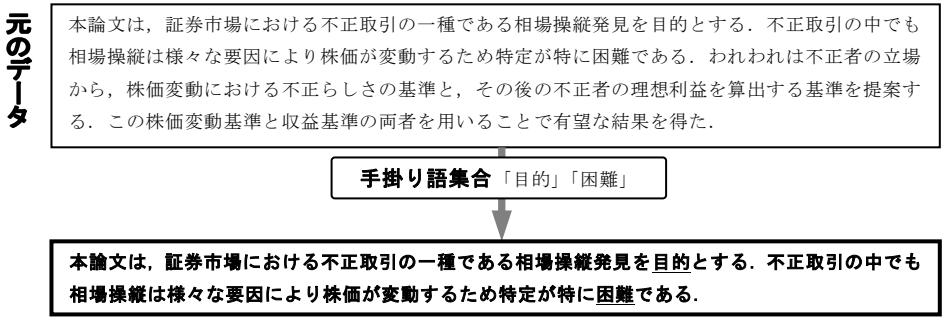

図 1　提案システムの例

[†1] 九州大学
Kyushu University

2. 論文概要・手掛り語・問題文・非問題文

2.1 論文概要

本研究では，電子情報通信学会(IEICE)の知能ソフトウェア工学研究会(KBSE: Knowledge-Based Software Engineering)における，2004 年から 2006 年に発表された 52 件の論文・発表概要のテキストファイルをデータとして用いた[a]．なお，知能ソフトウェア工学研究会は，1988 年に設立され，人工知能や知識工学，ソフトウェア工学に関する話題を研究対象としている．

各論文概要は 2 文以上，150～1000 字で構成されており，4～7 文，300～500 字のものが多い．また，内容としては主に研究が行われるきっかけとなった背景・状況・課題や従来の手法・装置，提案手法・装置やその補助情報（使用した器具・プログラミング言語など），得られた結果・例が記述されている．

2.2 手掛り語

本稿では，「手掛り解決法を表している文に含まれている特徴のある語または語句」と定義する．KBSE の 52 件の論文概要をすべて読み，図 2 に示すような語（句）を抜き出した．また，表 1 に手掛り語が文中に出現した回数を示す．

ない　本論文　課題　提案　着目　ことで　よって　本稿　重要　目的　対象
本研究　中でも　困難　として　本報告　ここでは　ために　効果的　難し
そこで　しかし　提示　可能となる　により　において　ならず　必要　少な
そのため　こととした　問題点　研究する　即ち　解析する　結果　これまで
てきた　視点から　手間　大変　遅　従来　求められ　異な　特に
近年　効果的　対し　着眼点　必ずしも　過ぎる　重点　この研究では

図 2　抜き出した手掛り語（全 54 語）

表 1　論文概要 52 件中の手掛り語出現回数（一部）

順位	手掛り語	出現回数	順位	手掛り語	出現回数
1	提案	44	9	により・そのため	7
2	しかし	16	11	難し・着目	6
3	本論文・本研究・ない	13	13	よって	5
6	本稿	12	16	必要・困難・ことで	4
7	問題・そこで	9	19	結果　他 6 語	3

2.3 問題文・非問題文

本稿では，「問題文」を「論文概要を構成する各文のうち，目的や課題について書かれているもの」と定義し，以後 P（＝Problem）という記号でも表すこととする．「非問題文」の定義はその逆であり，記号では NP（＝Non-Problem）と表すこととする．後の章で述べる評価のために，論文概要の各文について問題文か非問題文であるという判定を予め人手（筆者）で付与しておいた．

a) http://www.selab.is.ritsumei.ac.jp/kbse/

3. PNP 率

本研究では，目的や課題，解決法を捉えるために，手掛り語を用いて，論文概要の各文を目的・課題などを表している文とそうでない文に自動的に分別することを目的としている．そのため，対象とする複数の論文概要すべてにおいて，文を分別できることが望ましい．そこで，与えた手掛り語集合はどれほどの論文概要で文を分けられるか評価するため，PNP 率という指標を導入する．手掛り語集合 C を与えたときの PNP 率 $PNP(C)$を式(1)のように定義する．

