半教師あり系列ラベリングによるアブストラクトのセクション分割

平尾 努　　鈴木 潤　　磯崎 秀樹　　永田 昌明
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
{hirao,jun,isozaki}@cslab.kecl.ntt.co.jp,
nagata.masaaki@lab.ntt.co.jp

1　はじめに

多くの科学技術論文アブストラクトは，論文の目的，実験手法，実験結果，結論といったセクション構造，一種の修辞構造を持っている．こうした構造の解析は，検索や情報抽出，ブラウジングなどアプリケーションの基盤として重要である．

従来手法の多くは，これを解析するため，多値分類のアプローチで取り組んできた．すなわち，文がどのセクションの属するかを分類する問題として扱ってきた [7, 2, 10]．しかし，各セクションは独立に存在するのではなく，それらの間には遷移の構造があると考えることが適切であろう．よって，多値分類問題として考えるよりも系列ラベリング問題として考えた方がより自然である．実際，Hirohata らは条件付き確率場 (Conditional Random Fields: CRFs)[3] を用いることで，非常に高い分類精度を得ている [1]．

本稿では，より高精度な分類 (分割) を達成するため，半教師あり CRFs を用いたアブストラクトのセクション分割手法を提案する．ただし，このタスクにおいては，全ての N グラムを素性として利用すると素性数が爆発するという問題があり，半教師あり学習の枠組みを適用する際の大きな問題となる．そこで，ラベルありデータだけでなくラベルなしデータからもラベリングに有効な素性を効率的に抽出する手法を提案する．

2　半教師あり系列ラベリング

本稿では，文献 [8] にて提案された半教師あり CRFs を用いる．

$\mathcal{X}$, $\mathcal{Y}$ をそれぞれ可能な入出力の集合とする．$\boldsymbol{x} \in \mathcal{X}$ を入力サンプル，$\boldsymbol{y} \in \mathcal{Y}$ を出力ラベルとし，$\mathcal{C}$ を無向グラフ中のクリーク集合とする．$\boldsymbol{y}_c$ はクリーク $c \in \mathcal{C}$ から出力されるラベルである．ここで，c に対し，ポテンシャル関数 Ψ を導入すると，CRFs は以下の式で定義される．

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x})} \prod_{c \in \mathcal{C}} \Psi_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}) \qquad (1)$$

ただし，$\boldsymbol{\lambda}$ はパラメタベクトルであり，$Z(\boldsymbol{x}) = \sum_y \prod_{c \in \mathcal{C}} \Psi_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda})$ である．c に対する特徴ベクトルを $\boldsymbol{f}_c(\boldsymbol{y}_c, \boldsymbol{x})$ とあらわすと $\Psi_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}) = \exp(\boldsymbol{\lambda} \cdot \boldsymbol{f}_c(\boldsymbol{y}_c, \boldsymbol{x}))$ となる．

ここで，J 種の確率モデル (ここでは隠れマルコフモデル) を $p_j(\boldsymbol{x}_j, \boldsymbol{y}; \boldsymbol{\theta}_j)$ を導入すると，ポテンシャル関数は以下の式で再定義される．

$$\begin{aligned}\Psi'_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}'\Theta) &= \exp(\boldsymbol{\lambda} \cdot \boldsymbol{f}_c(\boldsymbol{y}_c, \boldsymbol{x})) \cdot \prod_j p_j(\boldsymbol{x}_{jc}, \boldsymbol{y}_c; \boldsymbol{\theta}_j)^{\lambda_{I+j}} \\ &= \exp(\boldsymbol{\lambda'} \cdot \boldsymbol{h}_c(\boldsymbol{y}_c, \boldsymbol{x}))\end{aligned}$$

ただし，$\boldsymbol{h}$ は，特徴ベクトル $\boldsymbol{f}$ と p_j の対数尤度をつなげたベクトルであり，それに対応するパラメタベクトルを $\boldsymbol{\lambda'}$ とする．また，$\boldsymbol{\Theta} = \{\theta_{j=1}^J\}$ である．これらを用いて，半教師あり CRFs は以下の式で定義される．

