社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－182 (10)
2007／11／20

部分的かつ曖昧なラベル付き構造データからのマルコフ条件付確率場の学習

坪井 祐太 †1,3　鹿島 久嗣 †1　森 信介 †2
小田 裕樹　松本 裕治 †3

†1 日本アイ・ビー・エム株式会社　†2 京都大学　†3 奈良先端科学技術大学院大学

コーパスのアノテーション作業では、重要な部分のみへのラベル付与が効率的であり，また一部のラベルが一つに決定できずラベルの曖昧性が残ってしまうことが発生する．本研究では，このような部分的かつ曖昧なラベルが付与された学習データを使用してマルコフ条件付確率場を学習する方法を提案する．日本語単語境界認識の分野適応における部分的アノテーションと英語品詞タグ付けデータに存在する曖昧なアノテーションに対して実験を行い，本手法の有効性を検証した．

キーワード：部分的かつ曖昧なアノテーション，構造出力問題，条件付確率場

Training Conditional Random Fields using Partial and Ambiguous Structured Labels

Yuta Tsuboi †1,3, Hisashi Kashima †1, Shinsuke Mori †2,
Hiroki Oda and Yuji Matsumoto †3

†1 IBM Japan, Ltd.　†2 Kyoto University　†3 Nara Institute of Science and Technology

We address corpus building situations which we only annotate important part of given data, or which we cannot resolve label ambiguities with reference to the linguistic context. We proposed a parameter estimation method for Conditional Random Fields (CRFs) using the partial and ambiguous annotations of structured data. We show the promising results of our method applied to the domain adaptation task of Japanese word segmentation and Part of Speech tagging task using the ambiguous tags in Penn Treebank Corpus.

Keywords: Partial and Ambiguous Annotations, Structured Output Learning, Conditional Random Fields

1. はじめに

近年，構造を出力とする問題を扱う識別モデルが発展し，多くの自然言語処理の問題に応用されるようになった．構造を出力する問題とは，例えば日本語の単語分割タスクは，２つの文字の境界を入力単位とし，文の各文字境界列に対応する単語境界・非境界ラベル列を出力する問題として扱うことができる．また，品詞タグ付与問題は文を構成する単語列を入力列とし，各単語に対応する品詞のタグ列を出力する問題として考えることが出来き，係り受け解析は入力文の単語列に対応する依存木構造を出力する問題と考えることができる．

構造出力を扱う識別モデルは条件付確率場（Conditional Random Fields:CRF)[9] に始まり，パーセプトロン[5]，マージン最大化[20] アルゴリズムなどが提案されている．これまでの構造の各要素を独立に学習する手法との大きな違いは，構造全体の整合性を考慮したパラメータ推定である．構造内の各要素に相互関係がある問題で有効な手法であるため，品詞タグ付け[9]，単語分割問題[8),12)]，構文解析[21)]，固有表現抽出[16)]，対訳語対応付け[22)] など，自然言語処理の様々な問題に応用されている．

識別モデルの教師付き学習では，入力と出力の対を学習データとしてモデルのパラメータを推定する．構造出力でも出力が構造を持っている点を除けば同じ枠組みで学習は定義される．しかし，自然言語処理の応用では入力と出力構造の対が完全に与えられず，出力構造の一部のみが与えられたり，出力構造の一部が２つ以上の候補を持ち曖昧性を持ったラベルで与えられることも想定される．

出力構造の一部のみが与えられる例としては，既存の自然言語処理システムの分野適応がある．適応したいテキストの文中でも，特に重要な部分にアノテーションをして学習データを作成する方が文全体にアノテーションをするより効率的である．2.1 節では統計

的日本語単語分割の分野適応時の作業コスト削減を実現する，部分的なアノテーション付与方法を紹介する．

また，タスク自身の曖昧性やアノテーション作業者の習熟度によって曖昧性を持ったラベルが構造の一部に付与されることがある．2.2節では，よく知られたPenn Treebank英語品詞タグ付きコーパスにおいて品詞タグが複数候補付与されている単語があることを例に曖昧性のあるアノテーションを説明する．

これらの不完全なアノテーションは，これまでの構造出力学習手法では想定されていなかった．そこで，3節で問題を定式化し，4節で部分的かつ曖昧なアノテーションを用いてCRFを学習する方法を提案する．提案手法は動的計画法を用いることでマルコフ性を仮定できる問題においては多項式時間で計算できる．

提案手法の有効性を確認するために，5節で提案手法を日本語単語分割の分野適応と品詞タグ付けに適用した実験結果を示す．5.1節では，選択的サンプリング法によって重要な単語のみをアノテーションすることで文全体をアノテーションするのに比べて1/10の作業量で大きな学習効果が得られたことを報告する．5.2節では，品詞タグ付けタスクにおいて曖昧なアノテーションのまま学習する事によって安定的な性能が得られることを示す．

6節で関連研究を紹介し，最後に7節では結論と今後の展望について述べる．

2. 部分的かつ曖昧なアノテーション

2.1 部分的なアノテーション

本節では，部分的なアノテーションを付与することで分野適応において効率的なアノテーション作業が行える例を示す．本稿での自然言語処理システムの分野適応の目的は，特定の（適応元）分野の学習データ、または既存の言語処理システムを活用し、処理対象の（適応先）データでの性能を向上させることである．森[11]の日本語単語分割のためのユーザーインターフェース（UI）を例として紹介する．

