MEDLINE 概要文の役割分類のための信頼度の異なるデータを用いた学習

ナタリー·アイゼンバーグ† 新保 仁† 原 一夫† 松本 裕治†

† 奈良先端科学技術大学院大学 情報科学研究科
E-mail: †{natali-a,shimbo,kazuo-h,matsu}@is.naist.jp

あらまし 医学生物学論文アブストラクト (概要) の各文は, 大きく分けて, 研究の背景, 目的, 実験手法, 結果および結論に分類できる. このような文の構造的な役割の推定は, 情報検索の際に, 絞り込みの手がかりとして用いることが可能である. 先行研究では, 文内および文脈の情報を表す種々の素性を用いて文の役割分類を行った結果, 高い精度を得たと報告されている. ただし, その際には十分な量の学習データを確保するために, 実際に運用されるデータとは異なる学習データ収集方法が取られており, このことに起因して, 運用データに対する精度が, (学習データと同じ分布のデータに対して適用した場合と比較して) 低い数値に留まったことが報告されている. 本論文では, この問題に対し事例の “データソース” に依存して, 異なるコストを割り当てることによって解決を試みる.

Learning from Data of Varying Quality for Sentence Role Identification in MEDLINE Abstracts

Natalia AIZENBERG†, Masashi SHIMBO†, Kazuo HARA†, and Yuji MATSUMOTO†

† Graduate School of Information Science
Nara Institute of Science and Technology
E-mail: †{natali-a,shimbo,kazuo-h,matsu}@is.naist.jp

Abstract The abstract of a scientific paper typically consists of sentences describing the background and objective of the study as well as its experimental methods and results, and conclusions. There has been an increasing interest in recent years in identifying such structural roles, with particular motivations from the information retrieval point of view. In previous research done with respect to MEDLINE abstract, various sentence feature combinations were used in order to achieve successful performance, but one important issue has not yet been addressed: the unrepresentativeness of the major part of learning data; i.e., the learning set samples tend to originate from different sources baring many differences, while the application data source distribution does not necessarily obey that of the learning set. In this paper we solve this issue by applying “example source” sensitive costs in the training process.

1. はじめに

医学生物学分野においては, 文献数の急増とそれにともなう文献検索の効率化の必要性から, 論文概要 (アブストラクト) を, あらかじめ修辞構造的な役割ごとに関連性のある “節” に分割して記述することが提唱されている (構造化アブストラクト; structured abstract). ここでいう修辞構造役割とは, たとえば, BACKGROUND (背景), OBJECTIVE (目的), METHOD (実験手法), RESULTS (実験結果), CONCLUSION (結論) である. 包括的な文献調査を基礎とする evidence-based medicine の立場からの需要といった背景もあり, 構造化アブストラクトは 1990 年台以降, 数多くの医学生物学系学術誌で採用されている.

とはいえ, 現在でも半数以上の医学論文や, 大多数の医学以外の分野の論文アブストラクトは, 構造が明示されていない, 非構造化アブストラクトである. もっとも, これら構造化されていない一般のアブストラクトにおいても, 多くの場合, アブストラクトは典型的な修辞構造を持ち, 各文は, 背景, 目的といった, 構造的な役割を担うことを意図して執筆されている.

文章を構成する各文が, その文章内でどのような構造的役割を担っているかを明らかにすることの有効性は, 自動要約や, 情報検索といった自然言語処理タスクにおいて示されている. し

かし, こういったタスクが対象とする文書は一般に数が膨大であり, 人手で対象文書の文すべての役割分類を行うことは, それがたとえアブストラクトのような比較的短い文章であったとしても, 現実的ではない. このため, 本研究では, 非構造化アブストラクトの各文に対して, その構造的役割の自動推定を試みる.

2. アブストラクト文の構造役割推定

2.1 研究の意義と想定される応用

アブストラクトの構造を自動的に明らかにすることは, 文献検索時の選択肢を増やし, 情報の効率的な検索手段の実現につながる. ユーザの検索の動機や, 所望の情報の種類に応じて, 各構造的役割を持つ文の重要度は異なる, というのが, その理由である. 例えば, 臨床医がある病気に対する治療法の有効性について知りたい場合, OBJECTIVE (目的) を記述した文により重みを付けた検索を行なう, といった状況が考えられる. そうすることによって, その病気が単なる副作用としてアブストラクトの RESULTS (実験結果) 部分に述べられている文献を検索結果から除外する, あるいは表示順位を下げることが可能になる.

PubMED [14] のような大規模文献検索エンジンに, 数個程度の検索キーワードのみをあたえると, 数十件もの文献が該当することが多々あるが, 単にキーワードを追加するだけで, 上記の臨床医の例にあるような絞り込みを実現することは大抵の場合困難であり, また, たとえ絞り込みに有効なキーワードがあったとしても, それがユーザにとって常に明らかとは限らない. 検索対象とする文の構造的役割を選択肢として提示することは, ユーザに代替の絞りこみ手段を与え, よりユーザの目的に即した検索を実現するための一助となる.