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\lambda'},\boldsymbol{\Theta}) = \frac{1}{Z'(\boldsymbol{x})} \prod_{c \in C} \Psi'_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) \qquad (2)$$

ラベルありデータ $\mathcal{D}_\ell = \{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$ が与えられた場合，$\boldsymbol{\Theta}$ 固定のもと，$\boldsymbol{\lambda}'$ の MAP 推定は以下の式で与えられる．

$$\mathcal{L}^1(\boldsymbol{\lambda}'|\boldsymbol{\Theta}) = \sum_n \log P(\boldsymbol{y}^n|\boldsymbol{x}^n;\boldsymbol{\lambda}',\boldsymbol{\Theta}) + \log p(\boldsymbol{\lambda}') \qquad (3)$$

ただし，$p(\boldsymbol{\lambda'})$ は $\boldsymbol{\lambda'}$ の事前確率分布でありる．

また，ラベルなしデータ $\mathcal{D}_u = \{\boldsymbol{x}^m\}_{m=1}^M$ が与えられた場合，Maximum Discriminant Functions sum' 推定法を用いると，$\boldsymbol{\lambda'}$ を固定のもとでは，以下の目的関数を最大化すればよい．

$$\mathcal{L}^2(\boldsymbol{\Theta}|\boldsymbol{\lambda}') = \sum_m \log \sum_{\boldsymbol{y} \in \mathcal{Y}} g(\boldsymbol{x}^m, \boldsymbol{y}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) + \log p(\boldsymbol{\Theta}) \qquad (4)$$

$p(\boldsymbol{\Theta})$ は $\boldsymbol{\Theta}$ の事前確率分布をあらわし，$g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda}', \boldsymbol{\Theta}) = \prod_{c \in \mathcal{C}} \Phi'_c(\boldsymbol{y}_c, \boldsymbol{x}; \boldsymbol{\lambda}', \boldsymbol{\Theta})$ である．

学習時には $\mathcal{L}^1(\boldsymbol{\lambda'}|\boldsymbol{\Theta})$ と $\mathcal{L}^2(\boldsymbol{\Theta}|\boldsymbol{\lambda'})$ を反復して最大化し，最適化を行う．詳しくは文献 [8] を参照されたい．

3　系列パタンマイニングによる効率的な素性抽出

アブストラクトのセクション分割は，一般的なテキスト分類とは異なり，1 文に対しクラスラベルを割り当てるため，ユニグラムだけでなく，バイグラム，トライグラムといった長い N グラムも素性として必要となる．しかし，コーパス中のすべての N グラムを列挙し，素性として利用することは素性抽出，その後の学習の計算量，精度の観点から現実的ではない．

本稿ではこうした問題点を解決するため，系列パタンマイニング手法として知られる PrefixSpan[5] を拡張し，効率的にラベルありデータ，ラベルなしデータからラベリングに有効だと考えられ N グラムのみを抽出する手法を提案する．提案手法の基本的な考えは，文献 [9, 4, 6] に基づく．

表 1　分割表

	L	$\bar{L}$	計
α	$O_{\alpha L} = Y(\alpha)$	$O_{\alpha \bar{L}}$	$O_\alpha = X(\alpha)$
$\bar{\alpha}$	$O_{\bar{\alpha} L}$	$O_{\bar{\alpha} \bar{L}}$	$O_{\bar{\alpha}}$
	$O_L = M$	$O_{\bar{L}} = N - M$	N

```
Procedure WTPS(α, R, K)
  if R = {}  return
  R' ← {}
  if α = [ ]
    R' ← R
    β ← itemset(R)
  else
    foreach d ∈ R
      subseq ← postseq(last(α), d)
      if subseq ≠ [ ]
        R' ← append(R', subseq)
    end
    β ← itemset(R')
  foreach b ∈ β
    α ← append(α, [b])
    if |A| ≤ K
      A ← append(A, [α, χ²(α)])
      if |A| = K
        A ← sort(A)
        τ_K ← χ²(A[K])
    if χ²(α) ≥ τ_K
      A ← lastdel(A)
      A ← append(A, [α, χ²(α)])
      A ← sort(A)
      τ_K ← χ²(A[K])
      call WTPS(α, R', K)
    elsif χ²_max(α) ≥ τ_K
      call WTPS(α, R', K)
  end
```

