グラフ構造を持つ条件付確率場による Wikipedia 文書中の固有表現分類

Graph-structured Conditional Random Fields for Named Entity Categorization in Wikipedia

渡邊 陽太郎
Yotaro Watanabe

奈良先端科学技術大学院大学
Graduate School of Information Science, Nara Institute of Science and Technology
yotaro-w@is.naist.jp, <http://cl.naist.jp/~yotaro-w/>

浅原 正幸
Masayuki Asahara

(同 上)
masayu-a@is.naist.jp, <http://cl.naist.jp/~masayu-a/>

松本 裕治
Yuji Matsumoto

(同 上)
matsu@is.naist.jp, <http://cl.naist.jp/staff/matsu/>

keywords: named entity acquisition, collective classification, conditional random fields

Summary

This paper presents a method for categorizing named entities in Wikipedia. In Wikipedia, an anchor text is glossed in a linked HTML text. We formalize named entity categorization as a task of categorizing anchor texts with linked HTML texts which glosses a named entity. Using this representation, we introduce a graph structure in which anchor texts are regarded as nodes. In order to incorporate HTML structure on the graph, three types of cliques are defined based on the HTML tree structure. We propose a method with Conditional Random Fields (CRFs) to categorize the nodes on the graph. Since the defined graph may include cycles, the exact inference of CRFs is computationally expensive. We introduce an approximate inference method using Tree-based Reparameterization (TRP) to reduce computational cost. In experiments, our proposed model obtained significant improvements compare to baseline models that use Support Vector Machines.

1. は じ め に

固有表現とは、人名、地名、組織名などの固有名詞や、時間表現、日付表現などの総称である。これらは、自然言語処理の応用分野である関係抽出や情報検索、質問応答などにおいて重要な役割を果たす。現実には多様な固有表現が存在することから、形態素解析時には辞書に登録されていない表現が多く出現し、未知語になりやすいという問題がある。そのため、解析誤りを避けるためには、多くの固有表現を辞書に登録しておくことが有効である。

固有表現抽出の手法は近年では機械学習に基づくものが主に用いられているが、テキスト中の固有表現を全て網羅できるわけではない。そこで、獲得できる固有表現については、既存の資源などを利用して獲得しておき、それを固有表現抽出の手がかりとして利用することで解析精度の向上が期待される。

固有表現獲得のための資源として我々は Web 上の多言

語百科事典である Wikipedia^{*1}に注目した。Wikipedia には数多くの固有表現が数多く登録されている。各記事には語釈文やカテゴリなどの情報が付与されており、これらは抽出の手がかりとして有用な情報である。また、各記事は HTML によって記述されているため構造を持ち、さらにアンカーによって別の関連する記事を得ることができる。このような特徴を持つ Wikipedia は、固有表現の獲得に適した資源であると考えられる。

Wikipedia 中の記事は HTML(半構造データ)であり、構造化されていないテキストには無い特徴がある。特に注目すべき点として、文書中に存在するアンカーの出現に依存関係があることが挙げられる。例えば、リストにおいて列挙されているアンカーは、同じ固有表現クラスに属する事柄の記事を参照しているといった傾向がある。このようなアンカー間の依存関係を捉えた分類をおこなうことで、より高い精度での固有表現の獲得が期待される。

これまでに、このような依存関係を考慮した分類手法がいくつか提案されてきており、これらは大きく 2 種類

*1 <http://ja.wikipedia.org/>

に分類することができる．一つは、局所的な分類を繰り返すことで全体の分類をおこなう Iterative Classification [Neville 00, Lu 03b, Lu 03a]、もう一つは大域的最適化に基づく分類手法である Collective Classification [Getoor 01, Taskar 02] である．本稿では、アンカー間の依存関係を考慮した分類をおこなうために、大域的最適化に基づく無向グラフモデルである条件付確率場 (CRFs) [Lafferty 01] を用いて Wikipedia 中の固有表現の分類をおこなう．CRFs を適用する上で重要であるのは、どのようなグラフ構造を構成するかであるが、本稿では、DOM 構造によって捉えられるアンカー間の依存関係をグラフ構造に反映させた CRFs のモデルを提案する．

本稿の構成は以下の通りである．まず 2 章にて条件付確率場について述べ、3 章にて条件付確率場を Wikipedia から固有表現を獲得する問題へ適用する方法について述べる．次に、4 章にて評価実験により提案手法を検証する．5 章にて、関連研究について述べ、最後に 6 章にてまとめと今後の課題について述べる．

2. 条件付確率場

条件付確率場 (Conditional Random Fields; CRFs) は観測系列 x が与えられた時のラベル系列 y の条件付確率 $p(y|x)$ を与える確率モデルである．

CRFs は以下のように定義される． $\mathcal{G} = \{V, E\}$ を確率変数 y, x 上の無向グラフとする．ここで V はグラフ $\mathcal{G}$ 内のノード集合、 E はエッジ集合である．クリーク集合 $\mathcal{C} = \{\{y_c, x_c\}\}$ が与えられたとき、CRFs は観測系列 x が与えられた時のラベル系列 y の条件付確率 $p(y|x)$ を定義する．

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \Phi(y_c, x_c) \quad (1)$$

ここで、 $\Phi(x_c, y_c)$ はクリーク上のポテンシャル関数、 $Z(x)$ は正規化項で、これは可能なラベル割り当ての各条件付確率についての和である．

ポテンシャル関数が事前に与えられる素性関数の集合 $F = \{f_1, \dots, f_K\}$ と、各素性に割り当てられる重みの集合 $\Lambda = \{\lambda_1, \dots, \lambda_K\}$ に分解可能であると仮定すると、以下の式で与えられる．

$$\Phi(y_c, x_c) = \exp \left(\sum_k \lambda_k f_k(y_c, x_c) \right) \quad (2)$$

CRFs のモデルパラメータ Λ の最適化は、 N 個の訓練データ $\mathcal{D} = \{y^{(d)}, x^{(d)}\}_{d=1}^N$ が与えられた時、一般に以下の条件付対数尤度を最大化するパラメータ Λ を推定

する．

$$\mathcal{L}_\lambda = \sum_{d=1}^N \sum_{c \in \mathcal{C}^{(d)}} \sum_k \lambda_k f_k(y_c^{(d)}, x_c^{(d)}) - \sum_{d=1}^N \log Z(x^{(d)}) - \sum_{k=1}^K \frac{\lambda_k^2}{2\sigma^2} \quad (3)$$

