学術論文概要中の各文の観点推定

廣川　佐千男*　　酒井　敏彦（九州大学）

Estimation of Viewpoints of Sentences in an Abstract of Scientific Article

Sachio Hirokawa*, Toshihiko Sakai (Kyushu University)

概要　関連研究調査では、検索結果で得られる多量の文献と自身の研究成果がどのような観点で関連があるかを明確にすることが重要である。本稿では、論文概要に含まれるそれぞれの文について、背景、問題、関連研究、目的、手法、結果などの観点を表す手掛り語があると考え、SVM の属性選択により最適な手掛り語集合を求め、各観点を表すモデルを構築し、識別を行った。300 件のデータを用いた 5 分割交差検定で、背景、問題、関連研究、目的、手法、結果の F 値はそれぞれ、0.53、 0.39、0.08、0.46、0.66、0.31 となった。これらはいずれも属性選択を行わないモデルによる性能よりも向上している。

キーワード：ＳVM，属性選択，分類精度，機械学習

(SVM, Feature Selection, classification accuracy, machine learning)

1.　はじめに

論文概要は自然言語で書かれていて、形式的な構造は決められていない。しかし、論文概要はその論文の要点を簡潔にまとめたもので、限られた文字数の中に著者が必要不可欠と考える事項を記述している。どのような学術論文でも、共通の必須項目はある。実際、国際会議やジャーナルの査読規定にはそのような項目が記載されている。ジャーナルによっては、論文概要の記載項目を XML として規定しているものもある。

このように、学術論文は自由に記述できる文章であるが、意味的、あるいは論理的には構造化文書を考えるべきである。しかしながら、自然言語で記述された文章から論理的構造を自動的に抽出するのは容易なことではない。そこで、本稿では、論文概要の意味的構成要素である背景、問題、関連研究、目的、手法、結果を表す文のモデル化を試みる。具体的には、人手で論文概要の各文を読み、どの観点を述べたものか判定してもらう。この人手による判定結果を学習データとして機械学習を適用し、各観点の推定を行った。

SVM は汎化性能が高く、判別のための属性が大規模であっても有効なので、さまざまなデータについて利用されている。しかし本稿で識別対象とするのは文なので、一つのデータに含まれる単語の数は３５個程度しかない。一般の文書分類の場合には、属性数が多い方が識別性能がよい。しかし、本稿で扱う文のような場合には、状況が違うのではないかと著者らは考えた。そこで、SVM の研究でも広く行われている属性選択により識別性能を向上できるのではないかと考えた。また、背景、問題、関連研究、目的、手法、結果など、それぞれの観点を表す特有の表現や手掛り語があるのではないかと考えた。著者らはこれまでに、問題を表す文の識別を行ってきたが、本稿では、他の観点でも同様に観点を表す手掛り語があり、それらを用いることで識別性能向上ができることを示す。

2.　関連研究

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[Angrosh2010]。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ている。[Sakai2012]では、学術論文で扱われる問題という観点での検索を実現するため、論文概要の中で問題を表す文を SVM を使って判定する方法を提案している。様々な手掛り語集合を検討しているが、特に手掛り語集合は限定せず全ての単語を利用する場合が最も判別性能がよいという結果を示している。野中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[Nonaka2010]。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．本稿では、[Sakai2012]の分析を詳細化することで、問題文判別性能が向上する手掛り語集合はどのようなものか、また、どのような状況で判別性能が向上するかを分析した。

少数の手掛り語やキーフレーズから始めて、ブートストラップによりエンティティを獲得する手法としては、Pantel 等の Espresso がある[Pantel2006]。反復過程で当初の話題からずれてしまうトピックドリフトを回避するために、色々な試みがなされている。[Komachi2008, 小町 2010, Radev2008]では、単語の共起関係をグラフとして捉え、単語と文書、あるいは単語と文という二種類の節点から成る二部グラフにおいて密に関連する部分グラフを求める問題として統一的な解釈が述べられている。廣川は、「しかし」というキーワードを手掛り語として企業の倒産状況を表す文書から、倒産理由を抽出する手法を提案している[Hirokawa2012]。貞光らは、トピックとは別に獲得したい対象文書が共通に持つ属性も合せて利用することで、トピックドリフトを回避する手法を提案している[Sadamitstu2011]。これらの研究は、文書*単語行列から、特定の文書群を特徴づける単語集合を求める問題を扱っていると見ることもできる。一方、識別すべき文書群の事例が学習データとして入手できる場合には、SVM で判別モデルを構築することができる。また、判別に必須の属性が何かを明かにする属性選択問題は、SVM に関する重要テーマとして多くの研究がある。著者らの知る限りでは、属性選択と手掛り語を共通の課題としてとらえ、手掛り語集合による判別性能を評価した研究はない。

