手掛り語に着目した論文概要からの課題抽出

酒井, 敏彦
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. 2012 (B-4-1), pp.1-6, 2012. 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

手掛り語に着目した論文概要からの課題抽出

酒 井 敏 彦[†1]　廣川 佐千男[†1]

関連研究や研究動向を調査するには，検索結果の論文概要を様々な観点で理解することが必要である．特に新しい分野を対象とするときやその分野の重要課題を見付けるには，単純な検索やクラスタリングでは困難で，人手でしっかり読まなければならない．本稿では，課題を表す特徴的な文のパターンがあると仮定し，複数の手掛かり語からどの手掛かり語集合が問題文を抽出できるかを考察する．

Extract problems from article abstract using clue.

Toshihiko Sakai[†1] and Sachio Hirokawa[†1]

In order to investigate related works and research field, it is required to understand the paper abstract of search results in various viewpoints. In paricular, In order to find the important problem of the research field and aimed at a new research field, simple search and clustering is not enough. we have to read the paper manually. In this paper, we assume that there is a pattern of characteristic problem sentence, and considers which clue set can extract problem sentence from multiple clues.

†1 九州大学
Kyushu University

1. はじめに

研究を始めるとき，あるいは研究成果をまとめるときには自分が研究を行っている分野やその内容に応じた関連研究を調べる必要がある．現在までに出版されている論文を全て読むことは困難である．通常は自分が探している条件に合致するように検索を行い，論文を探す場合が多い．例えば，キーワード，著者名，発表年，関連キーワードなどを条件として入力し検索を行い，絞りこまれた論文を調査する．しかし，検索条件が弱ければ検索結果が多すぎたり，逆に，検索条件がきつすぎると求める論文が見つからないこともある．我々はこういった問題に対して様々な観点，目的に応じた論文検索を行う研究を行なっている[1),2)]．

様々な観点，目的に応じた論文検索を行うために論文構成要素ごとの分類を行う研究がある[6)]．例えば，論文は「背景」,「問題」,「結果」のように複数の要素で構成されている．関連研究調査を効率良く行うには各論文で記述されている「問題」に着目する必要がある．特に，論文概要においてはその論文で解決すべき問題が記述されている．

そこで，本稿では論文概要から問題文の抽出を行うことを目的とした．あらかじめ人手で問題文の判定を行い，それを SVM に学習させ，問題文識別を行う．学習データの構築において，(a) 全ての単語，(b) 人手で選んだ 53 個の手掛かり語，(c) 問題文の特徴語上位 53 個 (TF-IDF 値上位) の 3 通りについて F 値での比較を行った．また，最も性能の高かった (a) について，ポジティブな単語とネガティブな単語の効果について分析した．

2. 関連研究

論文の関連研究章において論文構成要素ごとに分類を行う研究がある[3),5)]．Angrosh らは関連研究の章だけに着目することで，個別の論文では困難な関連分野の概略を求める方法を述べている．具体的には，引用の有無と文に現れるキーワード，フレーズに着目することで文を自動で分類するカテゴリーを定義し，このカテゴリーに従い文を論文構成要素ごとに分類する手法を提案した．分類結果をトレーニングデータとして CRF(Conditional Random Field) で学習させた．10-分割交差検定による評価で精度は 96.51%だった．[3)]

Sakai らは特許文書から技術課題情報を抽出するのに有効な手掛かり表現を自動的に獲得し，それを用いて抽出する手法を提案している．1 つの手掛かり表現から直前に出現する表現を抽出し，その表現から新たな手掛かり表現を獲得し，これからさらに直前に出現する表現を抽出する．これを繰り返し，手掛かり表現を自動獲得している．実験の結果，精度

表 1 問題文の例[⋆1]

	文
1	広域における対象追跡への利用を目的として，首振りカメラを用いた分散カメラシステムの確率的連結関係を効率的に推定する手法を提案する．
2	確率的連結関係とは，各カメラの撮影可能領域内のどの点から対象が出現・消失しやすいか，またそれらの点の間に経路が存在するか，および経路が存在する場合は各経路をどの程度の確率で使用するかを示すものである．
3	首振りカメラを利用することで固定カメラに比べて広範囲・高解像度の撮影が可能となるが，各瞬間には首振り方向しか観測できないために対象の出現・消失情報を取得しにくくなるという問題がある．
4	提案手法では，対象の出現・消失が観測されるたびに連結関係を逐次推定し，推定された経路の不確かさや各経路を対象が移動する確率を考慮して各カメラを首振り制御することで，効率的に対象の出現・消失情報を取得する．
5	シミュレーション実験により提案手法の有効性を確認した．