日本語や中国語のように分かち書きされていない言語では，単語境界を求めることは単純ではなく，周辺文脈を考慮した上で適切に決定する必要がある複雑な問題であるため，統計モデルが活用されている．しかし，統計的手法では言葉の使われ方の違いによって，しばしば学習データと異なる分野のデータで性能劣化が発生する．単語分割では特に元の分野では観測されなかった未知語の出現が分割誤りの主原因となる．

一方で，現実の適応時には適応先分野の専門用語辞書や製品名一覧が分野特有の単語リストとして使用可能な事が多い．森[11]は適応先の分野単語リストが出現する文脈を評価するKWIC（KeyWord in Context）形式のUIを提案した．分野単語リストのエントリ「すり傷」のUIでの表示例を**図1**に示す．アノテーション作業者は文字列「すり傷」が適応先テキストの各文脈で実際に単語として使用されている箇所に印を付ける．1行目は2単語「こする」と「傷つく」の一部，2行目は単語「こすり傷」の一部であるため，最後の2行にのみ正しい単語境界であることを示す印を付けている．このUIは文字列領域に対するアノテーションをYES/NO入力に単純化することで作業を効率化しており，他の自然言語処理タスクにおいても有用な方法である．例えば，品詞タグ付けは名詞か否か，係り受け付与は単語間の構文的依存関係の有無を文脈を見ながら二値判断する作業へと分解することが可能で，このUIを適用できる．このUIは重要な部分だけをアノテーションすることで文全体へのアノテーションに比べて作業量削減効果があるだけでなく，部分的なアノテーションを許すことで作業者が自信の無い部分は無理に判断する必要がないためノイズとなるアノテーションの追加を防止する効果も期待できる．一般的な分野適応時には，アノテーション作業は分野知識はあるが言語学的知識の無い複数人で行うことが多いため，この特徴は非常に重要である．

	が皮膚を強くこ	すり傷	ついてしまっ
	感染，角膜のこ	すり傷	，角膜潰瘍，
○	皮膚に切り傷や	すり傷	を負った場合
○	泥まみれの深い	すり傷	や，皮下深く

図1 KWIC形式アノテーションUI

また，分野適応においては既存のシステムが存在することが期待できるので，事例に対して既存のシステムで何らかの優先度を付与して部分アノテーションを付与することもできる．選択的サンプリングや能動学習の研究によると，優先度の高い事例に対して集中的に正解を付与して学習データを作成することで少ないアノテーション量で効率的に性能向上を得られることが知られている．構造出力タスクでは，各事例の構造内のアノテーションすべき要素を選択することになる[1),3),14),17)]．上記のUIの場合には，文脈を優先順位付けしてユーザに表示することで確認する文脈数を減らすことが可能になる．

上記の方法で，文全体の単語境界を指定する作業に比べ，部分的にアノテーションされた適応先の文は比較的低コストで作成することが出来る．単語分割以外の分野適応においても，構造全体のアノテーションに

表 1　Penn Treebank コーパスの頻度 3 以上の曖昧な品詞タグ例.

頻度	単語	品詞
15	data	NN|NNS
10	more	JJR|RBR
7	pending	JJ|VBG
4	than	IN|RB
3	trading	NN|VBG
3	broker-dealer	JJ|NN

比べ部分的アノテーションは効率的に作成することが出来ると考える.

2.2　曖昧なアノテーション

本稿で扱う曖昧なアノテーションとは，ラベル列などの構造を持った学習データのある事例の要素に付与されるべきラベルの候補が複数存在する場合に，その候補ラベル集合を指す.

例えば，Penn Treebank の品詞タグ付きデータでは，次のような曖昧なアノテーションが存在する：
That/DT suit/NN is/VBZ pending/VBG|JJ ./.
ただし，各単語のスラッシュ後の英字は品詞タグでそれぞれ DT：限定詞，NN：名詞単数，VBZ：動詞 3 人称単数現在形，VBG：動名詞または動詞現在分詞，JJ：形容詞を示す．縦棒はタグの曖昧性を示し，この例では “pending” の品詞タグは “VBG または JJ” としてアノテーションされていることを意味する☆.

曖昧なアノテーションは，タスクおよびアノテーション手順の曖昧さ，作業者の習熟度などにより発生する．理想的な学習データとしては全てのアノテーションの曖昧性が人手によって解決されていることが期待されるが，Penn Treebank コーパスの様に明確にタグ付け基準が定められている場合ですら，**表 1** に示すような品詞に曖昧性を含む文が 100 文以上存在する．ただし，NNS:名詞複数，RBR:副詞比較級，IN:前置詞または従属接続詞，RB:副詞である.

さらに，意味を扱う情報抽出などのタスクでは作業者が曖昧性を判断できないケースがしばしば発生する．このように，人手によるラベル付きデータの作成過程では曖昧なアノテーションが付与される可能性は高く，学習アルゴリズムはこれらに対応する必要がある.

3.　問題の定式化

本節では，部分的なアノテーションや曖昧なアノテーションを学習データに用いた，構造出力問題を定式化する．本研究では，特に入出力構造を列構造に限って説明するが，一般の木構造などにも適用できる.

入力列 $\boldsymbol{x}=(x_1, x_2, \cdots, x_T) \in \boldsymbol{X}$ を入力列 $x_t \in X$ を要素とする列構造，ラベル列 $\boldsymbol{y}=(y_1, y_2, \cdots, y_T) \in \boldsymbol{Y}$ をラベル変数 $y_t \in Y$ の列とすると，教師付き構造出力学習は写像 $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ を学習する問題として定義できる．例えば日本語の単語分割タスクでは $\boldsymbol{x}$ は文字境界を表す変数の列，ラベル列 $\boldsymbol{y}$ は単語境界の有無を表すラベル変数の列☆☆であり，品詞付与問題では，$\boldsymbol{x}$ は文の各単語を表す変数の列，$\boldsymbol{y}$ は対応する品詞タグの列である.