実際, Tbahriti ら [18] は, 類似論文検索のタスクにおいて, 特定の修辞的役割を持つ文の使用は, それ以外の文を用いた場合に比べて良い精度が得られることを示している. 彼らは, OBJECTIVE および CONCLUSION に属する文の類似度が高い論文は, EXPERIMENT の類似度が高い論文よりも, 内容的な類似度がより高いことを報告している.

さらに, Tuefel と Moens [19] は文の修辞的構造が, 文書要約にとっても有用であることを実証した. 彼らは, 抽出による要約 (extractive summarization) を行う際に, 文書の修辞構造のうち “新しい” 情報を重点的に抽出することが, 対象論文の成果を要約する上で有効であることを示したが, そういった情報は, BACKGROUND よりも OBJECTIVE や METHOD により出現しやすい.

2.2 問題設定

本研究で扱う “文の修辞構造役割推定” 問題は, 与えられたアブストラクトを構成する一連の文が, あらかじめ決められた複数の修辞構造役割のクラスのうちどれに属するかを決定する問題である.

分類先の役割クラス (集合) はあらかじめ固定しておく必要があるが, どのようなクラスを用意するかは, 先行研究においても相違が見受けられる. Graetz [7] はアブストラクトは, PROBLEM (INTRODUCTION), SOLUTION, RESULTS, CONCLUSION の 4 クラスに分割するのが自然と論じている.

一方, Salanger-Meyer [15] の分析によると, 論文コーパスに見受けられる傾向は Graetz の分類とは必ずしも一致しない. さらに Orasan's [13] の網羅的調査は, Graetz による提案よりもさらに詳細な分割の必要性を示している.

本研究では, Orasan による役割クラス分類を踏襲する. 同様の分類は, MEDLINE に収録の構造化アブストラクトの統計に基づく Yamasaki らの研究 [17] においても用いられている.

Mizuta と Collier [12] は, (アブストラクトではなく論文の本文を対象としている, という相違はあるが), 情報検索をアプリケーションとして想定した場合, 修辞構造分類においては, “古い” 情報と “新しい” 情報を区別することが重要であると論じている. これは, Graetz の分類における PROBLEM (INTRODUCTION) クラスを (主に古い情報について述べてある) BACKGROUND と (新しい情報を含む) OBJECTIVE に細分した, Orasan と Yamasaki らによる分類と合致する.

以上まとめると, われわれの役割クラス分類は, BACKGROUND, OBJECTIVE, METHOD, RESULTS, CONCLUSION の 5 種類となり, 与えられたアブストラクト内の文がこれらのうちどれに分類されるかを自動判別することが目的となる.

2.3 対象データ

本研究が対象とするアブストラクトは, MEDLINE [11] データベースからの抜粋したものである. MEDLINE は数千の学術誌に収録された 1400 万以上の文献からなり, 現存する医学生物学文献データベースのうち最大規模を持つ一つである.

MEDLINE に収録のアブストラクトは, その修辞構造がアブストラクト中に明示されているか否かに応じて, **構造化アブストラクト** (structured abstract) [2] と, **非構造化アブストラクト** (unstructured abstract) に分類できる. 構造化アブストラクトは, 修辞的構造がアブストラクト中に, 見出しとして明記されたアブストラクトである. したがって, ある見出し以降 (次の見出しまで) の文は, その見出しに記された役割を担っていることが容易にわかる. また, 非構造化アブストラクトとは, 構造化されていない通常のアブストラクトを指す.

文の構造役割推定タスクにおいては, おのずと (役割の明示されていない) 非構造化アブストラクトが推定の対象となる. 構造化アブストラクトは, 急速に普及しはじめているとはいえ, 全体としてみると, その数はいまだ非構造化アブストラクトにはるかに及ばないことから, 文の構造役割を自動推定する意義は失われていない.

本研究では, 構造役割判定のために, 教師つき機械学習によるアプローチをとる. アブストラクトという限定された文書を対象にしているとはいえ, 自由度の極めて高い自然言語文には相違なく, 人手で構造役割識別規則を設計することは 現実的ではないためである. (教師つき) 機械学習のアプローチでは, 機械学習器を訓練するために, 一定量の訓練データが必要とされる. 本タスクにおいてこのデータに該当するのは, 人手で役割ラベルを付与した非構造化アブストラクトの文である. ただし, 十分な量の役割タグづき文の準備は (人手の介在を必要とすることから) 大きな費用がかかり, 実用化に際して障害となる.