SVM を使った分類で属性選択は重要テーマとして多くの研究がある。[Alonso-Gonzalez2010] では、少数のサンプルしかないマイクロアレー遺伝子発現データの分類について、数千以上の属性に対し 40 個程度の属性選択でも同程度の分類性能が達成できると報告している。[Hermes2000]では、全データをトレーニングデータとして学習したモデルにおける各属性のスコアを考え、1.0 に近いもの少数の属性だけに限定して再度モデルを作っても、ほぼ同程度の分類性能が達成できることを示している。このように、属性選択についての多くの研究は、対象を少ない属性により理解することが目的であり、少数の属性でも全ての属性を使った場合と同程度の分類ができることを示しているもので、少数の属性に限定することで、分類性能が向上するというものではない。本稿では、少数の手掛り語を使った方が、全ての単語を使った場合よりも、判別性能が高くなること示す。

3. 各観点を表す文

今回データとして 2004 年から 2011 年に出版された電子情報通信学会の研究会論文 42,921 件を収集し、その中からランダムに 300 件を抽出した。300 件の論文概要を形態素解析した結果、全ての単語の総数は 3,915 個、総出現数は 58,418、文の数は 1,672 文、一文あたりの単語出現数の平均は 34.94 である。SVM で学習させる際には、ベル(-1 or +1) をつける必要がある。今回は論文概要の文においてある観点を表すか否かを人手で判定した。このラベル付けを行うため、3 人の被験者に 300 件の論文概要を読んでもらい判定をしてもらった。図 1 に論文概要例と、文判定作業の画面を示す。本稿では、問題、背景、関連研究、目的、手法、結果などの分類を行った。2 人以上が同じ観点と判定したもの正解とした．観点ごとの正解数を表 1 に示す。これを教師データとして線形カーネルによる SVM-light で各観点の識別を行った。

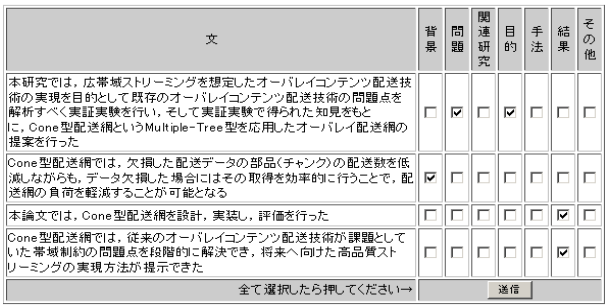


図 1 人手による文の分類作業画面

Fig. 1. Snapshot of Manual Sentence Classification

4. 全ての単語を使ったベクトル化での観点推定

まずナイーブな識別方法として、各文に現れる全ての単語を使ってベクトル化を行い、そのデータを使って SVM の学習を行った。人で判別を行った 1,672 文について、ランダムに５分割し、８割のデータを学習データとし、残りの２割のデータをテストデータとして５分割交差検定で性能評価を行った。表１の Prec,Rec,F 値,Acc がその結果である。「関連研究」と「その他」の観点についてはほとんど推定できていない。300 件の論文概要の中で、人手で行った判別作業の結果でも、この二つについてはそれぞれ３７個、１２個しか文がなかった。一方「手法」は５割、「背景」「問題」「結果」は３割の F 値となっていて、自動推定の可能性があるといえる。