ここで右辺の 3 番目の項はパラメータに対する事前分布 (Gaussian prior [Chen 99]) であり、 σ^2 はガウス事前分布の分散である．このような項を導入することによってパラメータの正規化がおこなわれ、過学習を防ぐ効果がある．パラメータの事前分布を考慮した最適化は、事後確率最大化 (MAP) と呼ばれる．

目的関数 $\mathcal{L}_\lambda$ の導関数は以下の式となる．

$$\frac{\partial \mathcal{L}_\lambda}{\partial \lambda_k} = \sum_{d=1}^N \sum_{c \in \mathcal{C}^{(d)}} f_k(y_c^{(d)}, x_c^{(d)}) - \sum_{d=1}^N \sum_{c \in \mathcal{C}^{(d)}} \sum_{y_c} f_k(y_c, x_c^{(d)}) p(y_c | x_c^{(d)}) - \frac{\lambda_k}{\sigma^2} \quad (4)$$

目的関数 $\mathcal{L}_\lambda$ は凸関数であるから、勾配法によって大域的最適解を得ることができる．この最適化には様々な手法を適用することができるが、準ニュートン法の一つである L-BFGS [Liu 89] がよく用いられる．

式 (4) の計算には周辺確率 $p(y_c|x) = \sum_{y \setminus y_c} p(y|x)$ を求める必要があるが、この計算はモデルのグラフ構造によってアルゴリズムが異なる．品詞タグ付けといった系列タギングの問題に用いられる Linear-chain CRFs では forward-backward アルゴリズム、木構造の場合は forward-backward アルゴリズムの一般化である Belief Propagation (BP) [Pearl 88] によって周辺確率を正確に計算できる．

BP を用いる場合、各クリークについて、ある特定のラベルに至るまでのポテンシャル関数の累積値 (メッセージと呼ばれる) のベクトル $m = \{m_i(y_j)\}$ を計算し、この値から周辺確率の計算をおこなう．定義するクリーク集合をノード (1 ノードからなるクリーク) とエッジ (2 ノードからなるクリーク) とした場合、ノード i から j へのメッセージ $m_i(y_j)$ の更新式は以下で与えられる．

$$m_i(y_j) = \sum_{y_i} \Phi_v(y_i, x) \Phi_e(y_i, y_j, x) \prod_{k \in \mathcal{N}(i) \setminus j} m_k(y_i) \quad (5)$$

$\mathcal{N}(i)$ はノード i に隣接するノード集合、 $\Phi_v(y_i, x)$ はノード $i \in V$ 上のポテンシャル関数、 $\Phi_e(y_i, y_j, x)$ はエッジ $(i, j) \in E$ 上のポテンシャル関数である．ここで、 Φ_v, Φ_e 内の観測系列を x と記述しているが、これは x は既に与えられており、全体を参照することができるためである．以後、 x_c と限定せず単に x と記述する．

式 (5) を用いて, ノード j の周辺確率 $p(y_j|x)$ は以下の式によって計算できる.

$$p(y_j|x) = \kappa \Phi_v(y_j, x) \prod_{i \in \mathcal{N}(j)} m_i(y_j) \quad (6)$$

ここで, κ は正規化項である.

このように周辺確率はメッセージから計算できるが, 閉路を含むようなグラフ構造 (e.g. Dynamic Conditional Random Fields (DCRFs) [McCallum 03, Sutton 04], Skip-Chain CRFs [Sutton 07]) の場合, 周辺確率の正確な値を求めることが困難であるため, 近似的に計算する手法が用いられる. 周辺確率の近似計算に用いられる手法としては, 閉路の存在を気にせずに BP を適用する Loopy Belief Propagation (Loopy BP) [Murphy 99], グラフから構成できる全域木の集合を列挙し, 各全域木ごとに BP を用いてメッセージを更新していくことで周辺確率の推定をおこなう Tree-based Reparameterization (TRP) [Wainwright 03] などがある.

3. 条件付確率場を用いた Wikipedia 文書中の固有表現分類

Wikipedia では, 一記事につきある特定の事柄に関して記述され, 各記事には, 見出し語, その事柄の定義文, その事柄が属する 1 つ以上の Wikipedia のカテゴリが付与されている. ここで, 定義文やカテゴリなどの情報を手がかりに, その文書で述べられている事柄が何であるかを個別に分類する文書分類の問題として扱うことで固有表現を獲得するという方法が考えられる.

一方, Wikipedia は HTML 文書 (半構造データ) であるため, 生テキストには無い特徴がある. その特徴の中で特に重要な特徴を持っているのは, リスト (など) およびテーブル (<TABLE>) であり, それらの中に出現する要素間には依存関係がある. 本稿では, Wikipedia にて頻繁に出現するリストに焦点を当てることにする. 図 1 は HTML 文書の一部と, 対応する DOM 構造の例である. ここで, タグ直下のリスト において最初に出現するアンカー (「ディラード&クラーク」「カーペンターズ」) は, 同じ固有表現カテゴリに属する. このようにリスト構造内において同列の関係にあるアンカー間には, 双方共に同じカテゴリに属する傾向がある. 実際, Wikipedia の記事の一部に対して人手でタグ付けし, 上記のようにリスト構造内で同列の関係にあるアンカーのペアに関して双方共に同じカテゴリに属するペアの数を調査したところ, 全体の約 79% という結果が得られており, これは分類精度の向上に寄与する特徴であると言える. また, Wikipedia の各記事にはその記事に関連する別の記事を参照するアンカーが存在し, その記事の内容は, アンカーに記述されているテキスト (アンカーテキスト) に表されている.

図 1 DOM 構造と定義したクリークとの対応関係.

ここで, 上記のようなリストの特徴およびアンカーテキストの存在を踏まえると, Wikipedia からの固有表現抽出の問題を, アンカーテキストに対するラベリング問題として扱うということが考えられる. リストなどの規則性によって出現するアンカーテキストの依存関係を分類に利用することで, 個別に分類をおこなうアプローチと比較して高精度な分類が期待できる.

依存関係を考慮した分類をおこなうための手法として CRFs を用い, アンカーテキストをノード集合 V , アンカーテキスト間の依存関係をエッジ集合 E とするような無向グラフ $G = (V, E)$ を構成し, 各ノード $v_i \in V$ へのラベル割り当てが全体で最適となるような割り当て $\arg \max_y p(y|x)$ を求める問題として扱うことで, Wikipedia 中の記事から固有表現を獲得する.

