学術論文の効率的利用のための意味的構造化

酒井佑太† 杉木健二† 松原茂樹‡

†名古屋大学大学院情報科学研究科
‡名古屋大学情報連携基盤センター

Semantically Structuring for Efficient Use of Electronic Technical Papers

Yuta Sakai† Kenji Sugiki† Shigeki Matsubara‡

†Graduate School of Information Science, Nagoya University
‡Information Technology Center, Nagoya University

1 はじめに

近年，WWW 上への公開や CD-ROM による配布をはじめとして，学術情報の電子的な流通が進みつつある．学術機関リポジトリや国立情報学研究所などの機関では，学術論文を WWW 上で提供しており，また，論文を CD-ROM で配布している学会や研究会も多い．これらの論文は貴重な知的資源であり，効率的な利用が求められている．例えば，論文データベースに対して，

1. 「論文検索システムの開発」を目的とする論文を探したい
2. 論文を作成するうえで，実験について記述する際，他の論文を参考にしたい

などの要求が存在していると考えられる．

しかし，学術論文検索システムの多くは書誌情報のみによる検索であり，上記の要求に対して満足のいく検索結果を得ることは難しい．このような要求に応えるために，論文の構成が意味的に構造化されている必要がある．例えば，どの章が，あるいは，どの段落が何について記述しているのかをタグ付けしておくことにより，先の要求に対して以下のように応えることが可能となる．

1. 「目的」について記述されている段落で「論文検索システムの開発」が記述されている論文を検索する
2. 「実験」について記述されている章を検索する

学術論文が意味的に構造化されることにより，論文検索システムや論文作成支援システムなどにおいて，学術論文の効率的利用が可能となる．

そこで本論文では，学術論文の効率的利用のための意味的構造化手法を提案する．論文の意味的構造化を，表層的構造化と意味タグ付与の 2 段階で実現する．表層的構造化では，PDF ファイルから論文中の章や段落などの論理要素を構造化する．意味タグ付与では，表層的構造化によって構造化された論理要素に対し，記述内容によって意味タグを付与する．

本手法では，テキスト分類の手法を用いて意味タグを付与する．段落をテキスト，意味タグをクラスと見なすことにより，意味タグの付与をテキスト分類として捉えることができる．分類器には SVM を多クラス分類器に拡張して用いた．

本手法の有効性を評価するために実験を行った．実験では，論文の第 1 章を対象とした．これは，第 1 章の意味的構造は，多くの論文において共通しているためである．実験の結果，タグ付与の精度は 48.9%，再現率は 47.2%，論文ごとの精度は 53.7%，再現率は 51.6%であった．

本論文の構成は以下の通りである．2 章で論文の構造について述べ，3 章で意味的構造化プログラムについて説明し，4 章で実験結果とその考察を行う．最後に 5 章で，まとめと今後の課題について述べる．

2 論文の意味的構造化

意味的構造化とは，論文中の各論理要素を取り出し，何について記述されているかを意味付けすることである．本論文の第 1 章に対する意味的構造化のイメージを図 1 に示す．学術論文の多くは構成がある程度共通しており，章や段落などの論理要素と，意味的構造との間の関係性が強い文書である．また，論理要素の意味内容もある程度限定されていると考えられる．そのため学術論文は，一般的なテキストに比べ意味的な構造をとらえやすいという特徴をもつ．

本研究では，論文の意味的構造化を，表層的構造化と意味タグの付与の 2 段階で考える．表層的構造化では，論文中のタイトルや，章，節，段落，図表，脚注，数式などの論理要素を抽出し構造化する．意味タグの付与では，抽出された論理要素の記述がどのような意味を持つのかを解析し，内容に即した意味タグを付与する．

2.1 表層的構造化

近年，WWW 上での公開や CD-ROM での配布などをはじめとして学術論文が電子的に流通するようになってきた．しかし，流通する学術論文のほとんどは PDF ファイルや PS（Postscript）ファイルである．これらのフォーマットには，LaTeX のような表層的構造が含まれていない．そのため意味的に構造化するにはまず，表層的構造化が必要となる．