次に，$\boldsymbol{y}$ の一部だけが与えられたデータを表現するために，$\boldsymbol{L}=(L_1, L_2, \cdots, L_T)$ を入力 $\boldsymbol{x}$ の各点 t が取り得るラベル変数の値集合 $L_t \in 2^Y - \{\emptyset\}$ の列とする．例えば 2.1 節の KWIC 形式の UI で長さ 6 文字 (文字境界数 = 5) の文字列に部分的にアノテーションしたとすると,

$$\boldsymbol{L} = (\{\bigcirc, \times\}, \underbrace{\{\bigcirc\}, \{\times\}, \{\bigcirc\}}_{\text{部分的なアノテーション}}, \{\bigcirc, \times\})$$

となる．ただし，$t=2$ から $t=4$ までがアノテーションした単語の文字境界であり，単語境界と非単語境界のラベルを ○ と × とする．また，2.2 節の曖昧な品詞タグの例を表現すると，$t=4$ にのみ複数のラベル (VBG,JJ) を含む $\boldsymbol{L}$ と書ける:

$$\boldsymbol{L} = (\{\mathrm{DT}\}, \{\mathrm{NN}\}, \{\mathrm{VBZ}\}, \underbrace{\{\mathrm{VBG}, \mathrm{JJ}\}}_{\text{曖昧なアノテーション}}, \{.\})$$

なお，列構造全体が全てアノテーションされたデータは，$t=1, \cdots, T$ の全てで要素サイズ $|L_t|=1$ である $\boldsymbol{L}$ で表現される.

以上のように定義すると，本稿で扱う教師付き学習問題は N 個の部分的で曖昧なアノテーションが付与された構造データ $D=\{(\boldsymbol{x}^{(n)}, \boldsymbol{L}^{(n)})\}_{n=1}^{N}$ を用いてラベル付与器を学習する問題と言える.

4.　周辺尤度最大化による条件付確率場の学習

本節では，曖昧かつ部分的なアノテーションを活用した CRF の学習方法を提案する．$\boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{X} \times \boldsymbol{Y} \rightarrow \Re^d$ を入力列 $\boldsymbol{x}$ とラベル列 $\boldsymbol{y}$ の組から d 次元の任意の素性ベクトルへの写像，$\boldsymbol{\theta} \in \Re^d$ をモデルのパラメータベクトルとする．CRF は $\boldsymbol{x}$ が与えられた時の $\boldsymbol{y}$ の条件付確率を次式でモデル化する.

$$P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta} \cdot \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{y})}}{Z_{\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{Y}}}, \qquad (1)$$

ただし，ベクトル $\boldsymbol{v}$ と $\boldsymbol{w}$ の内積を $\boldsymbol{v} \cdot \boldsymbol{w}$ とし，分母

☆ ただし，Penn Treebank コーパスでは，タグの候補が複数ある場合の記述の順番は重要ではなく，その順番には一貫性は無い[15].

☆☆ 文字の前が単語境界か否かを各文字に付与する Peng ら[12] の単語分割問題の定式化では，単語境界があることが自明である文の最初の文字にもラベルを付与することになり冗長であるため，本稿では上記の定式化を採用した.

は確率（和が 1）にするための正規化項

$$Z_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{Y}} = \sum_{\boldsymbol{y}\in\boldsymbol{Y}} e^{\boldsymbol{\theta}\cdot\boldsymbol{\Phi}(\boldsymbol{x},\boldsymbol{y})}$$

である．また，$\boldsymbol{\theta}$ が与えられたもとでは，ラベル列は $\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}\in\boldsymbol{Y}}{\operatorname{argmax}} P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ で予測する．

オリジナルの CRF では，学習データ $(\boldsymbol{x},\boldsymbol{y})$ にはラベル列 $\boldsymbol{y}$ が必要になるため，$\boldsymbol{y}$ の一部だけが与えられた $\boldsymbol{L}$ から直接学習することが出来ない．素朴には，$\boldsymbol{L}$ に適合するあらゆるラベル列の集合を $\boldsymbol{Y_L}$ としたとき，ラベル列 $\boldsymbol{y}\in\boldsymbol{Y_L}$ を適度に重み付けして全てを学習データとする方法が考えられる．しかし，ありえるラベル列の数は $|\boldsymbol{Y_L}| = |L_1|\times|L_2|\times\cdots|L_T|$ であり，$|L_t|>1$ となる t の数に対して指数的に増加してしまい，明示的にラベル列を列挙して学習データとする方法は汎用性が無い．

そこで，効率的に上記の計算を行う方法を考える．まず，集合 $\boldsymbol{Y_L}$ の分布をモデル化した周辺分布：

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y_L}|\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\boldsymbol{Y_L}} P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$$