3.1 DOM 構造に基づくグラフ構造の構築

HTML や XML といった半構造化された文書は, DOM (Document Object Model) 構造として表現される. ここで DOM とは, HTML 文書や XML 文書を扱うための Application Programming Interface (API) であり, DOM において HTML および XML 文書の論理構造は木構造の形で扱われる. DOM 構造は, 唯一の親を持ち, 子が順序をもつ順序木として扱うことができる. 図 1 では, 唯一の他と区別される頂点が に対応し, 各タグの子に対応するタグには出現順序があるため, 順序木の定義を満たす. よって以後, DOM 構造を順序木として扱う.

次に, DOM 構造から CRFs を構成する無向グラフ G 上のクリーク集合 C を定義するために, DOM 構造上において依存関係を持つようなアンカーの組について考えることにする.

図 1 の例において,

- (1) 「ディラード&クラーク」と「カントリーロック」は, DOM 構造上では兄弟 (Sibling) の関係であり, 同一のリストの中で出現している. この間には「ディラード&クラーク」の楽曲のジャンルが「カントリーロック」であるというような関係が成り立つ. このような出現の仕方には, 先に出現したものが持っている属性, または関連する事柄が後ろに来るという傾向がある.
- (2) 「ディラード&クラーク」と「カーペンターズ」は,

Sibling $E_S = \{(v_i^T, v_j^T) | v_i^T, v_j^T \in V^T, d(v_i^T, cpa(v_i^T, v_j^T)) = d(v_j^T, cpa(v_i^T, v_j^T)) = 1, v_j^T = ch(pa(v_i^T, 1), k), v_i^T = ch(pa(v_i^T, 1), \max\{l | l < k\})\}$

Cousin $E_C = \{(v_i^T, v_j^T) | v_i^T, v_j^T \in V^T, d(v_i^T, cpa(v_i^T, v_j^T)) = d(v_j^T, cpa(v_i^T, v_j^T)) \geq 2, v_i^T = ch(pa(v_i^T), k), v_j^T = ch(pa(v_j^T), k), pa(v_j^T, d(v_j^T, cpa(v_i^T, v_j^T)) - 1) = ch(pa(v_j^T, d(v_j^T, cpa(v_i^T, v_j^T))), k), pa(v_i^T, d(v_i^T, cpa(v_i^T, v_j^T)) - 1) = ch(pa(v_i^T, cpa(v_i^T, v_j^T)), \max\{l | l < k\})\}$

Relative $E_R = \{(v_i^T, v_j^T) | v_i^T, v_j^T \in V^T, d(v_i^T, cpa(v_i^T, v_j^T)) = 1, d(v_j^T, cpa(v_i^T, v_j^T)) = 3, pa(v_j^T, 2) = ch(pa(v_j^T, 3), k), v_i^T = ch(pa(v_i^T, 1), \max\{l | l < k\})\}$

図 2 Sibling, Cousin, Relative の各クリークの定義。

DOM 構造上では従兄弟 (Cousin) の関係になっており、これらは双方共に組織名である。このような出現の仕方は、双方が同じようなクラスに属するような傾向がある。

- (3) 「カーペンターズ」と「カレン・カーペンター」は、DOM 構造上では、「カーペンターズ」から見て「カレン・カーペンター」が兄弟の孫という関係 (本稿では Relative とする) になっている。これらの間には「カーペンターズ」の構成員として「カレン・カーペンター」が存在するという関係があり、このような出現の仕方は、先に出現したものを構成している要素が後ろに来る傾向がある。

このような、相互に依存関係があるようなアンカーの集合をクリークを構成するノードとすることで、依存関係を捉えた分類をおこなうことができると考えられる。ここで、上記の観察に基づき、DOM 構造上において以下で述べる 3 つのクリークの定義に該当するものを、CRFs のクリークとする。

DOM 構造に対応する順序木を $T^{ordered} = (V^T, E^T)$ とする。頂点 $v_i^T, v_j^T \in V^T$ 間の距離を $d(v_i^T, v_j^T)$ 、順序木の頂点 $v_i^T \in V^T$ から k 回辿った先祖を $pa(v_i^T, k)$ 、頂点 v_i^T の k 番目の子を $ch(v_i^T, k)$ 、頂点 $v_i^T, v_j^T \in V^T$ の共通の先祖を $cpa(v_i^T, v_j^T)$ 、と表すことにする。図 2 に示す 3 つの関係をクリークとして定義する。

ここで Sibling は 1 番目、Cousin は 2 番目、Relative は 3 番目の観察に対応したクリークであり、 E_S は Sibling の関係にある頂点の組の集合、 E_C は Cousin の関係にある頂点の組の集合、 E_R は Relative の関係にある頂点の組の集合を表している。図 1 に、DOM 構造とクリークの対応関係を示す。ここで、上記の関係の定義にて、関係として扱う頂点の組を、ある頂点とその頂点から最も近い頂点に限定している。その理由は、「最も近い」という条件を除いた場合、ある頂点と特定の関係にある頂点の数は複数個該当する可能性が出てくるが、それらの組を全てクリークとした場合、計算量の問題が生じるためである。

図 2 の定義を満たす頂点の組に対応するノード集合および各ノードを、構成する CRFs のクリークとする。すなわち、 $C = E_S \cup E_C \cup E_R \cup V$ である。

3.2 DOM 構造に基づくグラフ構造上の条件付確率場

前節で定義したクリーク C を持つ無向グラフ $G = (V, E)$ のクリークに対してポテンシャル関数を定義することで $p(y|x)$ を与える CRFs を構成する。観測集合 x が与えられたときのラベル集合 y の条件付分布を以下の式で与える。

$$p(y|x) = \frac{1}{Z(x)} \left(\prod_{(v_i, v_j) \in E_S, E_C, E_R} \Phi_{SCR}(y_i, y_j, x) \right) \left(\prod_{v_i \in V} \Phi_V(y_i, x) \right) \quad (7)$$

ここで $Z(x)$ は正規化項、 $\Phi_{SCR}(y_i, y_j, x)$ は Sibling, Cousin, Relative $\{(v_i, v_j) \in E_S, E_C, E_R\}$ に対応するクリークのポテンシャル関数、 $\Phi_V(y_i, x)$ はノード $v_i \in V$ に対応するポテンシャル関数であり、それぞれ以下の式で与える。

$$\Phi_{SCR}(y_i, y_j, x) = \exp \left(\sum_k \lambda_k f_k(y_i, y_j) \right) \quad (8)$$

$$\Phi_V(y_i, x) = \exp \left(\sum_{k'} \lambda'_{k'} f'_{k'}(y_i, x) \right) \quad (9)$$

