情報処理学会論文誌 Vol.57 No.8 1850–1860 (Aug. 2016)

語の出現と意味の対応の階層ベイズモデルによる教師なし語義曖昧性解消

谷垣 宏一[1,2,a)] 撫中 達司[3] 匂坂 芳典[4]

受付日 2015年11月11日, 採録日 2016年5月17日

概要：対象語を限定しない語義曖昧性解消（all-words WSD）のための新しい教師なし学習モデルを提案する．all-words WSD は，辞書知識を言語処理に活用する基礎技術として実用化が期待されるが，扱う語義の種類が膨大で，かつ分布がドメインに強く依存する性質があるため，ラベル付きコーパスの構築を前提とする教師あり学習では実用化を見込むことが難しい．提案法は，ラベルなしコーパスに出現する種々の語と膨大な語義の間に自然な対応を推定するため，2 つの制約をモデル化する．1) 同じ語の各出現における語義は，単語タイプごとの事前分布に従う．2) 類似した文脈に出現する種々の語の語義は，各語の語義割当てを平滑化して得られる分布に従う．これら 2 つの制約を階層モデルによって同時適用することで，教師なし all-words WSD を実現する．SemEval データセットを用いた実験結果より提案法の有効性を示す．

キーワード：語義曖昧性解消，all-words，教師なし学習，階層ベイズ，Gibbs サンプリング

Hierarchical Bayesian Mapping of Word Occurrences and Word Senses for Unsupervised Sense Disambiguation

Koichi Tanigaki[1,2,a)] Tatsuji Munaka[3] Yoshinori Sagisaka[4]

Received: November 11, 2015, Accepted: May 17, 2016

Abstract: This paper proposes a novel unsupervised model for all-words word sense disambiguation (WSD) to cope with the enormous number of sense classes inherent in the task. The proposed model is a hierarchical Bayesian model that incorporates two types of soft constraints and infers natural correspondence between unlabeled word occurrences and numerous senses: 1) senses of word instances follow the prior distribution of each word-type, 2) senses in a context follow the extrapolation from other words' senses in similar context. Experimental results applied to SemEval dataset confirmed the advantages of our hierarchical model.

Keywords: word sense disambiguation, all-words, unsupervised learning, hierarchical Bayes, Gibbs sampling

1. はじめに

語義曖昧性解消（Word Sense Disambiguation: WSD）とは，テキスト中の語が，辞書で規定されたいずれの語義で用いられているかを文脈に基づいて識別するタスクである．WSD タスクの中でも all-words タスクは，曖昧性解消の対象を特定の語に限定せず，文書中に出現するすべての語を対象に語義を識別するタスクであり[*1]，辞書知識を広く言語処理に活用するための基礎技術として実用化が期待される．all-words WSD はそのタスク設定から辞書に含まれるすべての語義を潜在的に識別の対象とし，膨大なクラ

1 三菱電機株式会社情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation, Kanagawa 247–8501, Japan
2 早稲田大学国際情報通信研究センター
Global Information and Telecommunication Institute, Waseda University, Shinjuku, Tokyo 169–8050, Japan
3 東海大学情報通信学部
School of Information and Telecommunication, Tokai University, Minato, Tokyo 108–8619, Japan
4 早稲田大学基幹理工学部
School of Fundamental Science and Engineering, Waseda University, Shinjuku, Tokyo 169–8555, Japan
a) Tanigaki.Koichi@ap.MitsubishiElectric.co.jp

[*1] たとえば評価型ワークショップ Senseval/SemEval の英語 all-words タスクでは，2〜3 の新聞記事，延べ 5,000 語程度が与えられ，記事中の内容語（辞書のエントリ）が対象語（対象トークン）として指定されることが多い．記事中の語は互いに意味的に関係しており，曖昧性解消結果の手がかりとして利用される [1].