を考える．このモデルのパラメータ $\boldsymbol{\theta}$ の最尤推定値は次式の対数尤度を最大化することで得られる:

$$\begin{aligned}\mathrm{LL}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \ln P_{\boldsymbol{\theta}}(\boldsymbol{Y}_{\boldsymbol{L}(n)}|\boldsymbol{x}^{(n)}) \qquad (2)\\ &= \sum_{n=1}^{N}\sum_{\boldsymbol{y}\in\boldsymbol{Y}_{\boldsymbol{L}(n)}} \ln P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}^{(n)})\\ &= \sum_{n=1}^{N}\left(\ln Z_{\boldsymbol{\theta},\boldsymbol{x}^{(n)},\boldsymbol{Y}_{\boldsymbol{L}(n)}} - \ln Z_{\boldsymbol{\theta},\boldsymbol{x}^{(n)},\boldsymbol{Y}}\right).\end{aligned}$$

式 (2) は，与えられた $\boldsymbol{x}$ に対する $\boldsymbol{y}$ の条件付確率 $P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ を重みとして $\boldsymbol{Y_L}$ 全てを学習データにしていることに相当する．

式 (2) はパラメータ $\boldsymbol{\theta}$ に関して凸であり，通常の CRF と同様に勾配法によって大域最適解が得られる[18]．CRF の学習では最尤推定による過学習を防ぐために，パラメータの事前分布 $P(\boldsymbol{\theta})$ と式 (2) を合わせて $\boldsymbol{\theta}$ の事後確率 $\mathrm{LL}(\boldsymbol{\theta})+\ln P(\boldsymbol{\theta})$ を目的関数とし，これを最大化する MAP 推定が用いられる．5 節の実験では，平均 0，分散 σ^2 の正規分布を事前分布とし，次式の正則化項付き対数周辺尤度関数を目的関数とした．

$$\begin{aligned}&\mathrm{RL}(\boldsymbol{\theta})\\ &= \sum_{n=1}^{N}\left(\ln Z_{\boldsymbol{\theta},\boldsymbol{x}^{(n)},\boldsymbol{Y}_{\boldsymbol{L}(n)}} - \ln Z_{\boldsymbol{\theta},\boldsymbol{x}^{(n)},\boldsymbol{Y}}\right) - \frac{||\boldsymbol{\theta}||^2}{2\sigma^2},\end{aligned} \qquad (3)$$

ただし $\boldsymbol{\theta}$ に依らない定数 $\ln 1 - \ln\sqrt{2\pi}\sigma$ は省略した．

式 (3) と共に，勾配法では必要になる式 (3) の偏微分は次式で与えられる．

$$\begin{aligned}&\frac{\partial \mathrm{RL}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\\ &= \sum_{n=1}^{N}\left(\sum_{\boldsymbol{y}\in\boldsymbol{Y}_{\boldsymbol{L}(n)}} P_{\boldsymbol{\theta},\boldsymbol{L}}(\boldsymbol{y}|\boldsymbol{x}^{(n)})\boldsymbol{\Phi}(\boldsymbol{x}^{(n)},\boldsymbol{y})\right.\\ &\qquad \left. - \sum_{\boldsymbol{y}\in\boldsymbol{Y}} P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}^{(n)})\boldsymbol{\Phi}(\boldsymbol{x}^{(n)},\boldsymbol{y})\right) - \frac{\boldsymbol{\theta}}{\sigma^2},\end{aligned} \qquad (4)$$

ただし，

$$P_{\boldsymbol{\theta},\boldsymbol{L}}(\boldsymbol{y}|\boldsymbol{x}^{(n)}) = \frac{e^{\boldsymbol{\theta}\cdot\boldsymbol{\Phi}(\boldsymbol{x},\boldsymbol{y})}}{Z_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{Y_L}}}$$

は $\boldsymbol{L}$ に適合するラベル列 $\boldsymbol{Y_L}$ のみで正規化した条件付分布である．式 (4) は最適解で 0 となるので，目的関数の最大化は $P_{\boldsymbol{\theta},\boldsymbol{L}}(\boldsymbol{Y_L}|\boldsymbol{x})$ と $P_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{x})$ の下での素性の期待値が等しくなるようにパラメータ $\boldsymbol{\theta}$ を選択していると言える．

式 (3,4) はラベル列の集合 $\boldsymbol{Y},\boldsymbol{Y_L}$ での和を含むが，先に述べたように明示的にラベル列集合を列挙して評価することは現実的でない．そこで，マルコフ性を仮定することで動的計画法により効率的に計算する方法を以下に示す．なお，以下の説明では 1 次のマルコフモデルを説明するが，2 次以上のマルコフモデルやセミマルコフモデル[16] にも容易に拡張可能である．

1 次のマルコフモデルでは，$(\boldsymbol{x},\boldsymbol{y})$ の各点 t において，入力変数とラベル変数の組の素性 $\boldsymbol{f}(x_t,y_t)$：$X\times Y$ と隣接するラベル変数の組の素性 $\boldsymbol{g}(y_{t-1},y_t)$：$Y\times Y$ を考え，その線形結合を $\boldsymbol{\phi}(x_t,y_{t-1},y_t) = \boldsymbol{f}(x_t,y_t)+\boldsymbol{g}(y_{t-1},y_t)$ と書く．すると，素性ベクトルは $\boldsymbol{\Phi}(\boldsymbol{x},\boldsymbol{y}) = \sum_{t=1}^{T+1}\boldsymbol{\phi}(x_t,y_{t-1},y_t)$ と分解できる．ただし，ラベル列の先頭と末尾を表す特別なラベルをそれぞれ S と E とするとき，$\boldsymbol{\phi}(x_t,y_{t-1},y_t)$ は先頭 $t=1$ で $\boldsymbol{\phi}(x_t,S,y_t)$，末尾 $t=T+1$ で $\boldsymbol{g}(y_{t-1},E)$ と定義する．