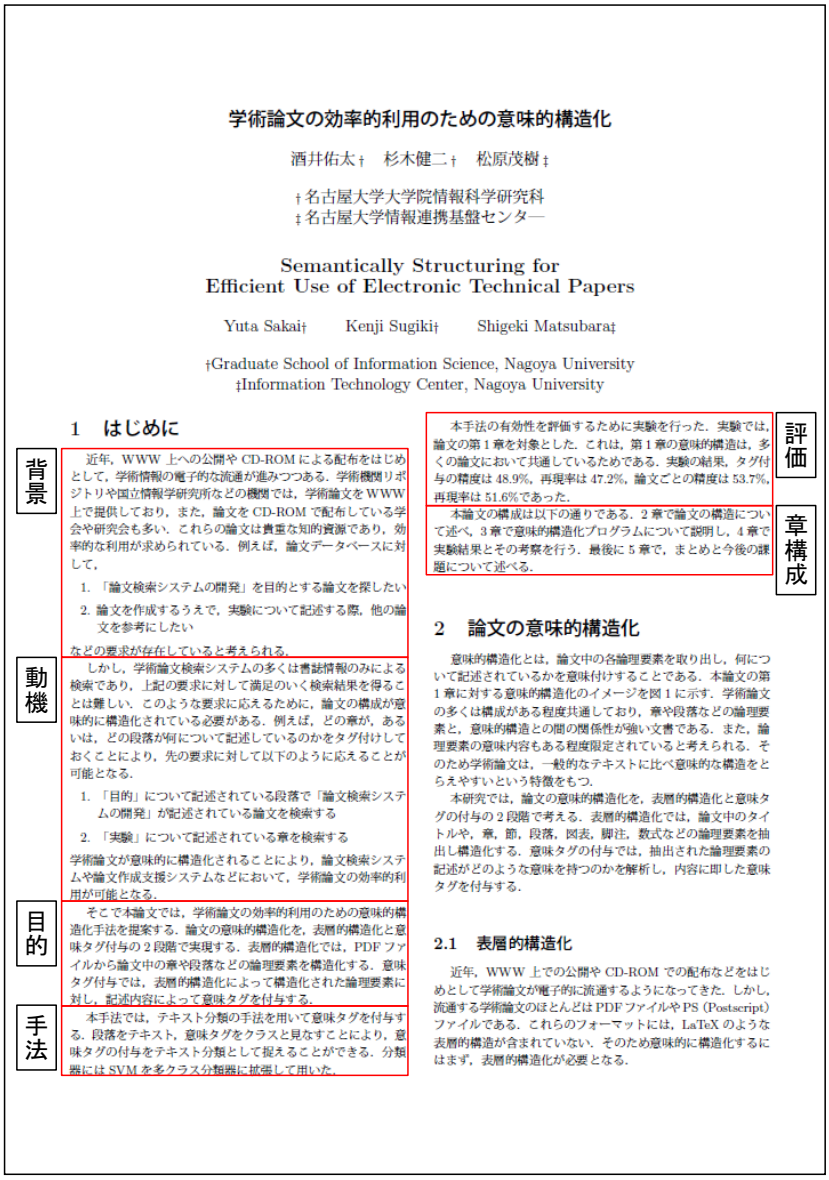

図 1: 意味的構造化のイメージ

我々はこれまで，レイアウト情報とテキスト情報を用いた表層的構造化の方法を提案している [1]．この手法では，PDF ファイルを pdftohtml[5] で XML 形式に変換し，その XML ファイルをもとに，記述された文の位置やフォントサイズなどの情報を用いて論理要素を抽出し，XML 形式で出力している．例として，本論文の表層的構造化により生成される XML ファイルを図 2 に示す．