ここで、 $k \in \{(y_i, y_j) | \mathcal{Y} \times \mathcal{Y}\}$ であり、これはラベル集合 $\mathcal{Y}$ の直積集合のある要素に対応する。つまり、 Φ_{SCR} は x には依存せず、単純にラベルの組み合わせのみによって決まる関数である。また、 $k' \in \{(y_i, x_j) | \mathcal{Y} \times \mathcal{X}\}$ であり、これは観測系列に出現する素性 x_j と、ラベル $y_i \in \mathcal{Y}$ の共起を捉えるための素性である。

このモデルは、基本的には Φ_V によって各ノードに付与されるべきラベルが何であるかを捉えた上で、 Φ_{SCR} によって各クリークの依存関係が考慮され、全体として最適となるようなラベル系列が選択されるような設計となっている。

パラメータ Λ の最適化をおこなうための条件付対数尤

表 1 訓練・評価用データの固有表現クラスと各クラスに属する Wikipedia の記事数一覧

固有表現クラス		SIZE
名前	イベント名 (EVENT)	121
	人名 (PERSON)	3315
	単位名 (UNIT)	15
	地名 (LOCATION)	1480
	施設名 (FACILITY)	2449
	称号名 (TITLE)	42
	組織名 (ORGANIZATION)	991
	職業名 (VOCATION)	303
	自然物名 (NATURAL_OBJECT)	1132
	製品名 (PRODUCT)	1664
	名前_その他 (NAME_OTHER)	24
時間・数値表現	時間・数値表現	2749
固有表現以外		1851
全体		16136

度関数は以下のように定義できる．

$$\begin{aligned}
 \mathcal{L}_\lambda = & \sum_{d=1}^N \left[\sum_{(v_i, v_j) \in E_S^{(d)}, E_C^{(d)}, E_R^{(d)}} \sum_k \lambda_k f_k(y_i^{(d)}, y_j^{(d)}) \right. \\
 & + \sum_{v_i \in V^{(d)}} \sum_{k'} \lambda_{k'} f_{k'}(y_i^{(d)}, \mathbf{x}^{(d)}) - \log Z(\mathbf{x}^{(d)}) \left. \right] \\
 & - \sum_k \frac{\lambda_k^2}{2\sigma^2} - \sum_{k'} \frac{\lambda_{k'}^2}{2\sigma^2} \quad (10)
 \end{aligned}$$

最後の 2 つの項はパラメータの正則化をおこなうための Gaussian prior によるものである．パラメータの最適化は L-BFGS などを用いておこなう．

4. 実験

本節では，前節で述べた手法を用いて，Wikipedia の記事から固有表現を獲得する評価実験の結果について報告する．

4.1 訓練・評価用データ

2005 年 10 月 29 日時点の Wikipedia の記事から，ランダムに選択した日本語の記事約 2300 内のリスト (タグ) に含まれるアンカー (<A>タグ) に対して，関根らによって提案された階層的な固有表現分類である拡張固有表現階層 [Sekine 02] version6.1.4 の対応するクラスを手で付与したデータを用いる．ここで，アンカーのリンク先の記事が存在しない場合についても，記事が存在する場合と同様にクラス付与をおこない分類対象とした．また，依存関係を考慮した分類の有効性を調査するため，Sibling, Cousin, Relative の関係にあるアンカーによって構成される連結グラフのノード数が 2 以下となるような事例については分類対象から除外した．その結果，全体で 16136 アンカーが分類対象となった．このうち，固有表現は 14285 である．データ中の Sibling, Cousin, Relative の各エッジの数はそれぞれ $|E_S| = 4925$, $|E_C| = 13134$, $|E_R| = 746$ である．

図 3 Sibling, Cousin, Relative を含めるか含めないかの各組合せによって構成されるグラフ．モデル名の S は Sibling クリーク, C は Cousin クリーク, R は Relative クリークが，構成されるグラフに含まれることを表している．

関根の拡張固有表現階層のクラス数は 200 以上であるが，本稿ではラベル数を限定し，拡張固有表現階層の深さ 2 までに限定したラベルを使用した．また，訓練・テストデータに存在する数が極端に少ない固有表現クラスについては，一つのクラスに統合した．その結果，訓練・評価用データに含まれるクラス数は，全体で 13 クラスとなった．拡張固有表現階層のクラスと各クラスに属する Wikipedia 記事数の対応を表 1 に示す．

4.2 実験方法

本稿における実験の目的は，1) 定義した 3 種のクリーク (Sibling, Cousin, Relative) それぞれを含むモデルと，含まないモデルとの精度比較をおこない，各クリークの分類精度への影響に関して調査すること，2) Support Vector Machines (SVMs) [Vapnik 98] をベースラインとして用い，CRFs との精度比較をおこなうことの 2 点ある．CRFs および SVMs による分類実験はそれぞれ以下に述べる方法を用いる．

§1 CRFs

CRFs のグラフを構成するために定義した 3 種類のクリークのうち，どのクリークが精度向上に寄与するかを調査するため，Sibling (S), Cousin (C), Relative (R) の各クリークを含むか含まないかの全組合せについて実験をおこなう (SCR 全て, SC, SR, CR, S のみ, C のみ, R のみ, 個別に分類 (I とした) の 8 通り)．各設定によって構成されるグラフの例を図 3 に示す．各クリークを除いた結果，グラフが全体で非連結となるものについては，部分連結グラフごとに分類をおこなう．本稿ではグラフの各部分連結グラフを事例とよぶ．

各モデルによって構成される事例の分布を表 2 に示す．SCR, SC, SR, CR の四つのモデルでは閉路を含む事例が存在するため，2 章にて述べた，周辺確率を近似的に計算する手法を適用する必要がある．

CRFs の学習および実行には，GRaphical Models in Mallet (GRMM) [Sutton 06] を用い，パラメータ推定および最適解の算出のための近似計算手法として TRP を，モデルの各素性に対する重みの事前分布には Gaussian Prior を用い，分散は $\sigma^2 = 10$ に設定した．

表 2 各モデルにおいて構築される事例の分布

モデル	SCR	SC	SR	CR
一つ以上の閉路を含む事例数	318 (36%)	324 (32%)	101 (1%)	42 (2%)
閉路を含まない事例数 (ノード数 2 以上)	555 (64%)	631 (62%)	2883 (27%)	1464 (54%)
ノード数 1 の事例数	0 (0%)	60 (6%)	7800 (72%)	1176 (44%)
全事例数	873	1015	10784	2682
1 事例あたりの平均ノード数	18.5	15.8	1.5	6.0
モデル	S	C	R	I
一つ以上の閉路を含む事例数	0 (0%)	0 (0%)	0 (0%)	0 (0%)
閉路を含まない事例数 (ノード数 2 以上)	2913 (26%)	1631 (54%)	237 (2%)	0 (0%)
ノード数 1 の事例数	8298 (74%)	1380 (46%)	15153 (98%)	16136 (100%)
全事例数	11211	3011	15390	16136
1 事例あたりの平均ノード数	1.4	5.4	1.05	1

表 3 分類に用いる素性.”✓” は該当する素性を与えたこと, CRFs の V はノードに対して, S は Sibling, C は Cousin, R は Relative クリークに対して素性を与えたことを意味している.