観点	件数	Prec	Rec	F 値	Acc
背景	234	0.2203	0.7838	0.3416	0.6158
問題	194	0.1930	0.8657	0.3151	0.5974
関連研究	37	0.0456	0.7862	0.0858	0.6559
目的	222	0.1451	0.6125	0.2345	0.5076
手法	532	0.3798	0.7651	0.5074	0.5625
結果	168	0.1695	0.9177	0.2856	0.5765
その他	12	0.0306	0.3190	0.0537	0.5312

表１ 全ての単語でベクトル化したときの観点推定性能

Table 1. Viewpoints Estimation Performance with respect to all words

5. SVM による各観点の特徴語

線形カーネルを使った SVM の学習では、ラベル $y_i=\{+1,-1\}$が付けられた l 個数の学習データ $x_i \in R^d$を分類するマージン最大の超平面を求めるため、$y_i*(w^Tx_i+b)>1$ $(i=1,...,l)$ を満足するように属性スコア $w = (w_1,...,w_d)$を求める。本稿では次元数 d は全ての単語の総数で d=3915 であり、w_j は問題文判定のための各単語の重要度を示すスコアと考えられる。全データで全単語を使って学習したときの単語のスコアについて、絶対値の大きい上位 5 個の単語を表２に示す。「その他」以外の観点については、それぞれの観点の特徴らしい単語が正のスコアの単語として現れている。

観点	ポジティブ上位 5個	ネガティブ上位 5個
背景	近年 重要 られる 普及 や	本 提案 しかし ない 時間
問題	しかし 問題 ない 困難 しかしながら	本 可能 も 者 し
関連研究	研究 き 方法 さ 認識	本 より 的 こうした システム
目的	論文 提案 研究 本 目的	機能 有効 及び 性 実験
手法	まず 次に 導入 具体 へ	論文 研究 が いる しかし
結果	結果 が % 示さ た	モデル 示す する 論文 について
その他	期待 ツール 音楽 1 3	を た 手法 から で

表２ SVM スコア上位語
Table ２. Top words of SVM score

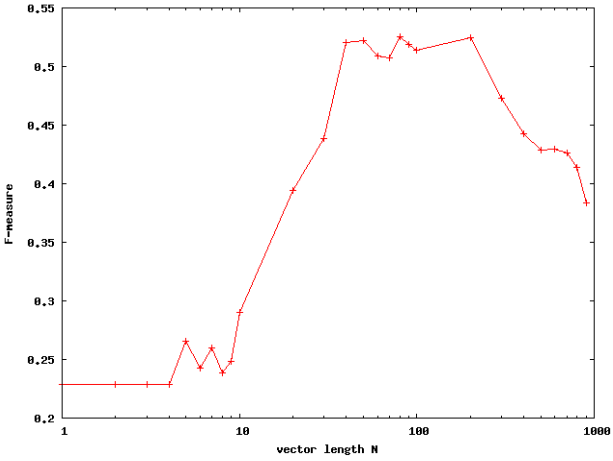
5. 観点推定性能に対する属性選択の効果

本章では、絶対値が SVM スコア上位 N 個の単語を使ってベクトル化を行い、観点の推定を行った。

観点	最適 N	Prec	Rec	F 値	Acc
背景	80	0.3775	0.8717	0.5253	0.7979
問題	200	0.2485	0.9469	0.3929	0.6847
関連研究	30	0.0447	0.9800	0.0846	0.5163
目的	100	0.2648	0.9692	0.4157	0.6652
手法	100	0.5582	0.8091	0.6595	0.7539
結果	70	0.1859	0.9838	0.3125	0.6019
その他	800	0.0615	0.3190	0.0956	0.5637

表 3 最適属性選択による観点推定性能
Table 3. Viewpoints Estimation Performance with respect to Optimal Feature Selection

図２は「背景」について N を変化させたときの F 値をプロットしたものである。付録に各観点についてのプロットを示す。「その他」を除いて上に凸のグラフとなっており、最適な N、すなわち、最適な属性選択があることが分かる。表３は、各観点について最適な N のときの判定性能をまとめたものである。



(a) 背景

図２ 選択属性数と F 値の関係（背景）
Fig. 2. F-measure with respect to the number of selected Features (Background)