2.2 意味的構造化

本研究では，表層的構造化によって抽出された論理要素に基づいて意味的構造化を行う．章や段落に意味内容を表す意味タグを付与する．

「はじめに」にあたる第 1 章中の段落に対する意味タグセットの例を表 1 に示す．本研究ではタグ付与の最少単位を段落としている．本来，文単位での意味タグの付与が望ましいが，実際には，意味内容の境界検出や文間関係の解析が必要となり困難である．一方，段落単位であれば意味内容に関わる情報が多く，表層的な素性でも意味タグの付与が現実的と考えられる．また，段落単位での意味タグの付与が実現すれば，意味タグを用いた段落単位での検索も可能となり，現状の論文検索システムに比べて有用性が高まることが予想される．

背景と動機が一つの段落で語られる場合や，一つの意味内容が複数の段落にまたがる場合が存在する．また，定義したタグセットに当てはまらない意味内容が語られることはほとんどない．上記の観察に基づき，タグの付与に以下の制約を設ける．

- 一つの段落に複数のタグを付与してもよい
- 一つのタグを複数の段落に付与してもよい
- あらゆる段落にタグを付与する

```
<conference-paper filename="sample.xml">
  <head>
    <title>学術論文の効率的利用のための意味的構造化</title>
    <title-en>Semantically Structuring for Efficient Use of Electronic Technical Papers</title-en>
    <author>酒井 佑太</author>
    <author>杉木 健二</author>
    <author>松原 茂樹</author>
    <author-en>Yuta Sakai</author-en>
    <author-en>Kenji Sugiki</author-en>
    <author-en>Shigeki Matsubara</author-en>
  </head>
  <body>
    <section>
      <st>はじめに</st>
      <p>近年，WWW 上での公開や CD-ROM での配布などをはじめとして学術情報が電子的に流通するようになってきた．
        電子図書館や国立情報学研究所をはじめとする多くの機関では，
        学術論文をWWW上で提供しており，論文をCD-ROMで配布している学会や研究会も多い．
        これらの論文は貴重な知的資源であり，効率的な利用がもとめられている．
        例えば下記のような要求が存在していると考えられる．
        <ol>
          <li>目的が「論文検索システムの開発」である論文を探したい</li>
          <li>論文作成中，実験の書き方について他の論文を参考にしたい</li>
        </ol>
      </p>
      <p>近年，WWW 上での公開や CD-ROM での配布などをはじめとして学術情報が電子的に流通するようになってきた．
        電子図書館や国立情報学研究所をはじめとする多くの機関では，
        学術論文をWWW上で提供しており，論文をCD-ROMで配布している学会や研究会も多い．
        これらの論文は貴重な知的資源であり，効率的な利用がもとめられている．
        例えば下記のような要求が存在していると考えられる．
        <ol>
          <li>目的が「論文検索システムの開発」である論文を探したい</li>
    ……
```

図 2: 表層的に構造化された XML ファイルの例

表 1: 意味タグセット

タグ名	意味
background	研究の背景
motivation	研究の動機
purpose	論文の目的
technique	手法の概要
evaluation	評価の概要
organization	論文の構成

論文を意味的に解析する研究は他にも行われている．佐波ら [2] は，論文の序章から質問応答のための知識を取り出すために，序論に含まれる情報として「目的」,「問題点」,「背景」,「手法」,「必要条件」の 5 種類を定め，それらの抽出を試みている．しかし，文単位での知識抽出は文間関係の複雑さから現実的でない．また，佐波らは取り出した知識による質問応答を目的としているが，我々は論文全体の構成を意味的に構造化することにより，論文検索や論文作成支援など，目的を限らない学術論文の効率的利用を目指している．

3 分類器による意味タグの付与

本研究では意味タグ付与の問題をテキスト分類として考える．テキスト分類とは，テキストをあらかじめ用意されたクラスに分類するタスクである．段落をテキスト分類におけるテキストとして，意味タグをクラスとして考えることにより，意味タグ付与のタスクをテキスト分類として捉えることができる．

本研究では，一つの段落に複数の意味タグが付されることを許可するため，ソフトクラスタリングを適用する必要がある．また，各段落は，どれかの意味タグに必ず分類される必要がある．

本手法では SVM[4] をテキスト分類として意味タグの自動付与を行う．SVM はもともと二値分類器であるため，多クラス分類に拡張する必要がある．SVM の多クラス分類への拡張方法として以下の代表的な二つの手法がある [3].