(78.0%)，再現率 (77.6%) ともに高い結果を示している．[4)]

本研究の特徴は文書全体ではなく文を分析したところである．当初は，論文での問題文に着目し，問題文を特徴付けるポジティブな手掛かり語が問題文の抽出に深く関わってると考えて研究を始めた．しかし，問題文の判別にはネガティブな手掛かり語も重要な役割を担っていることがわかった．

3. 問題文と特徴語

一般的な論文は複数の文で構成されており，必ずその論文で解決しようとした問題が書かれている．我々は問題文に着目し，SVM を適用することで論文概要の問題文，非問題文の判別を行う．例えば，表 1 の論文概要では 3 文目が問題文となる．

人手で選んだ問題文を用いて SVM による機械学習を行う．そこで学習させる特徴語として次の 3 種類を考え論文概要の各文を特徴付ける．

(a) 全ての単語: 論文概要に出現する全ての単語 4,548 個を学習データに用いる．

(b) 人手で選んだ 53 個の手掛かり語: 以前我々が提案した 53 個の手掛かり語[1)] を学習データに用いる．

表 2 に 53 個の手掛かり語を示す．

(c) 問題文の特徴語上位 53 個 (TF-IDF 値上位): 問題文の特徴語に関して TF-IDF 値を計算し，その上位 53 個を学習データに用いる．

表 3 に TF-IDF 値上位 53 個の単語を示す．手法 (b) と公平な比較にするために手法 (c) も 53 個とした．また，TF-IDF 値の上位は 1 文字の単語が多いことがわかった．TF-IDF 値の計算は次式を用いる．

$$TF\text{-}IDF_{(w,S,D)} = \frac{tf(w,S)}{\sum_{w \in S} df(w,S)} * log \frac{|D|}{df(w,D)}$$

w は単語，S は問題文集合，D は全ての論文概要集合である．$tf(w,S)$ は S における単語 w の頻度，$df(w,D)$ は単語 w を含む文書 D の数を表す．

表 2 53 個の手掛かり語[1)]

提案	本稿	ない	しかし	そこで	本論文
課題	着目	ことで	よって	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	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	対し
着眼点	必ずしも	過ぎる	重点	この研究では	

表 3 TF-IDF 値上位 53 個の単語

が	しかし	ある	ない	は	に
で	て	と	な	ため	れ
さ	いる	問題	なる	い	や
場合	通信	困難	の	を	する
た	し	こと	的	化	その
性	ネットワーク	など	必要	において	として
情報	この	から	という	従来	推定
における	システム	低下	でき	時間	特性
評価	により	処理	も	手法	

⋆1 寺下訓史, 浮田宗伯, 木戸出正継, 広域分散首振りカメラ群における確率的連結関係推定法の効率化, 信学技報, vol. 108, no. 484, pp. 241-248, 2009.

4. 実験と考察

データとして電子情報通信学会研究会[⋆2]で検索できた 2004〜2011 年の期間における 42,921 件 (2011 年 8 月 26 日現在) の論文計 300 件をランダムに選択し，その中から論文概要の抽出を行った．

4.1 SVM での学習データ

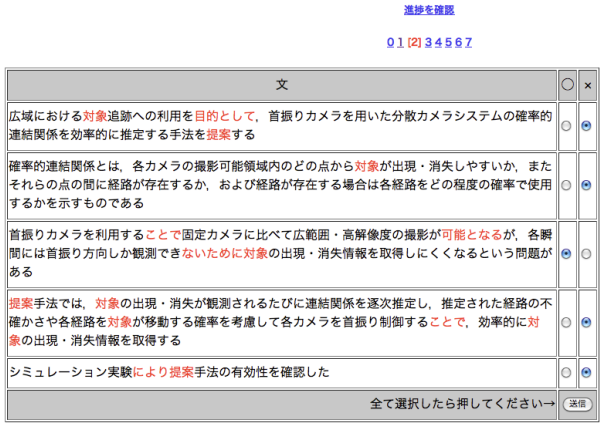

図 1 問題文判定画面

SVM で学習データを決定するため事前に論文概要の各文を読み，問題文か非問題文であるかを人手によって評価した．論文概要 300 件における合計 1,353 文を 3 人でチェックした．図 1 は問題文評価を行う画面である．図 1 のラジオボタン「◯」で問題文,「×」で非問題文の判定を行なってもらった．今回は 2 人以上が問題文であると判定した文を問題文とし，SVM での学習データとして用いた．2 人以上が問題文と判定した文は 1,353 文のうちの 149 文であった．