この問題を避けるため, われわれは, 役割の明記されたアブストラクト, すなわち構造化アブストラクトを訓練データとして利用する. 文の役割を表す構造化アブストラクトの見出しは, 慣例として OBJECTIVE: のように, すべて大文字でコロン (:) をともなって表記されるため, 文役割は容易に自動抽出できる. しかしながら, 構造化と非構造化アブストラクト内の文の性質は必ずしも一致するとは限らず, 少なくとも下記の相異点がある.

（1） **情報の提示順の相違**. 構造化アブストラクトにおいては, 同一役割見出しが 1 個のアブストラクト中に 2 回以上出現することは極めてまれである. したがって同じ役割を持つ文は一箇所に “塊” として出現する. 非構造化アブストラクトでも同様の傾向が見られるが, 構造化アブストラクトほど確固とした傾向ではない. 複数個の実験について報告されている場合, 実験内容 (METHOD に該当) 結果 (RESULT) が交互に複数回表記されることは珍しくない.

（2） **文法的な相違**. 見出しの存在により, 構造化アブストラクトには

> OBJECTIVE: To assess the efficacy of [a treatment] on [a disease].

のように不定詞句単独からなる文がしばしば見受けられる. 非構造化アブストラクトにおいては, 見出しが存在しないためそのような文は許されない.

（3） **見出しの信頼性**. 役割見出しの選択を著者に一任している学術誌がある一方, あらかじめ特定の役割見出しを用いることを強制している学術誌がある. 前者においては, 著者によって見出しの解釈にばらつきが生ずる可能性があり, また, 後者においては, 強制された見出し (役割分割) と今回われわれが用いる 5 クラスとは整合性がない可能性がある. OBJECTIVE, METHOD, RESULTS , CONCLUSION の 4 クラスを強制している学術誌論文においては, OBJECTIVE 中に BACKGROUND に関する情報が含まれている場合が多々見受けられる. このような構造化アブストラクトの文をそのまま訓練データに用いると, ラベル付けノイズとなってしまう.

そこでわれわれは, (i) 見出しを手がかりに自動的に役割ラベルを付与した構造化アブストラクトの文と, (ii) 見出しが存在しないため, 人手によって役割ラベルづけされた非構造化アブストラクトの文, を訓練データとして併用する. ただし, 前述のとおり, 大量のタグ付けデータは作成が困難なため, (自動ラベル付けされた) 構造化アブストラクトと (人手でラベル付けした) 非構造化アブストラクトの比率は前者が圧倒的に多く, 後述の実験においてはその比率は 5:1 程度である (図 1).

一方, 機械学習によって訓練した分類器を適用する文 (実験評価におけるテストデータ) はすべて非構造化アブストラクトから取られたものである. 訓練データとテストデータの分布は必然的に大きく異なったものとなる. たとえば, 5. 章の実験で用いるわれわれのデータの分布は, 図 1 のようになっている.

要約すると, われわれが直面する状況は以下の通りである.

- 非構造化アブストラクトからの訓練データ (文) は作成に費用がかかるため十分な量は確保できない.
- 構造化アブストラクトからの文は, 自動処理により大量に得ることができるが, 本来の分類対象である非構造化アブストラクトの性質を必ずしも表現していない.

これら (信頼性は高いが少量の) 非構造化アブストラクトおよび (信頼性は低いが大量にある) 構造化アブストラクトを役割分類精度を高めるためにいかに利用するか, が, 以下本論文の主題である.

次章では, この問題に対するわれわれの提案について説明する.

3. 信頼性の異なるデータを用いた学習

人手で役割ラベルを付与された非構造化アブストラクトは高価であり, 一方, 構造化アブストラクトは, 入手は簡単であるが, 本来の分類対象である非構造化アブストラクトとは異なる性質を持つことがある [17].

われわれは, これら双方を役割分類器の訓練データとして最大限利用するための手法を提案する. 学習の際に, 信頼性の高いデータ (文役割分類タスクにおいては, 非構造化アブストラクト) をより重視し, 信頼性の低いデータ (同, 構造化アブストラクト) に対する分類精度を犠牲にしても, 信頼性の高いデータの分類精度を重んじる, というのが提案手法の基本的な考え方である.

少量の信頼性の高いデータと大量の信頼性の低いデータが存在し, 訓練データと適用 (テスト) データの分布が異なる状況は, 文役割分類タスク以外にも多数見受けられる. われわれの提案する手法は, そのような問題一般に適用可能な手法である.

3.1 Support Vector Machine

われわれの手法は Support Vector Machine (SVM) [20] への拡張である. SVM に対する差異を明らかにするために, まず Vapnik による SVM の定式化について簡単に述べる.