式 (3,4) の第 1,2 項を効率的に計算するポイントは，ラベル列ごとの再計算を避けるためにある $\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}$ に対して行列 $\alpha_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}}[t,j],\beta_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}}[t,j]$ をあらかじめ計算することである．なお，$\boldsymbol{Y}$ の和の計算時は $\boldsymbol{L} = (Y,\cdots,Y)$ とすればよい．これはよく知られた Forward-Backward アルゴリズムを制約 $\boldsymbol{L}$ を満たすように拡張したものであり，α は $t=1,\cdots,T$，β は $t=T+1,\cdots,1$ の順で以下のように計算する．

$$\begin{aligned}&\alpha_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}}[t,j]\\ &= \begin{cases}0 & \text{if } j\notin L_t\\ \boldsymbol{\theta}\cdot\boldsymbol{\phi}(x_t,S,j) & \text{else if } t=1\\ \ln\sum_{i\in L_{t-1}} e^{\alpha[t-1,i]+\boldsymbol{\theta}\cdot\boldsymbol{\phi}(x_t,i,j)} & \text{else}\end{cases}\end{aligned}$$

$$\beta_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}}[t,j] = \begin{cases} 0 & \text{if } j \notin L_t \\ \boldsymbol{\theta}\cdot\boldsymbol{g}(j,E) & \text{else if } t = T+1 \\ \ln \sum_{k\in L_{t+1}} e^{\boldsymbol{\theta}\cdot\boldsymbol{\phi}(x_t,j,k)+\beta[t+1,k]} & \text{else} \end{cases}$$

以降，α, β, Z の添え字 $\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{L}$ は誤解の無い範囲で省略する．行列 $\alpha[t,j]$（$\beta[t,j]$）は各点 s でラベル y_s が L_s に含まれる点 t までの前（後）からの部分ラベル列の中で，$y_t = j$ である全ての部分ラベル列の指数の和の対数値☆を格納している．α, β の計算量は $O(T|Y|^2)$ である．

最後に，対数尤度（式 (3)）とその偏微分（式 (4)）を α, β を使って計算する方法を示す．式 (3) の Z の対数は α, β を使って次式で得られる：

$$\ln Z_{\boldsymbol{\theta},\boldsymbol{Y_L}} = \ln \sum_{j\in L_T} e^{\alpha_{\boldsymbol{\theta},\boldsymbol{L}}[T,j]} + \boldsymbol{\theta}\cdot\boldsymbol{g}(j,E),$$

また，式 (4) の第 1,2 項は同様に次式で計算出来る．

$$\sum_{\boldsymbol{y}\in\boldsymbol{Y_L}} P_{\boldsymbol{\theta},\boldsymbol{L}}(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{\Phi}(\boldsymbol{x},\boldsymbol{y}) = \sum_{t=1}^{T}\sum_{j\in L_t}\left(\gamma_{\boldsymbol{L}}(t,j)\boldsymbol{f}(x_t,j) + \sum_{i\in L_{t-1}} \varepsilon_{\boldsymbol{L}}(t,i,j)\boldsymbol{g}(i,j)\right) + \sum_{i\in L_T} \varepsilon_{\boldsymbol{L}}(T,i,E)\boldsymbol{g}(i,E)$$

ただし，$\gamma_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}}$ 及び $\varepsilon_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}}$ は次式の周辺確率である（上式で下付きの $\boldsymbol{\theta}, \boldsymbol{x}$ は省略した）．

$$\gamma_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}}(t,j) = P_{\boldsymbol{\theta},\boldsymbol{L}}(y_t = j|\boldsymbol{x}) = e^{\alpha[t,j]+\beta[t,j]-\ln Z_{\boldsymbol{Y_L}}}$$

$$\varepsilon_{\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{L}}(t,i,j) = P_{\boldsymbol{\theta},\boldsymbol{L}}(y_{t-1} = i, y_t = j|\boldsymbol{x}) = e^{\alpha[t-1,i]+\boldsymbol{\theta}\cdot\boldsymbol{\phi}(x_t,i,j)+\beta[t,j]-\ln Z_{\boldsymbol{Y_L}}}$$

以上の方法で，式 (3,4) は $O(T|Y|^2)$ で計算出来る．

5. 実　　験

5.1 部分的単語分割からの学習

本節では，日本語単語分割の分野適応のシナリオにおいて，部分的アノテーションを付与した際の提案法の性能を実データにより検証する．

本実験では適応元のデータとして日常会話辞書[7] の例文を用いた．適応先のデータには家庭用医療マニュアル[4] の文を用いた．適応元の全てのデータ（A，B）と適応先の 1000 文（C）は人手により単語に分割されている（**表 5.1** 参照）．本実験での分野適応の目的は，アノテーション済みの適応元のデータに加え，部分的アノテーションをした適応先データを使用して適応先での単語分割性能の向上である．

表 2　単語分割タスクのデータ．

	分野	分割済	文数	単語数
A	会話文	○	11,700	145,925
B	会話文	○	1,300	16,348
C	医療マニュアル	○	1,000	29,216
D	医療マニュアル	×	53,834	N/A