今回 SVM で学習させる際のツールとして SVM-Light[⋆3]を用いた．SVM-Light での学習データは図 2 の形式になる．論文概要の 1 文は学習データの 1 文と一意に対応している．行頭の数値はその文が問題文 (1) か非問題文 (-1) かのラベルである．コロンの直前は単語の番号であり，コロンの直後はその単語の文中での出現頻度である．1,353 文に対して，形態素解析ツール MeCab を用いた結果，単語の総数は 4,548 個であった．

```
1 1:1 23:1 27:2 28:2 50:1 55:1 56:2 57:2 111:1
83:1 2634:1 4152:1 4153:1 4154:1 4155:1 # @an20
-1 4:1 7:1 10:2 34:1 65:1 164:2 204:1 239:1 273
905:1 2906:1 # @ap2007-148.html-1
```

図 2 SVM-Light での学習データ形式

4.2 評価指標

評価指標として Precision，Recall，F 値を求めた．今回は閾値 α を決め，SVM の推定値が α 以上のものを「推定問題文」と判定する方法を取った．文 s_i に対し，SVM による推定結果値を $svm(s_i)$，s_i が問題文であるかどうかを $p(s_i)$ で表す．問題文である時は $p(s_i)$=1，問題文でないときは $p(s_i)$=0 である．閾値 α での $precision(\alpha)$，$recall(\alpha)$，$F(\alpha)$ は以下のようになる．

$$Precision(\alpha) = \frac{\#\{s_i | svm(s_i) \geq \alpha \ and \ p(s_i) = 1\}}{\#\{s_i | svm(s_i) \geq \alpha\}} \quad (1)$$

$$Recall(\alpha) = \frac{\#\{s_i | svm(s_i) \geq \alpha \ and \ p(s_i) = 1\}}{\#\{s_i | svm(s_i) = 1\}} \quad (2)$$

$$F(\alpha) = \frac{2}{1/Precision(\alpha) + 1/Recall(\alpha)} \quad (3)$$

評価として 10-分割交差検定を用いた．今回のデータでは 1,353 文を 10 等分し，9 つを学習データとし，1 つをテストデータとした．10 つのグループが一回ずつテストデータとなるように 10 回 SVM での実験を行い，10 回の結果の平均をとった．

4.3 実験結果

図 3 は「全ての単語」での結果を示す．図 3 より F 値の最大は α=-0.60 のところで約 0.93 であった．図 4 は「人手で選んだ 53 個の手掛かり語」での結果を示す．図 4 より α=-2.00 から α=-1.10 まで F 値の値は高々約 0.19 で全体的に F 値の値は低かった．図 5 は「TF-IDF 値上位 53 個の単語」での結果を示す．F 値は α=-0.60 のところで最大で約 0.67 だった．