N 個の訓練データ $\{(x_i, y_i)\}_{i=1}^{N}$ が与えられたとする. ここで x_i は i 番目の事例の入力素性ベクトルであり, $y_i \in \{+1, -1\}$ はそのクラスラベルである ($y_i = +1$ は事例 i が正例, $y_i = -1$ は i が負例であることを示す). SVM は素性ベクトル x の線形関数

$$f(x) = w \cdot x - b$$

を学習する. この中で, w, b は学習によって最適化されるパラメタであり, $f(x)$ によって定まる識別関数

$$\mathrm{sgn}(f(x)) = \mathrm{sgn}(w \cdot x - b) \qquad (1)$$

が, すべての訓練事例に対して正しいラベルを出力するように (すなわち, すべての $1 \leqq i \leqq N$ に対して $\mathrm{sgn}(f(x_i)) = y_i$ が成り立つように) 学習が行われる.

線形分離可能な問題 (上記制約をみたす $f(x)$ が少なくとも一つ存在する問題) については, 一般に, すべての訓練事例 $x = x_i$ に関して 1 をみたす $f(x)$ は無限に存在する. SVM はそれら

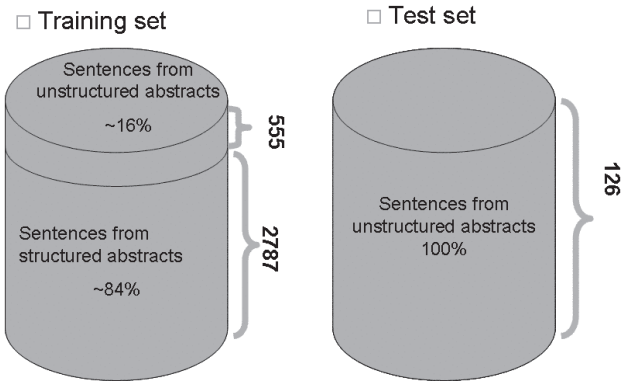

図 1　データソースの分布 (左側: 訓練データ, 右側: テストデータ).

の中から, “最大マージン原理” に基づいて, 2 クラス間の “マージン” が最大になる $f(x)$ を出力する. スケーリングによらない識別関数の一意性を保つため, 正例 x に対しては $f(x) \geqq +1$, 負例に対しては $f(x) \leqq -1$ が成り立つ, という制約を課すと, $f(x)$ によって定まる以下の二つの超平面 H_{-1} and H_{+1} を考えることができる.

$$H_{+1} : w \cdot x - b = 1,$$
$$H_{-1} : w \cdot x - b = -1.$$

マージンは, これら超平面間の距離として定義され,

$$2\frac{|w \cdot x - b|}{\|w\|} = \frac{2}{\|w\|}.$$

で与えられる. したがってこれら超平面間の距離の最大化は $\|w\|$ の最小化と等価であり, SVM におけるマージン最大化問題は,

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(w \cdot x_i - b) \geqq 1, \quad \forall i. \end{aligned} \tag{2}$$

という二次計画問題として定式化される.

実用上は, 訓練データ中に含まれるノイズ等の影響を軽減し, SVM の汎化性能を高めるために, 各事例 i に対する制約 $y_i f(x_i) \geqq 1$ にある程度の違反を許す**ソフトマージン SVM** が用いられる. ソフトマージン SVM の制約式は各事例に対して新規のパラメタ $\xi_i \geqq 0$ を導入し, $y_i f(x_i) \geqq 1 - \xi_i$ と緩和される, とする. 引き換えとして, 目的関数には, 違反した量 ξ_i の総和に応じた罰則を与える項が付加され, 以下のような最適化 (二次計画) 問題となる.

$$\begin{aligned} \underset{w,b,\xi}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C(\sum_i \xi_i) \\ \text{s.t.} \quad & y_i(w \cdot x_i - b) \geqq 1 - \xi_i, \quad \forall i \\ & \xi_i \geqq 0. \quad \forall i \end{aligned} \tag{3}$$

ここで $\xi = (\xi_1, \ldots, \xi_i \ldots, \xi_N)$, また, C は主目的関数 $\|w\|^2/2$ に対して制約違反をどの程度重視するに応じて事前に設定すべきハイパーパラメタである. 極限 $C \to \infty$ においては, ソフトマージン SVM の最適化問題 (3) は式 (2) で与えられる **ハードマージン SVM** と一致する.

3.2　信頼性が異なるデータに対する SVM の拡張

信頼性の高い訓練事例をより重視し, そうでない事例を相対的に軽視する, というバイアスを取り込むために, ソフトマージン SVM における制約に変更を加える. このためわれわれは, 全ての訓練事例に対して均一のハイパーパラメタ C を用いるのではなく, 事例に依存した値を C として用いる. 具体的には, 事例 i に対し

$$C(i) : \mathbb{N} \mapsto \mathbb{R}.$$

を

$$C(\text{信頼性の高い事例}) \gg C(\text{信頼性の低い事例})$$

が成り立つように定める. これを式 (3) 中の定数 C のかわりに代入し, 次式を得る.

$$\begin{aligned} \underset{w,b,\xi}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + \sum_i C(i)\xi_i \\ \text{s.t.} \quad & y_i(w \cdot x_i - b) \geqq 1 - \xi_i, \quad \forall i, \\ & \xi_i \geqq 0, \quad \forall i. \end{aligned} \tag{4}$$