入力文字境界を表す二値素性としては，各文字境界に隣接する文字及び字種ユニグラムと文字境界の前後 2 文字の領域に含まれる文字及び字種 n グラム（$n = 2, 3$）を使用した．字種素性は字種が異なる文字境界では単語境界になりやすい事前知識を反映したもので，字種はひらがな，カタカナ，漢字，英字，アラビア数字，記号からなる．たとえば，字種のひらがなを H，漢字を C とすると，文字列「やすり傷」の中央文字境界を表す素性は { す|，|り，やす|，す|り，|り傷，やす|り，す|り傷，H|，|H，HH|，H|H，|HC，HH|H，H|HC} である．ただし，「|」は注目する文字境界との相対位置を示す補助記号である．パラメータ数を減らすためにテストデータではない A と D での頻度を用いて素性選択をした．具体的には，C_A と C_D をそれぞれデータ A と D における素性の出現頻度としたとき，式 $C_A + 0.5C_D \geq 2$ を満たす素性のみを用いた．最終的な入力素性数は 298,363 である．

性能評価には標準的な再現率（R）と精度（P）の調和平均値 $F = 2RP/(R+P)$ を用いた．ただし，

$$R = \frac{\text{正解単語数}}{\text{全単語数}} \times 100,$$

$$P = \frac{\text{正解単語数}}{\text{システムの出力単語数}} \times 100.$$

また，実験では 1 次のマルコフモデルの CRF を実装した．適応元のモデルは学習用データ A を使って学習し，開発用データ B で性能を評価し最適なハイパーパラメータ σ を決定した．$\sigma = 1.75$ の時，データ B で最高の性能 $F = 97.56$ となり，以降このモデルを CRF^S とする．

適応先データ中での単語リスト中の単語の出現箇所をサンプリング手法により選択した部分的アノテーションから学習した CRF^T の性能を確認した．

本実験では，適応先データ C を，C1）部分的アノテーション用及び学習用 500 文と C2）評価対象 500 文に分け 2 分割交差検定を行った．単語リストとしては，適応元データ A には出現しない適応先 C1 の単語集合を使用した．この単語リストのエントリ数は交差検定の各分割の平均で 948.5 であった．単語リストの各単語についてアノテーションする出現箇所数を 1

☆ 数値計算時にはオーバーフローを防ぐための工夫が必要になる．詳細は文献 19) の 1.4.6 章を参照．

表 3 実験で変化させた単語リスト中の単語アノテーション箇所数（交差検定での平均値）.

	エントリ毎	総数	/全単語
100-900	最高 1 回	100-900	0.6%-6.1%
a1	1 回	948.5	6.5%
a2	最高 2 回	1,355	9.3%
comp	全出現	14,608	100%

回 (*a1*) または最高 2 回 (*a2*) と変化させ，部分的アノテーションが付与された文を学習データに追加し学習結果を評価した．また，単語リストの全てのエントリを確認できない場合を想定し，アノテーション数を 100 から 900 に変化させて性能を評価した．**表 5.1** に単語リストのエントリ毎のアノテーション箇所数（エントリ毎），平均総単語アノテーション数（総数），C1 の全単語出現数の平均（14,608）との比率（/全単語）を示す．

表 5.1 の *comp* を除き，選択的サンプリング法により各単語の出現箇所に優先順位付けをした．選択的サンプリングの指標にはラベルエントロピー[2]:

$$H(\boldsymbol{y}_t^s) = \sum_{\boldsymbol{y}_t^s \in \boldsymbol{Y}_t^s} P_\theta(\boldsymbol{y}_t^s|\boldsymbol{x}) \ln P_\theta(\boldsymbol{y}_t^s|\boldsymbol{x})$$

を用いた．ただし入力 $\boldsymbol{x}$ の t から s までの長さ $s-t+1$ の部分ラベル列を $\boldsymbol{y}_t^s = (y_t, y_{t+1}, \cdots, y_s) \in \boldsymbol{Y}_t^s$ とする．注意する点として，能動学習の文脈では学習とサンプル選択を交互に行うことを想定しているが，本実験ではラベルエントロピーが高い順に一度に設定数のアノテーション箇所を選択した．なぜなら，CRF の最適化には時間がかかるため，繰り返し学習を行うことはインタラクティブなアノテーションシステムでは現実的でないからである．

提案法（以下 *mrg*）と比較のため，部分的アノテーション $\boldsymbol{L}$ に適合するラベル列 $\boldsymbol{Y_L}$ 中，CRF^S で最も確率の高いラベル列を学習データとして使用する手法（以下 *argmax*）を実装した．適応先のデータは限られておりハイパーパラメータの開発用に適応先データは使えない場合を想定し，CRF^S と同じ $\sigma = 1.75$ を使用した．

まず最初に，CRF^S は適応元での性能（$F = 97.56$）に比べて，適応先では大きく性能が低下した（$F = 92.23$）．一方，文全体をすべて単語分割した適応先データ（C1）500 文のみを使用した学習により性能は $F = 95.91$ に向上した．さらに，文全体を単語分割済み適応先データ（C1）と適応元データ（A）を両方使用した学習が最も良い性能を示した（$F = 96.74$）．この性能を *comp* として適応先データを追加した性能の上限として参照点とする．

図 2 はアノテーション単語数を変化させた時の性能

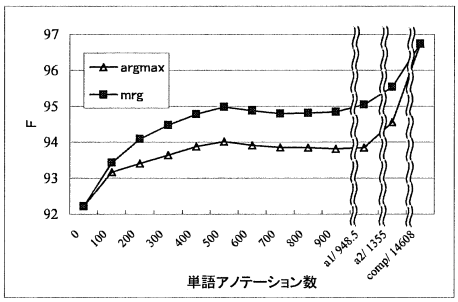

図 2 単語リストの一部にアノテーションした時の性能（交差検定での平均値）.