スを扱う[*2]．また，語義の分布はたとえば品詞と比べてドメインに強く依存することが知られている [3], [4]．こうした理由から all-words WSD は高コストなラベル付きコーパスの構築を前提とすることが難しく，辞書知識を利用した教師なし方式がさかんに研究されてきた．

辞書知識を利用した教師なし all-words WSD の典型的な方法は，テキスト中で注目する語（曖昧性解消の対象語）から一定の範囲に出現している語を文脈語とし，それらの文脈語と対象語の語義候補の間で，語釈文中の語の重複率や，辞書階層中の語義の近さなどに基づく意味的類似度を計算して，最大スコアを与える語義を見つけるというものである [1], [3]．また，語義を対象語ごとに独立推定する代わりに，テキストの一定範囲に出現する語群を対象に，personalized PageRank や最適化の手法を適用して，各語の語義を同時推定する研究もある [5], [6]．こうした教師なし WSD の先行研究は，いずれもテキストの一定範囲に出現する語を文脈語として利用する．そのような直接的な文脈語が曖昧性解消の手がかりとして有効であることは明らかであるが，一方で，手がかりを文脈窓の中に限定することは WSD の限界を狭めると見ることもできる．

そこで本研究では，語の出現と意味の対応関係を一般化し，文脈窓に出現する直接的な文脈語の代わりに，大量入手が容易なラベルなしコーパス中で対象語と類似した特徴を持つ他の語を多数参照して，手がかりとして利用するアプローチをとる．これまでこのような一般化を狙った先行研究としては，Tanigaki らが文脈・語義それぞれの類似度（非類似度）で定まる 2 つの距離空間の間で，対応の密度を最大化することにより，ラベルなしコーパスの各語の語義を同時推定する方式を提案している [7]．これは，似た文脈に出現する語は似た意味を持つ傾向があるとの仮説 [8] に基づく一般化である．同様の一般化に基づく教師なし WSD の先行研究としてはこのほか文献 [9], [10] があり，ドメインが一致するデータセットでは直接的な文脈語を利用するよりも優れた性能を得られたことが報告されている [10]．

本論文では，このような文脈の類似性に基づく先行研究の一般化をさらに推し進め，単語タイプの一致に基づく一般化（大域的制約）と統合する，新しい教師なし all-words WSD のモデルを提案する．本論文の貢献は以下である．

- Tanigaki らのモデル [7] を拡張した新しいモデルを提案した．拡張点は，all-words WSD における大域的制約のモデル化，および，階層ベイズモデルとしての定式化，の 2 点である．前者の拡張では，トークン[*3]の語義に対する大域的制約として，単語タイプごとの語義確率分布を適用するモデルとした．これにより，文脈からは有効な手がかりが得られないトークンにおいても曖昧性解消が可能となり，ベースモデル [7] よりも安定した性能が得られる．また，後者の拡張により，上述の大域的制約をラベルなしコーパスから語義と同時推定可能とした．all-words WSD におけるベイズ推定は本論文が初めて試みるものである．
- 上述した大域的制約の導入（モデルの階層化）により曖昧性解消性能が有意に向上することを実験的に示した．
- 同時計算する語を増やすことで性能が向上する傾向があることを実験結果より示した．これは，大量入手が容易なラベルなしコーパスを用いて文脈と意味の対応関係を一般化する本研究のアプローチの有効性を示唆している．

2. 曖昧性解消の手がかり

語の意味には，次の性質があることが知られている．

(I) 単語タイプが同じトークンは（文脈が異なっても）同じ意味を表す傾向がある [11].

(II) 似た文脈に現れるトークンは（単語タイプが異なっても）似た意味を表す傾向がある [8].