Lagrangian 乗数 $\alpha = (\alpha_1, \ldots \alpha_i, \ldots, \alpha_N)$, $\mu = (\mu_1, \ldots, \mu_N)$ を (4) に導入し ($\alpha_i \geqq 0$, $\mu_i \geqq 0$), 主 Lagrange 関数

$$\begin{aligned} L(w, b, \xi, \alpha, \mu) = & \frac{1}{2}\|w\|^2 + \sum_i C(i)\xi_i \\ & - \sum_i \alpha_i [y_i(w \cdot x_i + b) + \xi_i - 1] \\ & - \sum_i \mu_i \xi_i \end{aligned} \tag{5}$$

を得る. 二次計画問題 (4) の解は, L の鞍点において与えられるので, w, b, ξ に関する勾配を 0 とおいて (5) に代入すると, 双対関数

$$L_D(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j. \tag{6}$$

が得られる. 主問題 (4) の解は, 双対関数を α に関して最大化する Wolfe 双対問題

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & L_D(\alpha) \\ \text{s.t.} \quad & 0 \leqq \alpha_i \leqq C(i), \quad \forall i \end{aligned} \tag{7}$$

の解と一致する.

本論文で扱う文の構造役割判定においては, $C(i)$ は事例 i ごとに決められるのではなく, そのデータソース (構造化アブストラクトか, 非構造化アブストラクトか) のみに依存するので,

$$C(i) = \begin{cases} C_u & \text{事例 } i \text{ が非構造化アブストラクトの文の場合,} \\ C_s & \text{事例 } i \text{ が構造化アブストラクトの文の場合} \end{cases} \tag{8}$$

と書ける. ただし, 非構造化アブストラクトの制約違反 (分類誤り) をより厳しく罰するため,

$$C_u \geqq C_s$$

とする. 特別な場合として, $C_s = 0$ の場合, 非構造化アブストラクトのみを訓練データとして用いた一般のソフトマージン SVM ($C = C_u$) と一致する.

4. 関 連 研 究

McKnight and Srinivansan [8] は, 本研究同様 MEDLINE アブストラクトを用いた文の役割分類を試みた. SVM を大量の構造化アブストラクトのみで訓練した場合, 少量の非構造化アブストラクトのみで訓練した SVM と同等の精度しか得られなかったことが報告されている.

Lin ら [10] は, アブストラクトの生成モデルを仮定して文役割同定に適用し, SVM を用いた McKnight と Srinivansan と同程度の精度を得たが, 有意な差を持って優位性を示すことはできなかった.

Yamasaki ら [17] は, 時制などを含むさまざまな文内素性の有効性について調査すると同時に, 彼らは SVM を主に構造化アブストラクトを用いて訓練し, 構造化アブストラクトおよび非構造化アブストラクト双方に適用した. その結果, さまざまな素性は有効であるが, 構造化アブストラクトに適用した場合の精度に比べ, 非構造化アブストラクトに適用した場合の精度の低下が報告されている.

提案したデータソース依存した訓練法は上記の問題に対する解決法を与える.

コスト依存学習法が Abe らによって提案されている [1]. 彼らもわれわれ同様, ソフトマージンハイパーパラメタ C を調整することを提案している. しかし, 彼らの提案は, 正例, 負例それぞれの分類誤りに基づいてコストを変化させており, データソースへの依存性は考慮されていない.

Brefeld らは [3] 事例依存コスト学習の理論的枠組を与えている. 彼らが取り組んだ問題は本論文同様, 信頼性の低いデータが訓練データに含まれている場合である. 彼らは, 事例別およびクラス別のコストを導入した一般的な SVM 学習を提案しているが, 実験で用いられた問題は, 次元数の低い人工的な問題であり, 実際の応用についての議論はない.

5. 実　　験

5.1 実 験 設 定

前に述べたとおり, 本実験では, 修辞構造上の役割を表す 5 クラス, (BACKGROUND, OBJECTIVE, METHOD, RESULTS, CONCLUSION) に, 与えられた文を分類する.

訓練データは 3342 のラベルつき文からなる. うち 2787 文は構造化アブストラクトから, 見出しを元にラベル付けを自動的に行ったものであり, 残り 555 文は人手によってラベル付けを行った非構造化アブストラクトの文である. 訓練データ中の役割の分布を表 1 に示す. 一方のテストデータは非構造化アブストラクトの文のみで構成される.

訓練データとテストデータは以下のように作成した. MEDLINE 2003 から抽出した 2002 年発表論文の 103813 アブストラクトから, さらに 4000 文を抽出し, この中に含まれる非構造化アブストラクトについては, 人手によって役割クラスラベルを付与した. このうち, 80% を訓練データとし, 残り 20% をテストデータ候補とした. テストデータ候補から構造化アブストラクトをすべて取り除いた結果, テストデータ 126 文が残った.