結果として,「全ての単語」が一番高い値で,「人手で選んだ 53 個の手掛かり語」は低い結

⋆2 http://www.ieice.org/jpn/index.html
⋆3 http://svmlight.joachims.org/

果となった．

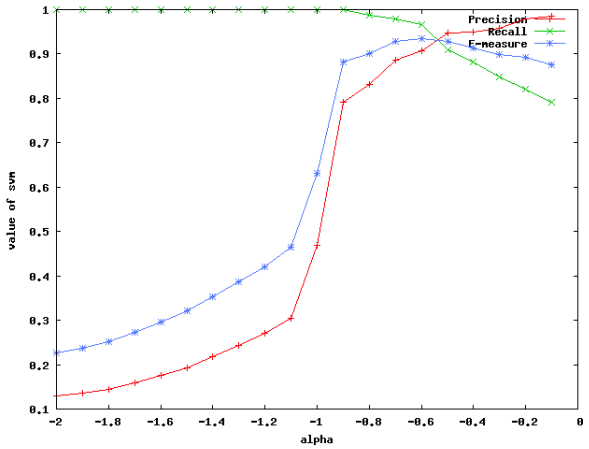

図 3 「全ての単語」の結果

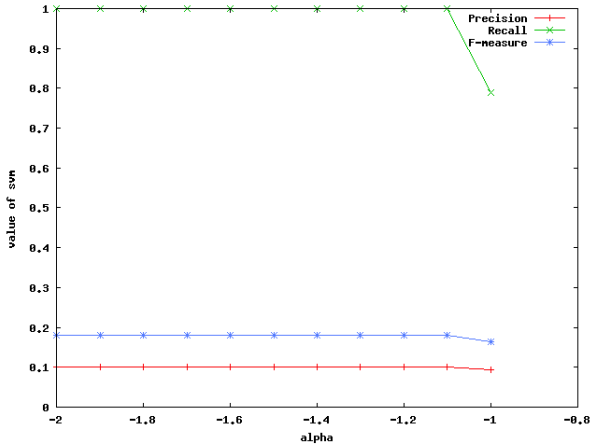

図 4 「人手で選んだ 53 個の手掛かり語」の結果

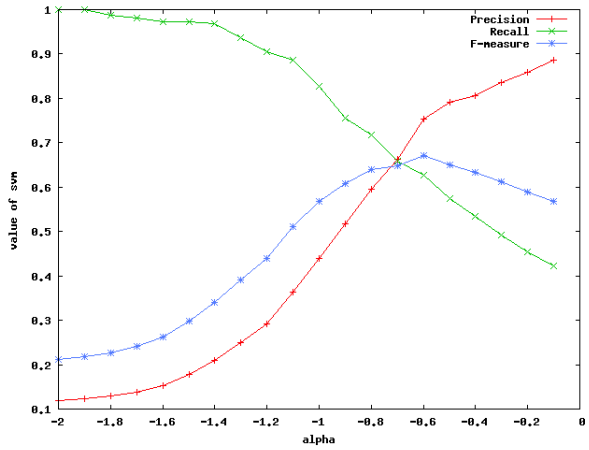

図 5 「TF-IDF 値上位 53 個の単語」の結果

表 4 実験結果

	F 値 (α=-0.60)	単語の総数
全ての単語	0.93	4,548
手掛かり語 53 個	-	53
TF-IDF 値上位 53 個	0.67	53

5. ポジティブな単語とネガティブな単語の影響度

前節までで論文概要から問題文を抽出するためには「全ての単語」が一番良い結果であることがわかった．しかし，SVM でどの単語が問題文判定に大きく影響しているかはわからない．そこで,「全ての単語」での単語の重みの分布を調べた．単語 w の重み $weight(w)$ は式 (4) で求めることが出来る．

$$weight(w) = \sum_{i=1}^{n} \frac{freq(w, s_i) * svm(s_i)}{freq(s_i)} \tag{4}$$

$freq(w, s_i)$ は文 s_i における単語 w の出現頻度を表す．$freq(s_i)$ は文 s_i での単語の個数

を表す．

図 6 がその分布の図である．縦軸は単語の重みの値，横軸は単語のランキングである．この図を見るとほとんどの単語は値 0 付近にあることがわかる．実験を行う前，我々はポジティブな単語が問題文判別に大きな影響力を持っていると考えていた．しかし，図 6 からポジティブな単語だけでなくネガティブな単語も存在していることがわかる．

5.1 結　　果

ポジティブな単語とネガティブな単語が SVM の学習にどう影響するかを調べるため，次の 3 つの条件で再度学習させ F 値の結果を調べた．

(1)　上位 5 個の単語
(2)　下位 200 個の単語
(3)　上位 5 個の単語+下位 200 個の単語

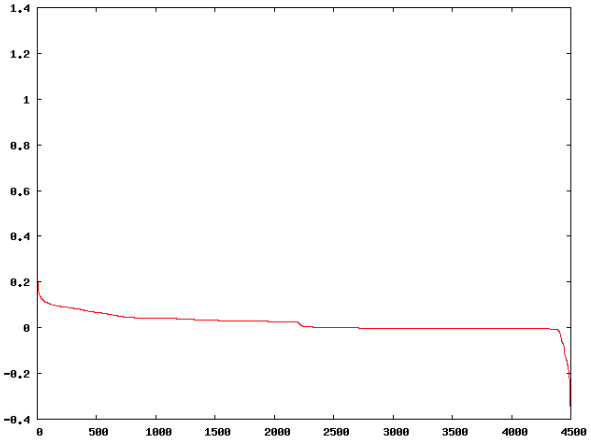

図 6　「全ての単語」での単語の重み

図 7 は以下の 3 つの条件での F 値の結果である．この図から α=-0.6 のとき (3) 上位 5 個の単語+下位 200 個の単語が一番高い F 値をとっている．このことから，SVM による単語の影響度というのは単語の重みが高いものと低いものによって強い影響を受けていることがわかった．