図 1　効率的な単語列抽出アリゴリズム

3.1　ラベルありデータからの素性抽出

PrefixSpan は与えられたデータベース (コーパス) から頻度が ξ 以上の系列パタン (単語列，本稿では N グラム) を抽出する手法である．具体的には単語をノードとし，それに接続可能な単語[*1]を子ノードとするトライを構築し深さ優先探索を行う．ただし，ルートから子ノードまでの単語列の頻度が ξ 未満であれば，単語を子ノードとして追加せず，それ以上の探索を行わない．ξ 以上であれば，単語を子ノードとして追加し，先の手続きを再帰的に繰り返す．これは，ノードを下方へと辿っていくと単語列の頻度が増加することがないという特性を活かした探索手法といえる．

しかし，この手法では，頻出する単語列しか抽出することができない，つまり，クラスラベルを考慮した単語列を抽出できないという問題がある．

[*1] 着目している単語の右隣に出現する単語の集合．より一般的には，その単語よりも右側に出現する単語の集合．

情報量基準による探索の打ち切り

単語列を α，クラスラベルを L とし，α と L との間の関係を表 1 に示す分割表で考える．このとき，単語列とラベルとの間の独立 (依存) 性を表す指標として $\chi^2(\alpha)$ 値が以下の式で定義される．

$$\chi^2(\alpha) = \frac{N(O_{\alpha L} \cdot O_{\bar{\alpha}\bar{L}} - O_{\bar{\alpha} L} \cdot O_{\alpha \bar{L}})^2}{O_\alpha \cdot O_{\bar{\alpha}} \cdot O_L \cdot O_{\bar{L}}}$$

χ^2 値の高い単語列ほど L との依存度が高い単語列だと考えることができるので，ラベル分類に有効な素性といえる．表中，N はコーパス中の文の総数，$O_L = M$ はあるクラスラベル L に属する文の総数であり，これらは，コーパスが与えられた時に決まる定数である．$O_\alpha = X(\alpha)$ は単語列 α を含む文の数，$O_{\alpha L} = Y(\alpha)$ は単語列 α を含み，かつ，クラスラベルが L である文の数である．

ここで，PrefixSpan における ξ の代わりに χ^2 値を用いることができれば，効率的にラベル分類に有効な単語列を抽出することができる．すなわち，ある閾値 τ を考え，$\tau \leq \chi(\alpha)$ を満たすように深さ優先探索を打ち切れば良い．しかし，頻度とは異なり，トライを下方へと辿っていく際，単語列の χ^2 値は単調に減少しない．つまり，あるノードにおける χ^2 値がそれより上位ノードの χ^2 値を超える場合がある．ここで，式 (5) が，$X(\alpha)$，$Y(\alpha)$ という依存関係にある 2 つの変数からなる関数となることに注意する．つまり，$\chi^2(\alpha) = \chi^2(X(\alpha), Y(\alpha))$ である．この時，ある単語列 α の接尾に対して何らかの単語 (列) を追加した単語列の取り得る χ^2 値の最大値は，以下の式で定義される [4].

$$\chi^2_{\max}(\alpha) = \max(\chi^2(Y(\alpha), Y(\alpha)), \chi^2(X(\alpha) - Y(\alpha), 0))$$