役割分類タスクは 5 クラスの多値分類問題であるため, 二値分類器である SVM をそのまま適用することはできない. このため, SVM を多値分類問題に適用するための標準的な方法のひとつである多数決法 (pairwise voting method) [16, Section 7.6] を採用した. この方法では, まず, すべての 2 クラスの組合せに対応する合計 $\binom{5}{2} = 10$ 個の二値分類器を作成する. 二値分類器の訓練の際には, 副問題に関連する 2 クラスに属する訓練データのみを用いる. テストデータに対して適用する際には, これらを二値分類器を同一データ (文) に対して適用して得られる 10 通りの二値分類結果の多数決によって最終的なクラスを決定する.

SVM の実装には, libSVM [4] に 3.2 節で説明した提案手法のための変更を加えたものを用いた. われわれは, 個々の二値分類 (合計 10 個) と, 最終的な多値分類の両方について精度を評価する. 多値分類については libSVM に付属の多値分類機能を用いず, 二値分類の結果を組み合わせるラッパーを別途準備して評価を行った. これは, libSVM に内蔵の多値分類ラッパーがどのように多値分類を行っているかについて詳細が得られず, 提案手法と整合性があるか確認が取れなかったためである.

事例を (素性) ベクトル表現するために, 以下の素性 (特徴量) を用いた.

（1） 表層単語
（2） 表層単語の原型 (lemmas)
（3） 品詞情報
（4） 上記 (1)–(3) のすべての可能な組合せ 4 通り.
（5） 連続する表層単語の二つ組 (bigram).

素性はすべて二値素性である. すなわち, 可能な限りの上記項目 (例: 単語) について個別のベクトルの次元を割り振り, その項目が, 文内に出現すれば素性ベクトルの該当次元 (要素) は 1, 出現しなければ 0 とした.

原型と品詞情報の取得には, 医学生物学分野を念頭に開発された GENIA Tagger [6] を用いた.

5.2 ベースライン

本研究と先行研究では, 実験に用いるデータや, 役割クラス分割が異なっているため, 直接の比較は困難である. このためま

表 1　訓練データ内の役割分布 (文の割合)

データソース	BACKGROUND	OBJECTIVE	METHOD	RESULTS	CONCLUSION	全体
非構造化アブストラクト	11.5%	7.9%	18.9%	47.7%	13.8%	100%
構造化アブストラクト	13.3%	10.0%	24.5%	35%	16.6%	100%
全訓練データ	13.0%	9.6%	23.5%	37.5%	16.1%	100%

表 2　ベースラインの多クラス分類精度

	ベースライン (1) 全訓練データを使った SVM ($C_u = C_s$ に相当)	ベースライン (2) 非構造化アブストラクトのみを訓練に 使った SVM ($C_s = 0$ 相当)
Accuracy%	64.2	67.4

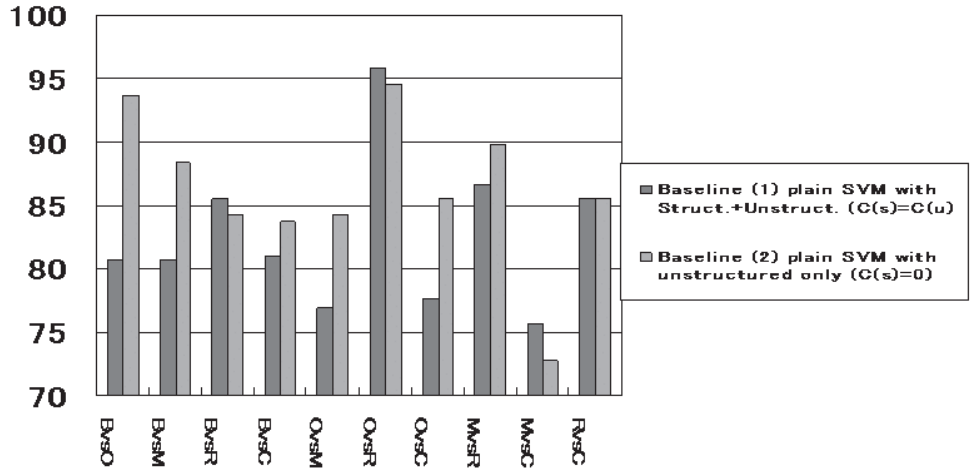

図 2　SVM による二値分類のベースライン. B = BACKGROUND, M = METHOD, O = OBJECTIVE, R = RESULTS, C = CONCLUSION.

ず, 先行研究で用いられている通常のソフトマージン SVM に基づく手法を, 前節で説明したわれわれの実験データに適用した. 後の実験では, これらベースラインを提案手法の評価の基準として用いる. 今回は, 通常の SVM を用いた二つのベースラインを用意した.