$$PNP(C) = \frac{Absts}{All} \tag{1}$$

ここで *All* は対象とする論文概要の件数，*Absts* は与えた手掛り語集合のうち 1 語でも含む文とそうでない文の両方で構成される論文概要の件数である．PNP 率は 0 から 1 までの値をとり，PNP 率が高いほど，より多くの論文概要で文を分別することができたといえる．

図 3 に PNP 率を適用した実例を示す．論文概要 3 件に対し，手掛り語集合として C_1={“B”, “C”}，C_2={“A”}を与えた場合を考える．なお，手掛り語集合を与える事前準備として論文概要を文単位で区切っておく．図 3 の例では $PNP(C_1) < PNP(C_2)$となり，C_2 の方がより良い手掛り語集合であるといえる．

論文概要 1	論文概要 2	論文概要 3
1-1) **********	2-1) **********	3-1) **********
1-2) **********A******	2-2) **********	3-2) **********
1-3) **********	2-3) **********A**********	3-3) *******A*****
1-4) ******B**********	2-4) **********C*****	

手掛り語集合 C_1={“B”, “C”}の場合

- 手掛り語を含む文 P’$_1$={1-4,2-4}
- 手掛り語を含まない文 NP’$_1$={1-1.2.3,2-1.2.3,3-1.2.3}
- $PNP(C_1)$=2/3=**0.667**

C_1	1	2	3
P’	○	○	×
NP’	○	○	○

手掛り語集合 C_2={“A”}の場合

- 手掛り語を含む文 P’$_2$={1-2,2-3,3-3}
- 手掛り語を含まない文 NP’$_2$={1-1.3.4,2-1.2.4,3-1.2}
- $PNP(C_2)$=3/3=**1.00**

C_2	1	2	3
P’	○	○	○
NP’	○	○	○

図 3　PNP 率の適用例

4. 手掛り語を用いた課題の抽出

本章では，前章で定義した PNP 率をさまざまな単語集合に適用してより良い手掛り語の集合を求め，評価を行う．たとえば，2.2 節に掲載した手掛り語すべての組み合わせについて PNP 率を求める事で最良の手掛り語集合を求めることが原理的には可能である．しかし 54 個の語句があり，全ての組み合わせをとると 2^{54} 通りとなり，非現実的である．

4.1 手掛り語 1 語のみの PNP 率（単一 PNP 率）

図 4 に上位 20 位についての結果をグラフに示す．なお，手掛り語 1 語のみの PNP 率を「単一 PNP 率」と名付けることにする．

手掛り語 1 語のみの場合，「提案」の PNP 率が 0.635 と，突出して高い値を示した．これは，「提案」という単語は出現回数 44 回と他の手掛り語と比べてもかなり多く，また論文概要において「○○という手法を提案する」といった，新しい手法（解決策）を提示する形で出現することが多いためであると考えられる．他の手掛り語においては 0.2 前後の値をとるものが多く，1 語のみでは文の分別に適していないといえる．

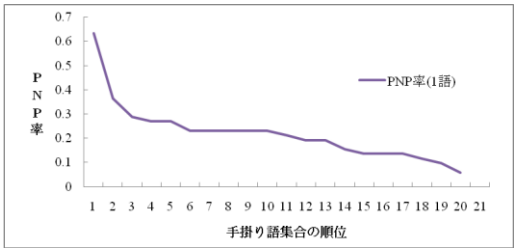

図 4 手掛り語 1 語のみの集合での PNP 率の結果

4.2 手掛り語 2~4 語での PNP 率

次に，手掛り語 2 語，3 語，4 語の集合で，全ての組み合わせで PNP 率を算出した．図 5 に結果を示す．

2 語, 3 語, 4 語と集合を増やすことによって, 最良の PNP 率の値も高くなっている．また，どの場合においても上位の組み合わせに「提案」という手掛り語が集合の要素の 1 つに入っていた．これは前節で述べた理由によるものだと考えられ，文を分別する基礎単語となっていると考えられる．