素性	SVMs	CRFs
記事の定義文 (bag-of-words)	✓	✓ (V)
記事の見出し (全体)	✓	✓ (V)
記事の見出し (形態素)	✓	✓ (V)
記事のカテゴリ (全体)	✓	✓ (V)
記事のカテゴリ (形態素)	✓	✓ (V)
リンク元のアンカーテキスト (全体)	✓	✓ (V)
リンク元のアンカーテキスト (形態素)	✓	✓ (V)
アンカーテキストの親タグ	✓	✓ (V)
アンカー直前のヘッダのテキスト (全体)	✓	✓ (V)
アンカー直前のヘッダのテキスト (形態素)	✓	✓ (V)
ラベル間素性		✓ (S, C, R)
前ラベル	✓	

§ 2 SVMs

SVMs による識別では, 各アンカー $a_j \in \mathcal{A} = \{a_1, a_2, \dots, a_N\}$ を個別に分類する手法 (以後 I) と, SCR によって構成されるノード集合に対応するアンカーのうち, 出現順序の早いものから順番に識別する手法 (以後 P) の 2 種類の手法を適用した. P ではアンカー a_j の識別に, アンカー a_{j-1} の識別結果を素性として用いた.

SVMs は 2 値分類器であるのに対して, 本稿で扱っている問題は多値分類問題である. そこで, 識別器を組み合わせて多値分類を行う one-vs-rest 法を用いて分類をおこなう. one-vs-rest 法は, $\mathcal{Z} = \{Z_1, Z_2, \dots, Z_n\}$ の多値分類問題を, Z_k に属するか, 属さないかという二値分類問題に分解して解く手法であり, SVMs を多値分類問題へ適用するためによく用いられる. SVMs の学習および実行には, TinySVM^{*2}を用い, 学習誤りとマージンとのトレードオフを設定するパラメータ C を, $C = \{0.01, 0.1, 1.0, 10.0\}$ の 4 種類の値に設定して実験をおこなった. また, カーネルは線形カーネルを用いた.

§ 3 分類に用いる素性

CRFs, SVMs で用いた素性を表 3 に示す. 形態素の素性を得るため, 日本語形態素解析器である MeCab^{*3} (version 0.96) を用いた.

4.3 評価

分類精度の評価は, Wikipedia データを 5 分割し, 訓練 4, テスト 1 の比率による交差検定によりおこなう. 交差検定のためのデータの分割は, SCR によって構成されるグラフの部分連結グラフ単位でおこない, 分類対象となるアンカー数が各データごとに, ある程度等しくなるよう分割をおこなった. 各固有表現クラスおよび全体の抽出精度の評価には精度と再現率の調和平均である F 値を用いた.

4.4 実験結果・考察

表 4 は, 4.2.1 節にて述べた CRFs による分類方法及び 4.2.2 節にて述べた SVMs による分類方法を Wikipedia データに対して適用した結果 (F 値) を示している. このうち, “全体” は全固有表現クラスでの F 値, “リンク先なし” は, 全体での分類結果のうち, アンカーのリンク先に文書が存在しないものに限定したデータでの F 値を示している. SVMs の実験結果については, SVMs(I), SVMs(P) 双方共に $C = 1.0$ の時に最も良い結果が得られたため, $C = 1.0$ の結果のみ掲載している.

§ 1 依存関係を考慮した分類と考慮しない分類での分類精度比較

表 4 の全てのアンカーを対象とした分類の結果, CRFs の Sibling, Cousin, Relative の各クリークを含めた場合と含めなかった場合の全ての組合せの中では, Cousin と Relative を含む場合 (CR) に最も高い F 値が得られ, 以降 SC, SCR, C, SR, S, R, I の順となっている.

依存関係を考慮した分類の統計的優位性を確認するために, CRFs のモデルの中で最も良い精度が得られた CR と依存関係を考慮しない I について検定をおこなった. 本稿では, 既存研究において分類対象に依存関係が存在するようなタスク (チャンキング [Sha 03] など) において, F 値の統計的優位性判定に用いられているマクネマー検定 [Gillick 89] を用いた. 具体的には, まず有意差の比較対象となる二つのモデル (m_1, m_2 とする) の出力結果から, (a) m_1 のみ正解した数, (b) m_2 のみ正解した数, の 2 項目を数える. 次に, 双方の誤り率が等しい, すなわち (a)=(b) であるという帰無仮説 H_0 をたて, テストする. 結果が棄却域 ($p < 0.01$) であれば有意差あり

*2 <http://www.chasen.org/~taku/software/TinySVM/>

*3 <http://mecab.sourceforge.net/>

表 4 CRFs と SVMs の分類精度 (全体, F 値) .