（1） 全訓練データを用いた SVM. 構造化アブストラクト/非構造化アブストラクトのどちらに対しても C は等しい (すなわち $C_u = C_s$).

（2） 非構造化アブストラクトの訓練データのみを用いた SVM. これは $C_s = 0$ の場合と等価である.

両方のベースラインと, 各二値分類問題の組合せそれぞれについて, 最適な C の探索を行った.

パラメタ C の値域は $[0, \infty]$ と広範に渡るため, $C = 10^n$ とおいて $n = -10, -9, \ldots, 0, \ldots, 10$ の範囲で最適な C の探索を行った. このように各二値分類問題について最適な C を求めた後で, 各々の二値分類問題における最適な分類器の出力を用いて多数決を行い, 5 クラス多値分類を行った.

表 2 に示すとおり, 最終的な多クラス分類では, 非構造化アブストラクトのみを用いた場合が, 非構造化に構造化を併用した場合を上回った. しかしながら, 表 3 に示した, 個別の二値分類精度を見てみると, 10 個の二値分類問題のうち, 三問題については, 後者が前者を上回っていることが見て取れる.

5.3　最適な C_s, C_u パラメタの設定

提案手法では, 二種類のアブストラクト (構造化および非構造化) のそれぞれに対し, ソフトマージンパラメタ C を個別に設定する. 具体的には, 構造化アブストラクトに対しては $C = C_s$, 非構造化アブストラクトに対しては $C = C_u$ である. 各二値分類問題に対して, 最適な (C_s, C_u) を求め, 5.2 節のベースラインと比較する.

そのために $C_u = 10^n$ の範囲として $n = -10, -9, \ldots, 0, \ldots, 9, 10$, および $C_s = 10^m$ の範囲として n を基準として $m/n = 0, -1, \ldots, -10$ について調査した.

C_u と C_s の範囲が同一でないのは, 非構造化アブストラクトの文の制約をより重視するためである (つまり $C_u > C_s$; パラメタ C_u, C_s は値が大きい程制約を破った際の罰則が大きくなることに注意). 以上のように, (C_s, C_u) の組を網羅的に探索し, もっとも高い精度を示した組を以下の実験に用いた.

6. 実験結果

二値分類の結果を, 表 4 に示す. 10 通りの二値分類問題のうち三つにおいて, ベースラインに対する精度の向上が見られた.

さらに, 最終的な 5 クラス多値分類についても検証を行った. 多値分類は, ベースライン同様, 10 個の二値分類副問題それぞれにおいて求めた最適なパラメタ (提案手法の場合 (C_u, C_s) の

表 3　ベースラインにおいて, 全訓練データを使用した ($C_u = C_s$) 精度が非構造化アブストラクトのみを用いて訓練した場合 ($C_s = 0$) を上回ったケース.

クラス 1	クラス 2	ベースライン (1) 全訓練データ使用 ($C_s = C_u$)	ベースライン (2) 非構造化アブストラクトのみ訓練に使用 ($C_s = 0$)
BACKGROUND	RESULTS	85.5	84.3
OBJECTIVE	RESULTS	95.8	94.5
METHOD	CONCLUSION	75.7	72.7

表 4　二値分類精度 (%)

クラス 1	クラス 2	ベースライン (1) 全データを使用して訓練した SVM ($C_s = C_u$)	ベースライン (2) 非構造化アブストラクトのみを訓練に用いた SVM ($C_s = 0$)	提案手法 データソースに依存した (C_s, C_u)
BACKGROUND	OBJECTIVE	80.7	**93.7**	**93.7**
BACKGROUND	METHOD	80.7	**88.4**	**88.4**
BACKGROUND	RESULTS	85.5	84.3	**86.7**
BACKGROUND	CONCLUSION	81.0	**83.7**	**83.7**
OBJECTIVE	METHOD	76.9	84.3	**88.4**
OBJECTIVE	RESULTS	95.8	94.5	**97.2**
OBJECTIVE	CONCLUSION	77.7	**85.5**	**85.5**
METHOD	RESULTS	86.7	**89.8**	**89.8**
METHOD	CONCLUSION	**75.7**	72.7	**75.7**
RESULTS	CONCLUSION	**85.5**	**85.5**	**85.5**

組) による分類器の結果を用いて, 多数決によって行った. 多値分類の結果を図 3 に示す. 図からわかるように, (C_u, C_s) を個別に設定した場合が, ベースラインの $C_u = C_s$ および $C_s = 0$ の双方 (いずれも各二値分類問題に対して最適な C を用いたが場合) を上回る結果となった.

7. 結論と今後の課題

以上, データソースの違いを考慮した学習の, MEDLINE アブストラクトを用いた構造役割分類に対する有効性を示した.

今回は, SVM のソフトマージンパラメタをデータソースに依存して別個に設定する有効性を検証することを主眼としたため, 単純な素性のみを使用した. 修辞構造役割推定のさらなる精度向上を目指して, 今後は以下のような拡張が考えられる.