表 4 品詞タグ付けタスクの学習データ.

	実験 1	実験 2
品詞が曖昧な単語を含む文	118 文	
品詞がすべて一意に定まった文	1480 文	2960 文

の変化を示している．提案手法を選択的サンプリング手法と組み合わせることで，6.5%(*a1*) 及び 9.3%(*a2*) のアノテーション数で *comp* の性能向上の内 63% 及び 73% が得られた．また，*argmax* との比較では $\boldsymbol{L}$ に適合するラベル列 y の分布を考慮した提案手法は，分布のピークの y のみを正解として使用した *argmax* に比べて常によい性能を示した．

5.2 曖昧な品詞タグからの学習

本節では曖昧な品詞タグを使用した品詞タグ付けの実験結果を示す．

本実験では，Penn Treebank コーパス品詞タグ付けデータ[10] の品詞が曖昧なアノテーションが含まれている 118 文を使用した．以降，この文を品詞曖昧文と呼ぶ．曖昧にアノテーションされた単語は全 82 種類である（例：表 1）．本実験での目的は，これらの単語に対して品詞が一意に定まっている文（品詞一意文）と品詞曖昧文を合わせて学習データとしたときの，品詞タグ付け性能の向上である．実験で学習に使用した文数を**表 4** に示した．

各単語を表す入力素性は，単語自身の文字列，単語末尾 1,2,3 文字列，先頭が大文字，先頭が数字，先頭が大文字かつドット（.）を含む，全て大文字，全て小文字，ハイフン（-）を含む，句読点を含む，文の最後が “,”，“?”，“!” で終わる，を表す二値素性である．頻度 2 以上の素性のみを用いて入力素性数は 14,391 であった．なお，Penn Treebank では句読点自身が品詞タグとなっているが記号タグ (SYM) としてまとめた．

提案手法の比較対象として，118 文に含まれる曖昧なタグをルールにより一意に定める方法を用いた．曖

表 5　品詞タグ付けタスクの実験結果（試行 5 回の平均）.

	正解率	提案法	ランダム	記述順	頻度順
実験 1	全体	**94.274**	**94.274**	94.262	**94.274**
	曖昧語	**73.272**	71.582	72.658	71.68
実験 2	全体	**94.982**	94.98	94.974	94.976
	曖昧語	**76.242**	74.276	75.28	74.326

昧なタグの決定には，1) ランダム，2) タグが記述された順☆，3) コーパスでのタグ頻度順，を用いた．評価には学習データとは別の品詞一意文 11,840 文を用い，評価指標には

$$\text{全体正解率} = \frac{\text{全単語品詞タグ正解数}}{\text{全単語出現数}},$$

$$\text{曖昧語正解率} = \frac{1}{|A|}\sum_{w\in A}\frac{w\text{ の品詞タグ正解数}}{w\text{ の出現数}},$$

を用いた．ただし，A は曖昧な品詞アノテーションが存在する単語 82 個の集合である．

表 5 にデータを換えた 5 回の試行の平均結果を示す．曖昧なアノテーションが付与された単語が学習データ数の中で少ないため，全体の正解率の向上にはそれほど寄与しなかった．しかし，曖昧な品詞アノテーションがあった個々の単語の正解率（曖昧語正解率）では，提案手法により性能向上が観察された．タグ記述順は，曖昧語正解率に関してはランダム，頻度順に比べて良い性能を示したが全体正解率では他に劣った．一方，提案手法は安定して他の手法より良いまたは同じ程度の性能を示した．

5.3 議　　論

両実験とも，提案法により部分的アノテーションや曖昧なアノテーションを用いて効果的に学習することが出来ていることが確かめられた．

5.1 節の実験では，部分的なアノテーション以外の部分のラベルの曖昧性を保持する方法が，既存のモデルでラベルを付与し固定的なアノテーションとして学習データとした方法に比べて良い性能を得ることが出来た．5.2 節の実験でも，決定的なルールによって曖昧なラベル候補から一つに決めて学習データとする方法に比べて，自然にラベルの曖昧性を扱う提案手法によって安定した性能を得られることが確かめられた．

一方で，5.1 節の実験では，単語リストによる部分的アノテーションの追加に関して必ずしも性能が単調増加しない現象が観察され，単語アノテーション数 500 ～ 800 間では逆に性能の一時的低下が発生した（図 2）．500 と 800 の分割結果を比較したところ，800 では短い単語同士を長い単語として認識する誤りにより性能低下が起きていた．理由の一つとしては，500 から 700 までの間にラベルエントロピーにより選択された単語の単語長が他の間隔より長いことがあげられる．単語内の文字境界は非単語境界としてアノテーションされるため，長い単語が学習データに多く出現することで単語境界に比べて非単語境界に重点を置いて学習されてしまったと考えられる．ラベルエントロピーは最優先で難しい文脈（たとえば，隣接する単語と字種が同じ場合）の単語出現箇所を選択し，第 2 にラベルのエントロピーが大きくなりやすい長い単語を選択するため，500 ～ 700 間に長い単語が選択されていた．結果的に，単語非境界ラベルが学習データに偏って与えられたため，性能が劣化したと考えられる．学習データがサンプリング法によって本来の分布から偏ってしまうことは一般的に起こる問題であるが，文全体のアノテーションに比べて部分的アノテーションでは文のどこにアノテーションを行うかによってもデータに偏りが生じるため，偏りの補正が今後の重要な課題である．