よって，以下の場合を考えて単語列の抽出と探索の打ち切りを考えれば良い．

(1)　$\tau \leq \chi^2(\alpha)$
(2)　$\tau > \chi^2(\alpha)$,　$\tau \leq \chi^2_{\max}(\alpha)$
(3)　$\tau > \chi^2(\alpha)$,　$\tau > \chi^2_{\max}(\alpha)$

(1) の場合，単語列 α の χ^2 値が τ を超えているため，α を抽出し，それ以下のノードの探索を行う．(2) の場合，単語列 α の χ^2 値が τ 未満であるため，α を抽出はしないが，α の接尾に何らかの単語 (列) を追加した場合，その χ^2 値が τ を超える可能性があるため，それ以下のノードの探索を行う．(3) の場合，単語列 α の χ^2 値もその接尾に何らかの単語 (列) を追加した場合の χ^2 値も τ は超えないので探索は行わない．

K ベスト単語列の効率的な抽出

上述した手続きをとることで，ラベル L に依存する単語列を効率的に抽出することはできるが，τ を適切に設定しなければ，探索が効率的にならないという問題が残る．素性抽出の観点からは，χ^2 値が高い上位 K 個の単語列のみを取り出せば良いので，本稿では，文献 [6] の考えに基づき，τ を動的に決定し，より効率的な探索を行う．

単語列を格納するための長さ K の配列 $A[1, 2, \ldots, K]$ を用意し，トライを探索した順に単語列を A に格納していく．この時，配列内の単語列はその χ^2 値でソートされているものとする．ここで，配列の K 番目の単語列の χ^2 値

表 2　実験結果

追加した単語列数	Baseline	3 千	1 万	2 万	3 万	4 万	5 万	6 万	7 万	8 万	9 万	10 万
正解率	84.1	83.6	84.5	**85.1**	84.4	84.8	84.2	84.1	84.0	84.3	84.5	83.5
アブストラクト正解率	26.5	27.0	28.5	**32.5**	30.0	28.5	28.5	28.0	28.5	29.0	29.0	27.0
F 値 (O)	58.0	61.0	60.0	62.0	**63.0**	62.5	59.0	62.5	62.0	62.0	61.0	61.5
F 値 (M)	34.4	35.7	37.2	**39.8**	36.9	36.9	36.1	35.3	35.0	35.4	35.8	34.3
F 値 (R)	38.7	38.2	41.2	**45.3**	42.4	41.9	40.1	40.9	43.2	43.0	44.3	41.0
F 値 (C)	75.5	74.9	75.9	**78.6**	77.3	77.3	76.7	76.0	76.7	78.0	76.8	74.9
F 値 (平均)	51.7	52.5	53.6	**56.4**	54.9	54.7	53.1	53.7	54.2	54.6	54.5	53.0

を閾値 τ_K とし，前節 (1)〜(3) の条件を考える．(1) の場合には，$A[K]$ を削除，α を A に追加し，τ_K を更新する．さらに，A の要素をその χ^2 値でソートし，探索を続ける．(2) の場合は，探索のみを続け，(3) の場合には探索を打ち切る．図 1 に疑似コードを示す．R，R' はコーパス (文集合)，$itemest$ は，コーパス中の全てのユニグラムを抽出する関数である．$postseq(last(\alpha), d)$ は，文を先頭から単語列 α の接尾となる単語が出現した位置までを削除する関数である．$append()$ は第 1 引数に第 2 引数を追加する関数である．

3.2　ラベルなしデータからの素性抽出

前節までで，ラベルありデータから，クラスラベル L に依存する単語列を効率的に抽出する手法を説明した．しかし，ラベルなしデータでは，当然ながら，L が与えられないため，先の手法をそのまま適用し，単語列の χ^2 値を計算することができない．本稿では，この問題を回避するため，ラベルありデータから分類器を構築し，データのクラスラベルへの帰属確率を求め，それを重みとしてデータの頻度を計算することでラベルなしデータからも効率的にクラスラベルに依存する単語列を抽出する手法提案する．すなわち，分割表における $X(\alpha)$，$Y(\alpha)$ を以下の式で定義する．

$$X(\alpha) = \sum_{d \in \mathcal{D}^u} F(d, \alpha)$$
$$Y(\alpha) = \sum_{d \in \mathcal{D}^u} P(C_d = L) \cdot F(d, \alpha)$$