NE CLASS	N	CRFs								SVMs ($C = 1.0$)	
		C	CR	I	R	S	SC	SCR	SR	I	P
人名	3315	.7419	.7429	.7453	.7458	.7507	.7533	.7981	.7515	.7383	.7386
時間・数値表現	2749	.9936	.9944	.9940	.9936	.9938	.9931	.9933	.9940	.9933	.9935
施設名	2449	.8546	.8541	.8540	.8516	.8500	.8530	.8495	.8495	.8504	.8560
製品名	1664	.7414	.7540	.7164	.7208	.7130	.7371	.7418	.7187	.7154	.7135
地名	1480	.7265	.7239	.6989	.7048	.6974	.7210	.7232	.7033	.7022	.7132
自然物名	1132	.3333	.3422	.3476	.3513	.3547	.3294	.3304	.3316	.3670	.3326
組織名	991	.7122	.7160	.7100	.7073	.7122	.6961	.5580	.7109	.7141	.7180
職業名	303	.9088	.9050	.9075	.9059	.9150	.9122	.9100	.9186	.9091	.9069
イベント名	121	.2740	.2345	.2533	.2667	.2800	.2740	.2759	.2667	.3418	.3500
称号名	42	.1702	.0889	.2800	.2800	.3462	.2083	.1277	.3462	.2593	.2642
名前・その他	24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0690	.0000
単位名	15	.2353	.1250	.2353	.2353	.2353	.1250	.1250	.2353	.3333	.3158
全体	14285	.7846	.7862	.7806	.7814	.7817	.7856	.7854	.7823	.7790	.7798
リンク先なし	3898	.5476	.5495	.5249	.5274	.5272	.5484	.5465	.5224	.5278	.5386

という判定になる．検定をおこなった結果，統計的に有意であった ($p < 0.01$)．これは，適切な依存関係を選択し，その依存関係をモデルの構造に反映させるすることによってモデルのパフォーマンスが有意に改善することを示している．また，定義した 3 種類のクリークのうち，Cousin が最も精度向上に寄与している．Cousin クリークは Sibling, Relative クリークと比較して出現頻度が高く，またクリークを構成するアンカーは双方共に同じ固有表現クラスに属しやすいため，対応するパラメータ $\Lambda_s = \{\lambda_k | f_k(y_i, y_j), y_i = y_j\}$ が他の素性と比較して大きな値を持つように学習される．よって，これらの素性が Cousin クリークを構成する事例の分類に寄与したものと考えられる．

各固有表現クラス個別の分類精度に関して，まず人名については Sibling, Cousin を含むもの (SC, SCR) にて高い F 値が得られた．データ中の Sibling の関係にあるラベルの組では，双方のアンカーのラベルが人名である事例は 4925 中 435, Cousin の関係にあるラベルの組で双方のアンカーのラベルが人名である事例は 13125 中 2557 と，それぞれ他のラベルの組と比較して相対的に高い頻度で出現しており，結果，これらのラベル間素性に対応するパラメータの値が大きくなるように学習され，精度が向上したものと考えられる．製品名，地名については Cousin を含むモデル (C, CR, SC, SCR) にて高い F 値が得られた．Cousin の関係にあるラベルの組で双方のアンカーのラベルが製品名である事例は 1072, 地名である事例は 738 とそれぞれ相対的に高い頻度で出現していたため，良い分類精度をもたらしたものと考えられる．

リンク先に文書が存在しないアンカーに限定した場合，見出語，定義文，Wikipedia のカテゴリの情報が欠落しているため，全てのモデルにおいて精度が大幅に低下している．その中で，CRFs(CR) が約 2%, CRFs(I) と比較して精度が高くなっている．これは，他の手がかりが少ないことから，アンカー間の依存関係が分類をおこなう上で重要な手がかりとなっていることを示している．しか

し，CRFs(SCR) と CRFs(CR) を比べると CRFs(SCR) の精度が若干悪いという結果が得られた．表 2 のように，CRFs(SCR) では閉路を含む事例数が多く存在するが，閉路を含む事例については TRP による周辺確率の近似が必要となるため，この近似によってパフォーマンスが低下したものと考えられる．

§ 2 CRFs と SVMs での分類精度比較

CRFs, SVMs のモデルの中でそれぞれ最も良い精度が得られた CRFs(CR) と SVMs(P) についてマクネマー検定をおこなったところ，統計的に有意であった ($p < 0.01$)．これより，全体としては SVMs モデルと比較し CRFs のモデルが優れていると言える．

個別の固有表現クラスに関しては，人名，地名，製品名など出現する頻度の高い固有表現については CRFs が SVMs を上回る結果となっているが，イベント名，単位名など正例数の少ない固有表現については，CRFs と比較して SVMs が高い精度となる傾向がみられる．これは SVMs の汎化性能によるものと考えられる．

§ 3 エッジに対する素性の与え方に関する比較実験

本稿で提案しているモデルは，エッジに対するポテンシャル関数は単純にラベルの組み合わせのみを捉えるものであるが，エッジに対して特定の観測素性を与えるという方法 ([Ghamrawi 05] など) を考えることができる．すなわち，提案モデルのエッジに対応するポテンシャル関数 Φ_{SCR} が，

$$\Phi_{SCR} = \exp \left(\sum_k \lambda_k f_k(y_i, y_j, \mathbf{x}) \right)$$

で定義されるようなモデルである．

ここで，2 ノード間のラベルを推定するためにどのような観測素性 $x_l \in \mathbf{x}$ を選択すべきかは明らかではない．エッジに対して与える観測素性は，双方のアンカーの組の関係を捉えられるものが適していると考えられるが，本稿で用いている分類のための素性のうち，それぞれの記事に出現する単語，カテゴリなどは，対応するノードに対して与えるのが自然である．なぜなら，単純に 1 ノー

図 4 分類精度の精度 (Precision)-再現率 (Recall) 曲線 (CR model) . 左はリンク先ありの事例のみ, 右はリンク先なしの事例のみ .

ドから構成されるグラフを考えた場合, このような素性の与え方は, 一般的な文書分類タスクにおける素性の与え方と一致するからである. アンカーの組の関係を捉えるようなエッジ素性として考えられるものとしては, ‘アンカーの構造上の関係’ と, ‘記事の内容に関する関係’ の二つが考えられる. このうち, ‘アンカーの構造上の関係’ を捉えるような素性を本稿で提案しているモデルに導入しても, 分類には寄与しない. なぜなら, リスト構造内での同列の関係, 前後関係などは, 本稿で提案している Cousin, Sibling クリークにおいて必ず成り立つ関係であるため, 常にエッジの素性として出現するからである. 常に出現するエッジ素性は, ラベルの組み合わせを捉える素性 $f_k(y_i, y_j)$ と等価であるため, 新たに追加される情報は無い. ‘記事の内容に関する関係’ として考えられる素性としては, 例えば, クリークを構成する 2 文書間で共通に持つカテゴリが考えられる. Wikipedia 中のカテゴリを共有していれば同じ固有表現クラスに属しやすいと推測できるからである.