全ての単語によるベクトル化での推定性能と属性選択で最適な場合の推定性能を比較したのが図３である。いずれの観点についても属性選択が有効であることが分かる。特に手法、背景、目的、問題については属性選択により F 値が 0.4 以上となっている。

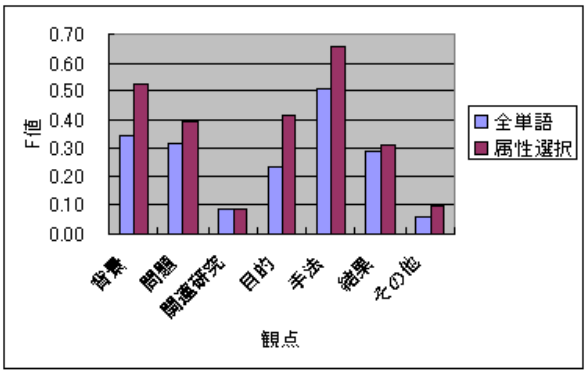
6. まとめと今後の課題

問題や技術や成果などの観点に応じた論文検索システムを実現するため、論文概要中のそれぞれの文がどのような観点を表すか識別する方法を提案した。観点としては、問題、背景、関連研究、目的、手法、結果の６つを対象とした。300 件の論文概要の各文を３人に読んでもらい、どの観点に該当するか判定してもらい、正解データとした。このデータに対して SVM を使って 5 分割交差検定で各観点のモデルを構築した。各文のベクトル化としては、全ての単語を使う方法と、SVM スコアの絶対値上位 N 個の単語でベクトル化する方法の二通りで評価を行った。後者が学習に利用できる例題が少数の場合、全ての単語を利用して学習したモデルよりも、スコアの絶対値の高い単語に限定することで、判定性能が向上できることを示した。

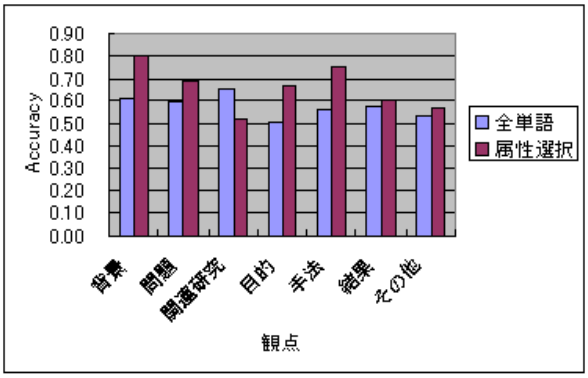
SVM を使った分類について、属性選択の効果の研究が多

数あるが、このような明らかな判別性能向上を示したものはない。その違いは、従来の研究がある文書を対象としていたのに対し、本稿で分類対象としたものが短い文であることに由来すると予想している。

本稿で得られたＦ値は十分なものとはいえない。しかし、評価実験に用いたデータ数と Ｆ 値の関連をプロットした図４を見ると、性能向上が期待できる。すなわち、正例が多い手法では７割の Ｆ 値となっている。つまり、人手による学習データを今の２～３倍準備することで、より一層の性能向上が期待できる。