- one vs. rest 法
  - k 個の各クラスに対して，あるクラスか，それ以外かという 2 分類器 $f_c(x)$ を k 個構築する
  - ある分類対象 x に対し $f_c(x)$ が最大となる分類器に対応するクラスを x のクラスとする
- one vs. one 法
  - k 個のクラスから任意の 2 クラスに対する 2 分類器を，全組み合わせ数 ${}_kC_2$ 個構築する
  - ある分類対象 x に対してすべての分類器により分類を行う
  - その分類結果を統合して x の分類クラスを決定する

本研究では one vs. rest 法を用いる．ただし，複数のタグ付与を許可するため one vs. rest 法を拡張し，ある段落 p に対して $f_c(p)$ が正の数となる分類器に対応するクラスの意味タグを付与する．また，全ての段落に必ずタグを付与するために，全ての分類器で $f_c(p)$ が負の数の場合は，その中で $f_c(p)$ が最大となる分類器に対応するクラスの意味タグを p に付与する．

SVM の素性として，Bug of words を用い段落中の単語頻度を与える．ただし，1 文字の単語と，全正解データ中で頻度 1 の単語，記号や数字は除く．また，単語にはステミングを行う．

4 評価実験

4.1 実験概要

本手法のタグ付与の有効性を評価するために，評価実験を行った．データとして，ACL2007 の 40 論文を用いた．データには，第 1 章の抽出と意味的構造化を人手で行った．意味的構造化は表 1 のタグセットをもとに行った．

学習と評価には 40 分割の交差検証法を用いた．すなわち，全 40 論文中，39 論文を学習セット，1 論文をテストセットとし，学習と実験を合計 40 回試行した．SVM 分類器として TinySVM を用い，カーネルは多項式カーネルを用いた．

評価は以下の二つの観点で行った．

- タグごとの評価

  各意味タグ $t_i(1 \leq i \leq 6)$ ごとの性能を，タグ付与の精度および再現率によって評価した．

$$\text{精度 (\%)} = \frac{\text{正解した段落数}}{\text{タグ } t_i \text{を付与した段落の数}} \tag{1}$$

$$\text{再現率 (\%)} = \frac{\text{正解した段落数}}{\text{正解データでタグ } t_i \text{が付与された段落の数}} \tag{2}$$

- 論文ごとの評価

  各論文の第 1 章に含まれる段落に対して，正しく意味タグが付与されたかどうかを，精度および再現率により評価を行った．

$$\text{精度} = \frac{\text{正解したタグの数}}{\text{論文に付与したタグの数}} \tag{3}$$

$$\text{再現率} = \frac{\text{正解したタグの数}}{\text{正解データの論文に付与されたタグの数}} \tag{4}$$

表 2: 論文単位の評価

	精度 (%)	再現率 (%)	F 値
平均	53.7	51.6	0.512

表 3: タグ単位の評価

	精度 (%)		再現率 (%)		F 値
background	55.7	(34/61)	70.8	(34/48)	0.624
motivation	42.6	(23/54)	37.1	(23/62)	0.397
purpose	54.0	(34/63)	59.7	(34/57)	0.567
technique	27.3	(6/22)	21.4	(6/28)	0.240
evaluation	22.2	(2/9)	12.5	(2/16)	0.16
conclusion	91.7	(22/24)	81.5	(22/27)	0.863
平均	48.9		47.2		0.475