これら 2 つの性質は，トークンの語義を推定するうえで，文脈と単語タイプ，双方の観点から相補的な制約を与える．本論文はこれらの性質を手がかりとして，ラベルなしコーパスで語義の曖昧性を解消する．先行研究においては，Yarowsky が “one sense per collocation” と “one sense per discourse” と呼ぶ 2 つの語義の性質を用いて，コーパスに複数回出現する単一の多義語の曖昧性を解消し，優れた性能が得られることを示した [12]．本論文が利用する性質 (I) はこの “one sense per discourse” に相当するが，(II) は “one sense per collocation” とはやや異なり，all-words WSD の手がかりとするために，単一の語の one sense ではなく種々の語の類似した語義を関連付けて扱うものである．

たとえば，“Exotic plants ...” や “Exotic trees ...”，“... of these plants ...” などの文を含むコーパスが与えられたとする（**図 1**）．最初の文に含まれる語 plant は，語義として *flora*（植物），*factory*（工場），*decoy*（桜客）を持ち，語義との対応に曖昧性がある．同様に 2 番目の文に含まれる語 tree も，*living-tree*（樹木）や *tree-diagram*（樹形図）の語義を持ち，曖昧性がある．ここでもし，語 plant がこのコーパスでは *flora* の語義で使われやすいことがあらかじめ分かっていれば，性質 I に基づいてこれらの文でも plant が *flora* の語義で使われていると解釈できる．このような事前知識は，特に最後の例文 “... of these plants ...” のように，曖昧性解消の有効な手がかりが文中に含まれない語に対し有効である．一方，最初の例文 “Exotic plants ...” に含まれるトークン plant は，同様に exotic（外来種の）の修飾を持つ，類似した文脈に出現するトーク

[*2] たとえば SemEval の WSD タスクで用いられてきた WordNet [2] の英語 3.1 版は，11 万種類の概念（synset）で語義を規定している．

[*3] 本論文では語の出現（インスタンス）をトークン，異なりをタイプと表記して区別する．

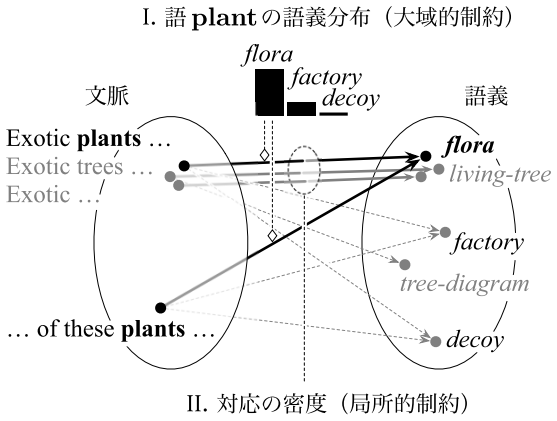


図 1 対応の曖昧性を解消するためラベルなしコーパスから推定する 2 つの統計量

Fig. 1 The two statistics to be estimated from unlabeled corpora for correspondence disambiguation.

ン tree が見つかること，さらに，plant の語義 *flora* と tree の語義 *living-tree* が比較的似た概念であることを考慮すれば，上述の性質 II に基づいてこれらのトークンの語義がそれぞれ *flora*，*living-tree* であると解釈できる．また，このような解釈を集めることで，上述の事前知識として用いたコーパス全体で使われやすい語義をある程度推定することもできる．

上述した簡単な例では，曖昧性解消の対象語に対し，I, II の性質をヒューリスティックな規則として逐次的に適用して曖昧性を解消した．しかし実際には，これら 2 つの性質が表している規則性は，いずれも他の語の解釈に互いに依存しているため，手続き的な方法で適用することはできない．また，両性質ともあくまで傾向にすぎないため，逸脱が許容される緩い制約として適用する方法が望ましい．そこで本論文は，データセット中のすべての語と語義の同時分布をモデル化することで単語間の依存性を扱う．このとき，上述した語義の性質は，同時分布に関する仮定として以下のように読み替えることで確率モデルとなる．

(I) 同じ語の各出現（トークン）における語義は，単語タイプごとの事前分布に従う．
(II) 類似した文脈に出現する種々の語の語義は，各語の語義割当てを平滑化して得られる分布に従う．

本論文では，これらの仮定を大域的制約・局所的制約として単一の階層ベイズモデルに統合する方式を提案する．2 つの制約を適用することで，ラベルなしコーパスに出現する種々の語と膨大な語義の間に，語義の性質に即した自然な対応を求めることができる．