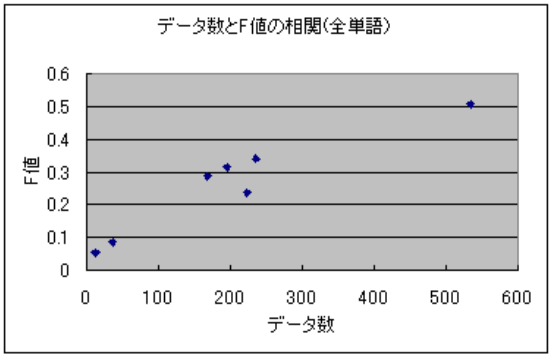


(a) F 値の比較

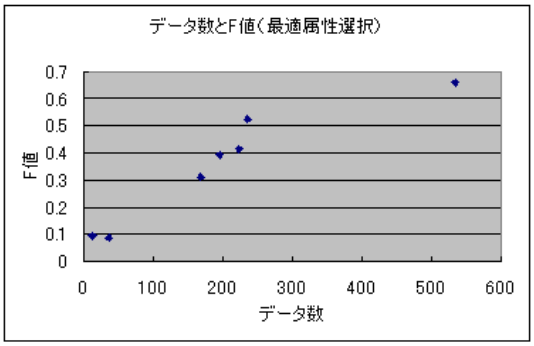


(b) Accuracy の比較

図３　全単語でのベクトル化と最適選択属性での性能比較



(a)　全単語でベクトル化



(b) 最適属性選択によるベクトル化

図４　データ数と Ｆ 値

本研究は JSPS 科研費 24500176 の助成を受けた。

文　　献

[Alonso-Gonzalez2010]n Alonso-Gonzalez,C.J., Moro, Q.I., Prieto, O.J.,Simon, M. A., Selectiong Few Genes for Microarray Gene Expression Classification, Springer LNCS 5988, pp.111-120 (2010)

[Hermes2000] Hermes, L., Buhmann, J.M., Feature Selection for Support Vector Machines, Proc. Pattern Recognition, Vol.2, pp.712-715 (2000)

[Hirokawa2012] Hirokawa, S., Feature Extraction using Restricted Bootstrapping, Proc. International Symposium on Innovative E-Services and Information Systems (IEIS 2012), pp.283-287 (2012)

[Komachi2008] Komachi,M., Kudo, T., Shimbo, M., Matsumoto, Y., Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms, Proc. EMNLP 2008, pp.1011-1020 (2008)

[小町 2010] 小町守, 工藤拓, 新保仁, 松本裕治, Espresso 型ブートストラッピング法における意味ドリフトのグラフ理論に基づく分析, 人工知能学会論文誌 Vol.25, No.2, pp.233--242 (2010)

[Nonaka2010] Nonaka, H., Kobayashi, A., Sakaji, H., Suzuki, Y., Sakai, H., Masuyama, S., Extraction of the effect and the technology terms from a patent document, Proc. Computers and Industrial Engineering(CIE), pp.1-6 (2010)

[Pantel2006] Pantel, M., Pennacchiotti, M., Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, Proc. ACL 2006, p.113-120 (2006)

support vector machine based classifications of reflectance data Optics Express, Vol. 19, No. 27, pp. 26816-26826 (2011)

[Radev2008] Radev, D., Mihalcea, R., Network and Natural Language Processing, AI Magazine, pp.16-28 (2008)

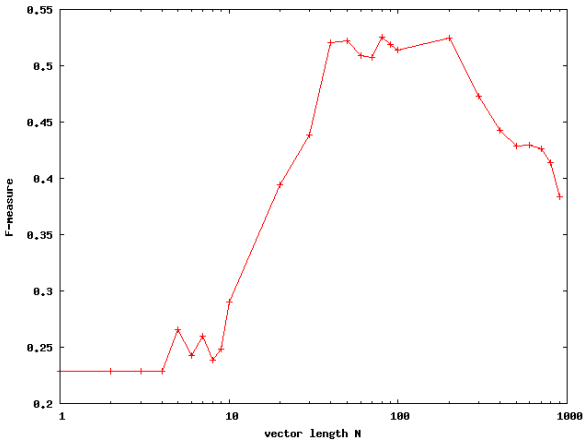
[Sadamitsu2011] Sadamitsu, K., Saito, K., Imamura, K., Kikui, G.,

Entity Set Expansion using Topic information, Proc. ACL 2011, pp.726-731 (2011)

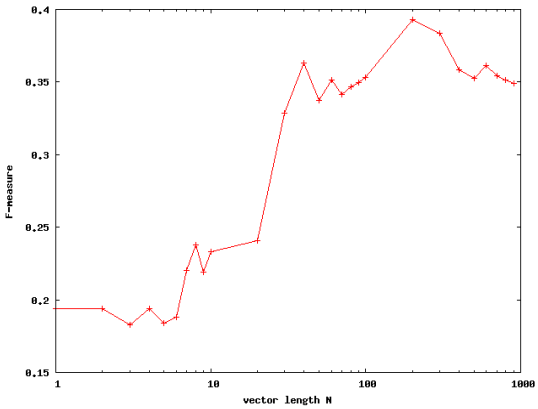
[Angrosh2010] Angrosh, M. A., Craneeld, S., and Stanger, N.: Context identication of sentences in related work sections using a conditional random field:towards intelligent digital libraries, Proc. of the 10th annual joint conference on Digital libraries, JCDL '10, pp. 293-302 (2010)