そこで, 記事の内容に関する関係が提案手法において寄与するか否かを検証するために, CRFs(CR) モデルにおいて, クリークを構成する 2 文書間で共通に持つカテゴリ素性 (全体, 形態素) を追加した場合とそうでない場合での比較実験をおこなった. 結果, カテゴリ素性を追加した場合に F 値 0.7797 と, 追加しない場合での F 値 0.7862 と比較して良い結果は得られなかった. その理由として, Wikipedia のカテゴリ素性自体はノードに対しても与えられており, 情報としては重複しているため, 精度向上に貢献しなかったものと考えられる. また, 固有表現クラスと直接関係の無い Wikipedia カテゴリの素性が分類に悪影響を与えた可能性がある.

アンカーの組に関する有効な素性を発見することで更なる精度向上の可能性はあるが, リスト構造の依存関係

を捉える上では単純に, 提案モデルで用いているようなラベルの組み合わせを捉える素性で十分であると思われる.

§ 4 周辺確率を用いた固有表現候補のフィルタリング

本稿の手法により抽出した固有表現候補から固有表現辞書を構築する場合, 少ない人手のコストでおこなえることが望ましい. そこで, CRFs の結果のうち一番良い結果が得られた CR の実験結果について, モデルが出力した各ラベル割り当ての周辺確率 $p(y_i|x)$ を特定の閾値 t でフィルタリングをおこなった. 図 4 は, CRFs(CR) の実験結果について, 各ラベル割り当ての周辺確率を閾値 t でフィルタリングした時の精度と再現率の関係をプロットしたものである. リンク先有りの事例のみの場合, 閾値 $t = 0.968$ の時, 再現率約 0.75, 精度はおおよそ 0.97 であった. これは, データ中に存在する固有表現全体のうちの $3/4$ を, 精度 97%, つまり 30 固有表現候補中誤りが 1 程度と非常に高い精度で獲得できることを示している. 一方, リンク先なしの事例に限定した場合, 分類の手がかりが少ないためにリンク先ありの場合と比較して全体的に再現率が低くなっている. 人手のコストを考慮すると高い精度での獲得が望ましいため, 手がかりの少ない事例について精度を維持した上で再現率をいかに向上させるかが今後の課題となる.

5. 関連研究

近年, Wikipedia を利用した研究が盛んにおこなわれており, 固有表現抽出 [Kazama 07], 固有表現の同定および曖昧性解消 [Bunescu 06, Cucerzan 07], 2 つの概念の意味的な関係の推定 [Strube 06], 共参照関係の同定 [Ponzetto 06] などのタスクにて利用されている.

CRFs などの無向グラフモデルを Web 文書からの情報抽出に用いた研究としては, Web 文書の階層性を考慮

したモデルである TCRFs [Tang 06], Web 文書からの製品情報の抽出のためのモデルである 2D Conditional Random Fields (2D CRFs) [Zhu 05], 階層構造をもつ条件付確率場である Hierarchical Conditional Random Fields (HCRFs) [Zhu 06], Dynamic Hierarchical Markov Random Fields (HMRFs) [Zhu 07] があるが, これらのモデルと本稿で提案した CRFs のモデルでは構造が異なる.

事例間の依存関係を考慮した確率モデルがこれまでにいくつか提案されてきている. これらは, Iterative Classification と, Collective Classification に大別される. Iterative Classification の手法としては, ネットワークのリンク構造を考慮したロジスティック回帰に基づく分類手法 [Lu 03b, Lu 03a] がある. この手法では, 局所的な分類を収束するまで繰り返すことで全体の分類結果を得るため, 最終的な分類結果が局所的な分類の順序に依存する点に問題がある. 一方, Collective Classification の手法としては, 関係を扱えるようにページアンネットワークを拡張した Probabilistic Relational Models (PRMs) [Getoor 01], 条件付マルコフネットワークをデータ間に依存関係のあるような問題へ適用するために拡張した Relational Markov Networks (RMNs) [Taskar 02] がある. RMNs は条件付きマルコフネットワークのクリークが, SQL に類似した式によって決定されるという点で, 条件付確率場の特殊型と解釈できる.

CRFs の周辺確率を利用した研究としては, 中国語のわかち書き問題で生じる未知語の同定に周辺確率を用いた研究 [Peng 04], 単語分かち書きにおいて, 周辺確率を用いて分割の曖昧性を保ったまま扱う手法 [工藤 05], 周辺確率を用いた確率的単語分割コーパスの構築 [岡野原 06], 文字単位の BI タグの周辺確率を利用し, 確率が閾値以上となる場合に単語の区切りとして決定することによって高速な形態素解析をおこなった研究 [岡野原 07] などがある.

6. お わ り に

本稿では, Web 上の多言語百科事典である Wikipedia から精度よく固有表現を獲得し分類する手法を提案した. 具体的には, リスト構造内のアンカー間の持つ依存関係を 3 種類のクリーク (Sibling, Cousin, Relative) として定義し, それらをグラフ構造に反映させることで, 依存関係を考慮した分類をおこなうことを可能にする CRFs のモデルを提案した. 評価実験にて, 汎化性能に優れた多くの分類問題に適用されている SVMs に対する提案手法と比較して高い精度で固有表現の分類が可能であることを示した. また, CRFs が出力した周辺確率を, ある特定の閾値でフィルタリングすることによって高い精度で固有表現を獲得できることを示した.

今後の課題としては, 付与する固有表現クラスの粒度

の細分化が挙げられる. 分類した固有表現を質問応答や関係抽出などの応用に用いることを考えた場合, 固有表現クラスの粒度の細分化は非常に重要である. 例えば, 質問応答システムへの応用を考えた場合, 固有表現が細分化されていることで回答候補を少数に絞ることが可能となり, 回答の精度向上が期待できる. しかし, 一般的に付与するラベル集合が大きい場合, 各ラベルごとの事例数は少なくなるため, 統計的手法による分類は難しくなる. この問題をどう解消するかは今後の課題である.

また, 本稿で使用した拡張固有表現階層のようなラベルの階層構造を分類に利用することで分類精度が向上するという報告があり [Rousu 06], このような手法を導入することでさらに精度よく固有表現を獲得できると考えられる.