$P(C_d = L)$ は文 d が L に属する確率であり，$\mathcal{D}^\ell$ を用いた分類器 (たとえば，ナイーブベイズ，最大エントロピー法など) で d をデコードすれば求めることができる．$F(d, \alpha)$ は d は単語列 α を含む時に 1，それ以外は 0 をとる関数である．これらの関係を用いると，M，N は以下の式となる．

$$M = \sum_{d \in \mathcal{D}^u} P(C_d = L)$$
$$N = |\mathcal{D}^u|$$

これらの値を用いることでラベルなしデータであっても，単語列の χ^2 値を計算することができるので，効率的に L に依存する K ベスト単語列を抽出することができる．

4　評価実験

4.1　コーパス

PubMed より，8 文以上からなる論文アブストラクトを 1,000 件抽出し，それらを人手で，目的 (Objective:O)，手法 (Method:M)，結果 (Results:R)，結論 (Conclusions:C) という 4 つ領域 (クラス) に分割した．これより，ラベルありデータとして 600 件，開発，テストデータとしてそれぞれ 200 件を得た．ラベルなしデータは，上記データと重ならぬよう PubMed から 6 万件のアブストラクトを選んだ．

4.2　評価指標と素性

評価指標としては，多値分類問題 (本稿では 4 クラスの分類問題) として正解率，1 つのアブストラクト全体での文のクラス分類が全て正解した割合であるアブストラクト正解率を用いた．さらに，本研究では，アブストラクトの領域分割を系列ラベリング問題として扱ったので，それらを正しく分割できた場合の精度，チャンキング F 値も用いた．

素性は，文献 [1] に従い，単語列，すなわち N グラム (範囲は前後 2 文) と文の出現位置を用いた．まず，ラベルありデータを用いて教師あり CRFs を評価した．3.1 節で説明した手法を用い，OMRC のそれぞれのクラスラベルに対し，χ^2 値の高い上位 K 件の単語列を抽出し，それらの和集合を素性として用いた[*2]．なお，式 (2)，式 (3) にそれぞれ，Gaussian priors，Dirichlet priors を導入し，K を変化させ，開発セットで最適化を行った．その結果，最も成績がよかった $K = 10,000$ を採用した．以降，これをベースラインとした．

次に，ラベルありデータを用いて最大エントロピー法[*3]により 4 クラスの多値分類器を構築する．分類器を用いて，ラベルなしデータ中の各文に対し，OMRC の各クラスラベルに属する確率を求める．これを用い，ラベルありデータの場合と同様に，3.2 節で説明した手法を用い，ラベルなしデータから単語列を抽出した．既にラベルありデータから抽出した単語列とラベルなしデータから抽出した単語列の和集合を最終的な素性として用いた．なお，K は，3 千件と 1 万件以降は，1 万きざみで 10 万件まで変化させた．

[*2] 通常，各ラベルに関して抽出した単語列の間には重なりがあるので最終的な素性の数は $4 \times K$ よりも少ない．

[*3] CRFs ではアブストラクト中の各文が OMRC のどのラベルの属するかという確率を与えることができないため，最大エントロピー手法を用いた．

表 3　特徴的な単語列

	OBJECTIVE	METHOD	RESULTS	CONCLUSIONS
LAB	The aim of	randomized	P <	results suggest that
	Our objective(s)	To test	+/−	demonstrate that
	This study	we designed	The results showed	results contribute
	The purpose	The method	significanty	These results indicates that
	This article	microarray	than	study confirmed
ULAB	we analyzed	compared with	(95 % confidence	According to
	prospectively	procedure	observation	To address this
	first report	detection is	probabilistic	evidence suggests that
	we determine	KO mice	mean values	analysis revealed that
	focus on	zebrafish	$n =$	was detected in