[Sakai2012] Toshihiko Sakai, J. Zeng, B. Flanagan, T. Nakatoh and S. Hirokawa, Descriminant Words for Problems in Scientific Articles, Proc.IIAI/ACIS International Symposium on Innovative E-Services and Information Systems (IEIS 2012), pp.267-271 (2012)

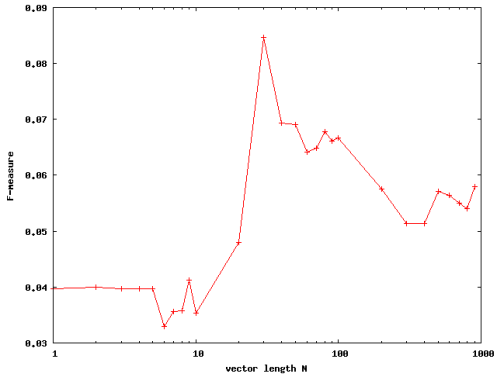
付録　選択属性数と F 値の関係



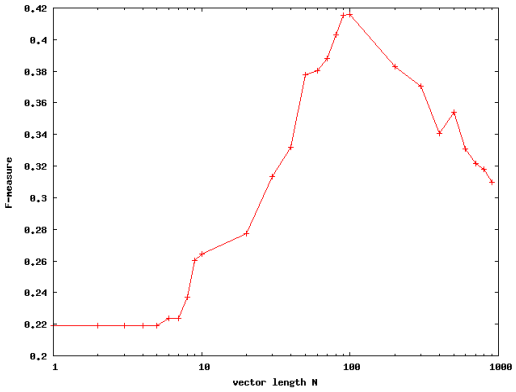
(a)　背景



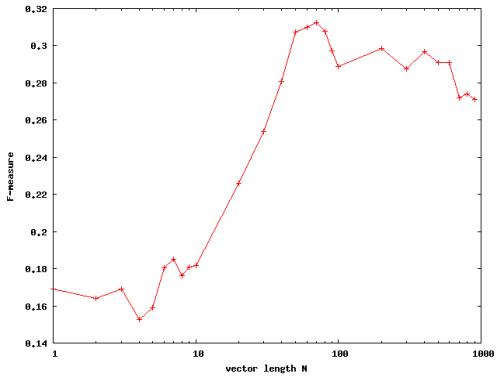
(b) 問題



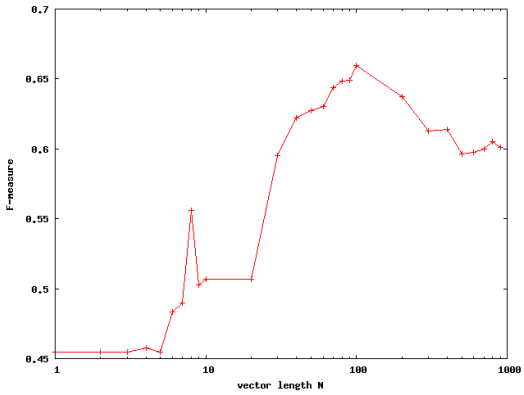
(c)　関連研究



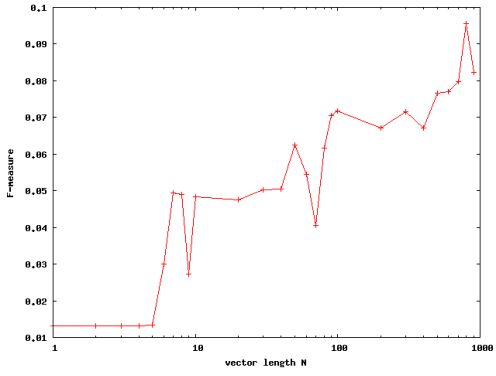
(d) 目的



(e)　手法



(f) 結果



(g) その他