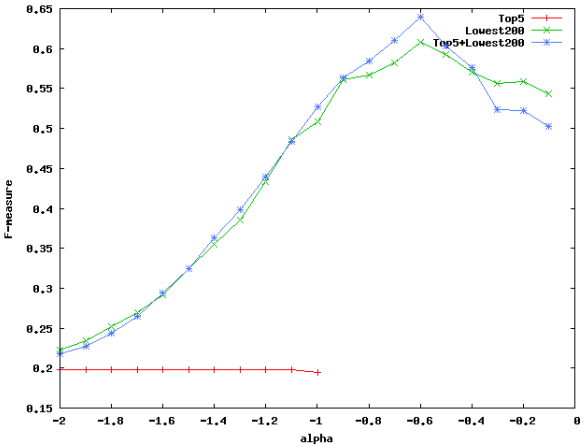

図 7　3 つの条件での F 値の結果

6. おわりに

本稿では，論文概要の中でその論文で扱う問題を記述した文 (問題文) を SVM を用いて抽出する方法を提案した．(a) 全ての単語，(b) 人手で選んだ問題文を表す手掛かり語集合，(c) 問題文の特徴語上位 (TF-IDF 上位) の 3 通りの特徴語集合について SVM で学習させた．その結果全ての単語を用いた場合が最も良い F 値 (0.93) を得ることができた．

この理由として学習させる単語が他の手法より多かったということが考えられる．しかし，TF-IDF 値上位 53 個で学習させた場合は F 値が 0.67 だった．単語の総数は「全ての単語」は 4,548 個だったのに対して，TF-IDF 値では 53 個であった．これは「全ての単語」に比べて非常に少ない個数である．このことから学習させる単語において SVM で問題文の分類を行う場合何らかの影響の違いがあると考えられる．

そこで，一番結果が良かった「全ての単語」においてどの単語が SVM における問題文判定に大きく影響しているかを調べた．単語の重みを分析することでポジティブな単語だけでなくネガティブな単語も存在することがわかった．その後，単語の重み上位 5 件と下位 200 件を SVM において適用し，問題文判別に良い影響を与えることがわかった．

発表者は今後の方針として論文概要を論文構成要素ごとに分類を行うことを考えている．そのためには論文から正確に論文構成要素を抽出するための論文構成要素ごとの良い手掛

かり語を発見する必要がある．今回は SVM を用いたが，他の機械学習の CRF，赤池情報量規準などを用いて問題を抽出できているかの精度を比較することを考えている．また，問題を抽出したあとはその問題の抽象度，具体度を考慮する．問題の抽出が終わったあとは解決，手法などの項目を抽出し，他の論文構成要素である「目的」や「解決」などにおいても同様のことを行う予定である．

参 考 文 献

1) 倉本佑太, 中藤哲也, 廣川佐千男, 手掛り語を用いた論文概要から課題の自動抽出, 情報処理学会 九州支部 火の国情報シンポジウム,2011.
2) 廣川佐千男, 二つの観点に基く検索結果の分析方法 Double Rank について, 第一回テキストマイニングシンポジウム,2011.
3) Angrosh, M. A., Craneeld, S., and Stanger, N.: Context identication of sentences in related work sections using a conditional random field:towards intelligent digital libraries, in Proceedings of the 10th annual joint conference on Digital libraries, JCDL '10, pp. 293-302, 2010.
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) 亀田尭宙, 武田英明, 相澤彰子, 関連研究に関する記述の分析による論文間の意味的関係の抽出, 人工知能学会全国大会 (第 25 回),2011
6) 白井宏和, 春原将寿, 中村勝一, 横山節雄, 宮寺庸造, 論文構成要素に着目した論文間関係把握支援システムの開発, 電子情報通信学会技術研究報告,vol.108, pp.23-28, 2009.