4.3　結果と考察

実験結果を表 2 に示す．まず，正解率について議論する．ベースライン (教師あり CRFs を用い，10,000 件の単語列を利用) に対し，半教師あり学習の枠組みを用いることで精度の向上は見られる，最も良い場合 ($K = 20,000$) であってもその差は 1 ポイント程度にとどまっている．また，ベースラインよりも精度が劣化している場合もみられる．

一方，アブストラクト正解率では，概ね精度が向上しており，K を 2 万に設定した場合では，およそ 6 ポイントの向上があった．それ以外の場合でもベースラインを下回ることはなく半教師あり系列ラベリングと提案した単語列抽出手法の有効性がわかる．ただし，抽出する単語列は 2 万件の場合が最も良く，それ以上増やしていくと精度が劣化する傾向にある．これは，素性数を増やすことで過学習が起っていると考えられる．

チャンキング F 値でもアブストラクト正解率と同様，単語列の数を 2 万件と設定した場合の成績がよく，F 値平均ではベースラインと比較して約 5 ポイント程度向上している．クラスラベルごとにみていくと，手法 (Method) と結果 (Results) では大きく F 値が向上しており，提案手法の有効性がよくわかる．

表 3 にラベルありデータから抽出した単語列，ラベルなしデータから抽出した単語列の例を示す．各クラスともそれを特徴づける単語列が抽出できていることがよくわかる．

5　まとめ

本稿では，半教師あり学習による論文アブストラクトのセクション分割手法を提案した．この際，必須となる N グラム素性の爆発を防ぐため，ラベルありデータ，ラベルなしデータから効率的にラベル分類に有効な単語列を抽出する手法を提案した．多値分類としての分類精度は約 85% を達成し，チャンキング問題としての F 値は約 56% を達成した．従来の教師あり学習による系列ラベリング手法と比較してチャンキング F 値では，約 5 ポイント近い改善が得られ，提案手法の有効性を確認した．

参考文献

[1] K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. Identifying sections in scientific abstracts using conditional random fields. In *Proc. of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, pages 381–388, 2008.

[2] T. Itoh, M. Shimbo, T. Yamasaki, and T. Matsumoto. Semi-supervised sentence classification for medline documents. In *IPSJ SIG Technical Report 2004-ISC-138*, pages 141–146, 2004.

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the 18th ICML*, pages 282–289, 2001.

[4] S. Morishita and J. Sese. Traversing itemset lattices with statistical metric pruning. In *Proc. of ACM SICACT-SIGMOD-SIGART Symp. on Database Systems (PODS'00)*, pages 226–236, 2000.

[5] J. Pei, J. Han, B. Mortazavi-Asl, and H. Pinto. Prefixspan: Mining sequential patterns efficiently by prefx-projected pattern growth. In *Proc. of the 17th International Conference on Data Engineering (ICDE 2001)*, pages 215–224, 2001.

[6] J. Sese and S. Morishita. Answering the most correlated n association rules efficiently. In *Proc. of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 410–422, 2002.

[7] M. Shimbo, T. Yamasaki, and T. Matsumoto. Using sectioning information for text retrieval: A case study with the medline abstracts. In *Proc. of the Second International Workshop on Active Mining*, pages 32–41, 2003.

[8] J. Suzuki and H. Isozaki. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proc. of the ACL-08:HLT*, pages 665–673, 2008.

[9] J. Suzuki, H. Isozaki, and E. Maeda. Convolution kernels with feature selection for natural language processing tasks. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistic (ACL*, pages 119–126, 2004.

[10] Y. Yamamoto and T. Takagi. A sentence classification system for multi-document summarization in the biomedical domain. In *Proc. of the International Workshop on Biomedical Data Engineering*, pages 90–95, 2005.