◇ 参 考 文 献 ◇

- [Bunescu 06] Bunescu, R. and Pasca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation, in *Proc. of EACL* (2006)
- [Chen 99] Chen, S. F. and Rosenfeld, R.: A Gaussian Prior for Smoothing Maximum Entropy Models, Technical report, Carnegie Mellon University (1999)
- [Cucerzan 07] Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data, in *Proc. of EMNLP-CoNLL* (2007)
- [Getoor 01] Getoor, L., Segal, E., Taskar, B., and Koller, D.: Probabilistic Models of Text and Link Structure for Hypertext Classification, in *IJCAI Workshop on Text Learning: Beyond Supervision* (2001)
- [Ghamrawi 05] Ghamrawi, N. and McCallum, A.: Collective Multi-Label Classification, in *Proc. of CIKM* (2005)
- [Gillick 89] Gillick, L. and Cox, S.: Some Statistical Issues in the Comparison of Speech Recognition Algorithms, in *Proc. of ICASSP* (1989)
- [Kazama 07] Kazama, J. and Torisawa, K.: Exploiting Wikipedia as External Knowledge for Named Entity Recognition, in *Proc. of EMNLP-CoNLL* (2007)
- [Lafferty 01] Lafferty, J., McCallum, A., and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in *Proc. of ICML* (2001)
- [Liu 89] Liu, D. C. and Nocedal, J.: The Limited Memory BFGS Methods for Large Scale Optimization, in *Mathematical Programming* 45 (1989)
- [Lu 03a] Lu, Q. and Getoor, L.: Link-based Classification using Labeled and Unlabeled Data, in *Proc. of ICML* (2003)
- [Lu 03b] Lu, Q. and Getoor, L.: Link-based Text Classification, in *Proc. of IJCAI* (2003)
- [McCallum 03] McCallum, A., Rohanimanesh, K., and Sutton, C.: Dynamic Conditional Random Fields for Jointly Labeling Multiple Sequences, in *NIPS Workshop on Syntax, Semantics, and Statistics* (2003)
- [Murphy 99] Murphy, K., Weiss, Y., and Jordan, M.: Loopy Belief Propagation for Approximate Inference, in *Proc. of UAI* (1999)
- [Neville 00] Neville, J. and Jensen, D.: Iterative classification in relational data, in *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13–20. AAAI Press, 2000. (2000)
- [Pearl 88] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann (1988)
- [Peng 04] Peng, F., Feng, F., and McCallum, A.: Chinese Segmentation and New Word Detection using Conditional

- Random Fields, in *Proc. of COLING* (2004)
- [Ponzetto 06] Ponzetto, S. P. and Strube, M.: Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution, in *Proc. of HLT-NAACL* (2006)
- [Rousu 06] Rousu, J., Saunders, C., Szedmak, S., and Shawe-Taylor, J.: Kernel-Based Learning of Hierarchical Multi-label Classification Models, *Journal of Machine Learning Research*, Vol. 7, pp. 1601–1626 (2006)
- [Sekine 02] Sekine, S., Sudo, K., and Nobata, C.: Extended Named Entity Hierarchy, in *Proc. of LREC* (2002)
- [Sha 03] Sha, F. and Pereira, F.: Shallow Parsing with Conditional Random Fields, in *Proc. of HLT-NAACL* (2003)
- [Strube 06] Strube, M. and Ponzetto, S. P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia, in *Proc. of AAAI* (2006)
- [Sutton 04] Sutton, C., Rohanimanesh, K., and McCallum, A.: Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data, in *Proc. of ICML* (2004)
- [Sutton 06] Sutton, C.: GRMM: A Graphical Models Toolkit, <http://mallet.cs.umass.edu> (2006)
- [Sutton 07] Sutton, C. and McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning, in Getoor, L. and Taskar, B. eds., *Introduction to Statistical Relational Learning*, MIT Press (2007)
- [Tang 06] Tang, J., Hong, M., Li, J., and Liang, B.: Tree-structured Conditional Random Fields for Semantic Annotation, in *Proc. of ISWC* (2006)
- [Taskar 02] Taskar, B., Abbeel, P., and Koller, D.: Discriminative Probabilistic Models for Relational Data, in *Proc. of UAI* (2002)
- [Vapnik 98] Vapnik, V.: *Statistical Learning Theory*, Wiley Interscience (1998)
- [Wainwright 03] Wainwright, M., Jaakkola, T., and Willsky, A.: Tree-based Reparameterization Framework for Analysis of Sum-Product and Related Algorithms, *IEEE Transactions on Information Theory*, Vol. 45, No. 9, pp. 1120–1146 (2003)
- [Zhu 05] Zhu, J., Nie, Z., Wen, J.-R., Zhang, B., and Ma, W.-Y.: 2D Conditional Random Fields for Web Information Extraction, in *Proc. of ICML* (2005)
- [Zhu 06] Zhu, J., Nie, Z., Wen, J.-R., Zhang, B., and Ma, W.-Y.: Simultaneous Record Detection and Attribute Labeling in Web Data Extraction, in *Proc. of ACM SIGKDD* (2006)
- [Zhu 07] Zhu, J., Nie, Z., Zhang, B., and Wen, J.-R.: Dynamic Hierarchical Markov Random Fields and their Application to Web Data Extraction, in *Proc. of ICML* (2007)
- [岡野原 06] 岡野原 大輔, 工藤 拓, 森 信介: 形態素周辺確率を用いた確率的単語分割コーパスの構築とその応用, NLP 若手の会第 1 回シンポジウム (2006)
- [岡野原 07] 岡野原 大輔, 辻井 潤一: 大規模コーパスを用いた高速な形態素解析, NLP 若手の会第 2 回シンポジウム (2007)
- [工藤 05] 工藤 拓: 形態素周辺確率を用いた分かち書きの一般化とその応用, 言語処理学会全国大会 NLP-2005 (2005)

〔担当委員: 奥村 学〕

2007 年 12 月 13 日 受理

著者紹介

渡邊 陽太郎

2004 年信州大学工学部情報工学科卒業。2007 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了。現在、同大学院大学情報科学研究科博士後期課程在学中。自然言語処理の研究に従事。情報処理学会員。

浅原 正幸

1998 年京都大学総合人間学部基礎科学科卒業。2001 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了。同年より日本学術振興会特別研究員。2003 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。2004 年同大学助教、現在に至る。博士 (工学)。自然言語処理の研究に従事。

松本 裕治 (正会員)

1977 年京都大学工学部情報工学科卒。1979 年同大学院工学研究科修士課程情報工学専攻修了。同年電子技術総合研究所入所。1984～85 年英国インペリアルカレッジ客員研究員。1985～87 年 (財) 新世代コンピュータ技術開発機構に出向。京都大学助教授を経て、1993 年より奈良先端科学技術大学院大学教授、現在に至る。工学博士。専門は自然言語処理。情報処理学会、日本ソフトウェア科学会、言語処理学会、認知科学会、AAAI、ACL、ACM 各会員。