☆ Penn Treebank コーパスでは曖昧な場合のタグ記述順は優先度を表すものではない[15] としているが，このルールではアノテーション作業者が何らかの意思を持ってタグの順番を決めていると仮定している．

6. 関 連 研 究

構造出力タスクにおいて部分構造を選択する選択サンプリングまたは能動学習手法の先行研究[1),3),14),17)]は，いずれも各部分構造は独立に学習することが出来る問題を仮定していた．提案手法は要素間の依存関係がある構造出力問題ではより適切な学習アルゴリズムであると言える．Culotta[6] らは CRF を活用したインタラクティブなアノテーションシステムの研究において，部分的なアノテーションを満たすラベル列を修正し作業量を減らす方法を提案している．この手法では，最終的には文全体のラベルが修正されることを仮定しており，CRF は通常通り学習される．作業者がすべてのラベルを修正できない場合には部分的アノテーションが残るため，提案手法を組み合わせることが有効である．Pereira ら[13] は，部分的に構成素領域がアノテーションされた構文木から確率的文脈自由文法の獲得方法を提案した．Pereira らの手法は木構造出力における部分的アノテーションからの生成モデルの学習と考えられ，提案法に置き換えることで識別モデルによる構文解析が可能になる．

7. 結論と今後の展開

本研究では，現実でのアノテーション作業により付与されうる曖昧かつ部分的なアノテーションを使用した条件付確率場のパラメータ推定法を提案した．計算

機実験により，提案法が大きくアノテーション作業を減らすと共に性能を向上させることを確かめられた．

提案する学習法は，構文解析，情報抽出など，その他の構造出力タスクにも応用可能である．しかし，対訳語対応付け[22]などマルコフ性を仮定できない他の自然言語処理の問題では，あらゆるラベル構造の割り当てスコアの総和を動的計画法で計算することが出来ない．提案法は確率モデルに基づいており総和の計算を含むため適用することは出来ないが，これらの識別構造問題で曖昧で部分的なアノテーションが与えられた際の学習方法は興味深い今後の研究テーマである．

参考文献

1) Anderson, B. and Moore, A.: Active Learning for Hidden Markov Models: Objective Functions and Algorithms, *Proceedings of the 22nd International Conference on Machine Learning*, pp.9–16 (2005).
2) Anderson, B., Siddiqi, S. and Moore, A.: Sequence Selection for Active Learning, Technical ReportCMU-IR-TR-06-16, Carnegie Mellon University (2006).
3) Argamon-Engelson, S. and Dagan, I.: Committee-Based Sample Selection for Probabilistic Classifiers, *Journal of Artificial Intelligence Research*, Vol.11, pp.335–360 (1999).
4) Beers, M.H.: メルクマニュアル医学百科—最新家庭版，日経 BP 社 (2004).
5) Collins, M.: Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2002).
6) Culotta, A., Kristjansson, T., McCallum, A. and Viola, P.: Corrective Feedback and Persistent Learning for Information Extraction, *Artificial Intelligence Journal*, Vol.170, pp.1101–1122 (2006).
7) Keene, D.，羽鳥博愛，羽鳥博愛，伊良部祥子(編)：会話作文英語表現辞典，朝日出版社 (1992).
8) Kudo, T., Yamamoto, K. and Matsumoto, Y.: Appliying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of Empirical Methods in Natural Language Processing* (2004).
9) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data, *Proceedings of the 18th International Conference on Machine Learning* (2001).
10) Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol.19, No.2 (1993).
11) Mori, S.: Language Model Adaptation with a Word List and a Raw Corpus, *Proceedings of the 9th International Conference on Spoken Language Processing* (2006).
12) Peng, F., Feng, F. and McCallum, A.: Chinese Segmentation and New Word Detection using Conditional Random Fields, *Proceedings of the International Conference on Computational Linguistics* (2004).
13) Pereira, F. C. N. and Schabes, Y.: Inside-Outside Reestimation from Partially Bracketed Corpora, *Proceedings of Annual Meeting Association of Computational Linguistics*, pp.128–135 (1992).
14) Roth, D. and Small, K.: Margin-based Active Learning for Structured Output Spaces, *Proceedings of the European Conference on Machine Learning*, Springer, pp.413–424 (2006).
15) Santorini, B.: *Part-of-Speech Tagging Guidlines for the Penn Treebank Project*, University of Pennsylvania, 3rd revision, 2nd printing edition (1995).
16) Sarawagi, S. and Cohen, W.W.: Semi-Markov Conditional Random Fields for Information Extraction, *Advances in Neural Information Processing Systems* (2005).
17) Scheffer, T. and Wrobel, S.: Active learning of partially hidden markov models, *Proceedings of the ECML/PKDD Workshop on Instance Selection* (2001).
18) Sha, F. and Pereira, F.: Shallow Parsing with Conditional Random Fields, *Proceedings of Human Language Technology-NAACL*, Edmonton, Canada (2003).
19) Sutton, C. and McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning, *Introduction to Statistical Relational Learning* (Getoor, L. and Taskar, B., eds.), MIT Press (2006).
20) Taskar, B., Guestrin, C. and Koller, D.: Max-Margin Markov Networks, *Proceedings of the Conference on Neural Information Processing Systems Conference* (2003).
21) Taskar, B., Klein, D., Collins, M., Koller, D. and Manning, C.: Max-Margin Parsing, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2004).
22) Taskar, B., Lacoste-Julien, S. and Klein, D.: A Discriminative Matching Approach to Word Alignment, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2005).