（1） 使用する素性の拡張. 今回の実験では, 単語や品詞といった単純な素性のみを用いたが, Mizuta ら [12] や Yamasaki ら [17] は時制と構造役割の間に強い相関があることを指摘している. より複雑なこれらの素性を取り込んだ上で, 提案手法の有効性の検証が必要である.

（2） 系列ラベル付け手法の導入. Yamasaki らは, 文の構造役割の系列とアブストラクト内の出現位置が重要な素性であると指摘している. アブストラクト先頭には BACKGROUND あるいは OBJECTIVE が出現しやすく, METHOD の後には RESULTS が出現しやすい, といった情報をとらえるために三通りの方策が考えられる.

（a） 文のアブストラクト内の相対位置を表す素性を SVM のような系列タグ付けに特化していない分類器に 対して導入する.

（b） Mizuta らの手法に基づく重みつけ関数を定義し, 事前知識として SVM に取り込む [21].

（c） 条件付き確率場 [9] や Collins の系列タグ付け用パーセプトロン [5] といった系列タグ付けアルゴリズムに, データソースに依存した制約重み付けを取り込んで拡張する.

文　献

[1] Naoki Abe, Bianca Zadrozny, and John Langford. An iterative method for multi-class cost-sensitive learning. In *Proceedings of ACM KDD'04*, 2004.

[2] Ad Hoc Working Group for Critical Appraisal of Medical Literature. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine*, 106(4):598–604, 1987.

[3] Ulf Brefeld, Peter Geibel, and Fritz Wysotzki. Support Vector Machines with example dependent costs. In *Proceedings of the European Conference on Machine Learning (ECML 2003)*, volume 2810, pages 167–178, 2003.

[4] Chih-Chung Chang and Chih-Jen Lin. A library for support vector machines. `https://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] Michael Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002.

[6] GENIA Tagger. `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger`. Department of Information Science, Faculty of Science, University of Tokyo.

[7] N. Graetz. Teaching EFL students to extract structural information from abstracts. *Reading for Professional Purposes: Methods and Materials in Teaching Languages*, pages 123–135, 1985.

[8] McKnight L. and Srinivasan P. Categorization of sentence types in medical abstracts. *Proceedings of the 2003 AMIA conference*, 2003.

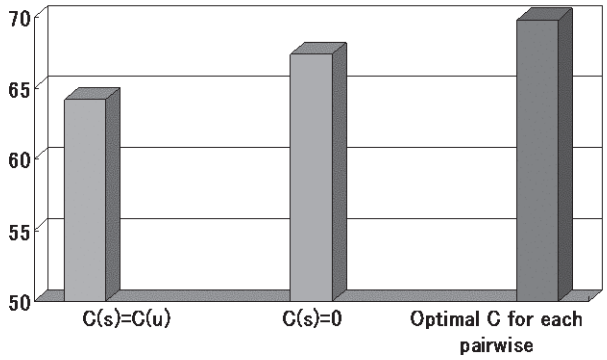

図 3　5 クラス分類精度 (%)

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, pages 282–289. Morgan Kaufmann, 2001.

[10] Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. Generative content models for structural analysis of medical abstracts. *Proceedings of the HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing*, pages 65–72, 2006.

[11] MEDLINE. `http://nlm.nih.gov/databases/databases_medline.html`, 2003. U.S. National Library of Medicine.

[12] Y. Mizuta and N. Collier. Zone identification in biological articles as a basis for information extraction. *Proceedings of the the Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004.

[13] Constantin Orasan. Patterns in scientific abstracts. *Proceedings of Corpus Linguistics*, pages 433–443, 2001.

[14] PubMed. `http://www.ncbi.nlm.nih.gov/PubMed/`. U.S. National Library of Medicine.

[15] F. Salanger-Meyer. Discoursal flows in Medical English abstracts: A genre analysis analysis per research-and text-type. *Text*, pages 365–384, 1990.

[16] Bernhard Schölkopf and Alex J. Smola. *Learning with Kernels*. MIT Press, 2002.

[17] Masashi Shimbo, Takahiro Yamasaki, and Yuji Matsumoto. Sentence role identification in medline abstracts: Training classifier with structured abstracts. *Lecture Notes in Computer Science*, pages 236–254, 2005.

[18] I. Tbahriti, C. Chichester, F. Lisacek, and P. Ruch. Using argumentation to retrieve articles with similar citations from MEDLINE. *Proceedings of the the Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004.

[19] S. Tufel and M. Moens. Experiments with relevence and rhetorical status. *Computational Linguistics*, 28(4), 2002.

[20] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[21] Xiaoyun Wu and Rohini K. Srihari. Incorporating prior knowledge with weighted margin support vector machines. *Proceedings of ACM KDD'04*, pages 326–333, 2004.