4.2 実験結果

論文単位の評価を表 2 に，タグ単位の評価を表 3 に示す．実験の結果，タグごとの評価として 47.2%の再現率と 48.9%の精度，論文ごとの評価として 53.7%の再現率と 51.6%の精度を得た．精度・再現率は同程度の評価となっている．単語のみを素性とした SVM 分類器を用いることで，この程度の性能を実現できることは，各タグ間で単語の分布がある程度異なっており，段落の意味を解析する際に，単語を利用することが有用であることを示している．

タグの種類別に F 値を比べると，evaluation が最も低く，organization が最も高い．

evaluation タグは表 4 に示すとおり，正解データ中での付与数が最も少ない．これにより学習量が不足していることが原因として考えられる．また，同一の段落に他のタグと重複してタグを付与しているため，他のタグとの単語の分布が似てしまうためと考えられる．表 5 に，タグごとにそれが他のタグと重複して付与された割合 (**重複付与率**) を示す．表 5 に示すとおり，正解データ中で evaluation は他のタグと同時に付与されることが多い．これにより evaluation に対する分類のための素性として単語のみでは不十分であると考えられる．

organization タグは F 値が最も高いが，正解データ中での付与数は evaluation の次に少ない．これは organization のほとんどが，以下のような特徴的なフレーズを含んでおり，学習量が少なくても，これらの表現に含まれる単語を素性として学習した．結果，F 値が高くなったものと考えられる．

- The structure of the rest of the paper is as follows.
- The remainder of the paper is organized as follows.
- Section 2 describes ...
- Section 3 presents ...

本手法では，意味タグを付与する際，その順序は考慮しなかった．しかし，今回使用したデータ 40 論文のうち 36 論文は，「background」,「method」,「purpose」,「technique」,「evaluation」,「organization」の順でタグが付与されていた．40 論文のうち 19 論文に対するタグ付与結果はこの順序に従っていなかった．この順序情報を制約として用いることにより，更なる性能の向上が期待できる．

タグ	付与数
background	48
motivation	62
purpose	57
technique	28
evaluation	16
organization	27

表 4: 各タグの付与数

タグ	重複付与率	
background	18.8	(9/48)
motivation	27.4	(17/62)
purpose	49.1	(28/57)
technique	50.0	(14/28)
evaluation	68.8	(11/16)
organization	25.9	(7/27)

表 5: 重複して付与されたタグの割合

5 まとめ

本論文では，論文の意味的構造化について論じ，機械学習による意味タグ付与手法を提案した．本手法では，意味タグ付与タスクをテキスト分類タスクとしてとらえ，機械学習による分類器を用いて意味タグの付与を行う．本手法の有効性を評価するために，40 分割交差検証法により英語論文 40 論文を用いて意味タグの付与実験を行った．実験の結果，タグごとの評価として 47.2%の再現率と 48.9%の精度，論文ごとの評価として 53.7%の再現率と 51.6%の精度を得られた．単語のみを用いた単純な手法としては良好な結果であると考えられる．

今後の課題としては，意味タグ付与の連続性や，順序制約などについても考慮することにより，意味タグ付与の性能向上を図る予定である．

参考文献

[1] 杉木 健二, 松原 茂樹, 吉川 正俊: レイアウト情報とテキスト情報を用いた学術論文の構造化, 電気関係学会東海支部連合大会講演論文集 (2005).

[2] 佐波智也, 日高宏紀, 渡辺靖彦, 岡田至弘: 論文検索のための知識の論文の序論からの抽出, 言語処理学会第 13 回年次大会 (2007).

[3] 山田 寛康, 松本 裕治: Support Vector Machine の多値分類問題への適用法について, 情報処理学会研究報告, NL-146-6, pp.33-38 (2001).

[4] TinySVM, http://chasen.org/ taku/software/TinySVM/

[5] pdftohtml, http://pdftohtml.sourceforge.net/