3 語，4 語の場合に着目してみると，『「提案」と「結果」と「他の手掛り語」』いう組み合わせが上位に現れている．「他の手掛り語」には「により」や「において」といった手法や状況を表すときに用いる語句が多く組み合わされており，「状況＋新手法に関する提起＋新手法による成果」という形式が文の分別に寄与しているものと考えられる．

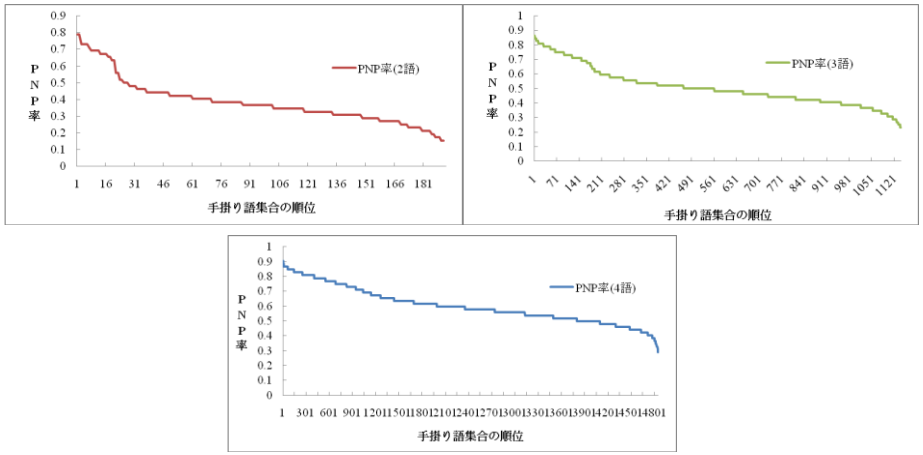

図 5 手掛り語 2~4 語での集合の PNP 率の結果

4.3 出現頻度上位 15 位までの手掛り語における PNP 率

本節では手掛り語の出現頻度に着目し，上位 15 位までの手掛り語において考えられる組み合わせすべてについて PNP 率を算出した．図 6 に結果を示す．

最も高かった PNP 率は 0.923 であり，これまでの実験で最も良い数値となった．分析してみるとおよそ 140 通りの組み合わせがあり，手掛り語集合の個数では 7~13 個であった．このことより手掛り語を増せばよい訳ではないことが理解できる．これは手掛り語が多ければ多いほど各文にいくらかの手掛り語が含まれている確率が高くなるためであると考えられる．

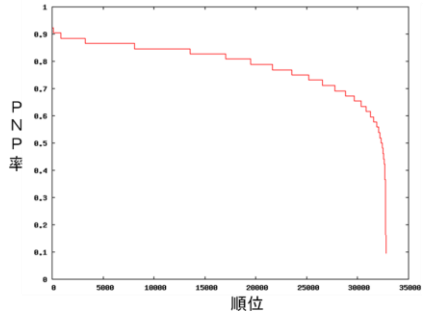

図 6 出現頻度上位 15 位までの手掛り語における PNP 率

5. SVM を用いた手掛り語の推定

5.1 SVM

SVM（Support Vector Machine）とは 1995 年に AT&T の V.Vapnik によって統計的学習理論の枠組みで提案された学習機械のことである[b]．SVM は，特にパターン認識の能力において，最も優秀な学習モデルの 1 つであることが知られている．

この SVM を用いて，論文概要の各文を問題文と非問題文に分別する最良の手掛り語集合を推定する．問題文と非問題文とに人手で分類した結果と単語の出現回数のデータを用意し SVM を実装した SVM-Light に与え，文中に出現する単語の重み W を逆算する．その上で重み上位の単語で手掛り語集合を作り PNP 率を求める[c]．単語の切り出しには形態素解析ツールである茶筌を用いた[d]．

なお，単語の重み W は(2)式に表すように SVM によって求まる各文のスコア *sentence_score* を単語の出現回数 *occurrence* で割ったものの総和であると定義する．

$$W = \sum_{n} \frac{sentence\_score_n}{occerrence_n} \qquad (2)$$

5.2 評価実験

名詞，接頭詞，動詞，形容詞，副詞，連体詞，接続詞の範囲で手掛り語の推定を行う．前節と同様に出現単語の重み付けを行い，上位 10 位の単語で集合を作り PNP 率を算出した（表 2）．名詞のみの場合に比べれば分野依存の単語は減少した．なお「本」はほぼ「本論文」や「本稿」といった形で出現するので分野依存の単語ではないこととする．また，0.962 という PNP 率はこれまでの結果では最も優れた値となった．

表 2　重み上位 10 位の単語で算出した PNP 率

順位	最小の単語集合	PNP 率
1	「実行　本　ため」・「実行　提案　本」	0.962
2	「本　ため」・「提案　本」	0.942
3	「こと　本　ため」	0.923

6. まとめと今後の課題

本研究では，関連研究調査の支援を目的として，論文概要において目的や課題などを表す文によく使われている語句（手掛り語）に着目し，それを手掛りに課題を表す文（問題文）とそうでない文（非問題文）とを自動的に分離する方法を提案した．今回データとして利用した知能ソフトウェア学会（KBSE）の論文概要 52 件においては，最小 7 個の手掛り語で約 92%の分離に成功した．手掛り語の個数の違いは見られるものの，SVM を用いた手掛り語の推定の結果（約 96%）に近いものとなった．

今後の課題としては，問題文と非問題文とに正確に分離できているか，また他の分野の論文概要でも手掛り語が適用できるか分析すること，最終的には分離できた問題文を可視化したり，文中のキーワードを抜き出し提示したりするような読解支援システムの構築などが挙げられる．

b) http://arx.ee.utsunomiya-u.ac.jp/research/svm/index.html
c) http://svmlight.joachims.org/
d) http://chasen.naist.jp/hiki/ChaSen/

参考文献

1) 新森昭宏，奥村学，丸川雄三，岩山真：手がかり句を用いた特許請求項の構造解析，情報処理学会論文誌，pp.891-905，vol.45，No.3（2004）
2) 高木慎也，新森昭宏：特許書類の可視化とハイパーテキスト化，第 9 回情報科学技術フォーラム，第 2 分冊 pp.333-336（2010）
3) 石川大介，石塚英弘，宇陀則彦，藤原譲：特許文献における因果関係の抽出と統合，情報知識学会誌，vol.14，No.4，pp.105-118（2004）
4) 酒井浩之，野中尋史，増山繁：特許明細書からの技術課題情報の抽出，人工知能学会論文誌，vol.24，No.6，pp.531-540（2009）
5) 西山莉紗，竹内広宜，渡辺日出雄，那須川哲哉：新技術が持つ特長に注目した技術調査支援ツール，人工知能学会論文誌 24 巻 6 号 SP-A，pp.541-548 （2009）
6) 近藤友樹，難波英嗣，奥村学，新森昭宏，谷川英和，鈴木泰山：論文データベースからの研究動向情報の抽出，言語処理学会第 13 回年次大会，pp.470-473（2007）
7) 村田真樹，一井康二，馬青，白土保，金丸敏幸，井佐原均：過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査（2007），言語処理学会ホームページ（http://www.nak.ics.keio.ac.jp/NLP/trend-survey.html）
8) 村田真樹，Stijin De Saeger，橋本力，風間淳一，山田一郎，黒田航，馬青，相澤彰子，鳥澤健太郎：論文データからの重要情報の抽出と可視化，第 23 回人工知能学会全国大会（2009）