3. 問題の定式化

具体的なモデル化の準備として，本章では教師なし all-words WSD の問題をベイズ推定の枠組みで定式化する．いま，WSD の対象トークン N 語からなる順序集合を $X = \{x_i\}_{i=1}^{N}$ とし，各トークンの語義候補集合からなる順序集合を $\mathcal{Y} = \{Y_i | Y_i = \{y_{ij}\}_{j=1}^{M_{w_i}}\}_{i=1}^{N}$ とする．ただし Y_i はトークン x_i の語義候補集合である．M_{w_i} は単語タイプ w_i の語義候補数，w_i は x_i の単語タイプであり，語彙 V の要素（$w_i \in V$）とする．また，任意のトークン $x_i, x_{i'} \in X$ の文脈距離（文脈の非類似度）が距離関数 $d_{\mathrm{x}}(x_i, x_{i'})$ によって定義され，同様に，任意の語義 $y_{ij}, y_{i'j'} \in \bigcup_{\iota=1}^{N} Y_\iota$ の語義距離（語義の意味的な非類似度）が距離関数 $d_{\mathrm{y}}(y_{ij}, y_{i'j'})$ で定義されるとする．これら X，$\mathcal{Y}$（および d_{x}，d_{y}）が与えられた下で，各トークン x_i の正しい語義 $y_{ij} \in Y_i$ を推定することを考える．なお，ここでは教師なし方式を考えるため，いずれの x_i に対しても正しい語義は与えられないものとする[*4]．

いま，トークン x_i の正しい語義が y_{ij} であるとする仮説を，変数 z_i を用いて $z_i = j$ と表すこととする．本論文における all-words WSD の目的は，データ X，$\mathcal{Y}$ が与えられたとき，それぞれの z_i（$i = 1, \ldots, N$）について最適解

$$j^*|_i \equiv \arg\max_j P(z_i = j | X, \mathcal{Y}) \tag{1}$$

を求めることである．

右辺の確率についてモデルを考えるとき，そのモデルパラメータを $\boldsymbol{\Theta}$ で表すこととする．また，X 全体の語義割当て仮説をベクトル $\boldsymbol{z} = (z_1, \ldots, z_N)$ で表す．このとき，式 (1) による条件付き確率の最大化の解は，次式 (2) に示す，周辺化した同時分布 $p(X, \mathcal{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ の最大化の解に一致する．

$$\begin{aligned} P(z_i = j | X, \mathcal{Y}) &= \sum_{\boldsymbol{z}: z_i = j} \int_{\boldsymbol{\Theta}} p(\boldsymbol{z}, \boldsymbol{\Theta} | X, \mathcal{Y})\, d\boldsymbol{\Theta} \\ &\propto \sum_{\boldsymbol{z}: z_i = j} \int_{\boldsymbol{\Theta}} p(X, \mathcal{Y}, \boldsymbol{z}, \boldsymbol{\Theta})\, d\boldsymbol{\Theta} \end{aligned} \tag{2}$$

ただし $\sum_{\boldsymbol{z}: z_i = j}$ は，$z_i = j$ となる $\boldsymbol{z}$ で求める総和を表す．

このように，あるトークンに関する z_i の解を求めるために，他のトークンの語義割当て $z_{i'}$（$i' \neq i$）やモデルパラメータ $\boldsymbol{\Theta}$ までも含めた同時分布 $p(X, \mathcal{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ を考えるのは，語の間の依存性を利用するためである．依存性をモデル化するため，モデルパラメータ $\boldsymbol{\Theta}$ には次章で述べるように元の WSD の問題には直接含まれない潜在変数を導入するが，最終的には式 (2) の周辺化によって潜在変数による条件付けは消去されるため，WSD として合目的的な解が得られる．

同時分布 $p(X, \mathcal{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ の推定にはサンプリング法を適用できる．変数を十分な回数 T 回ずつサンプリングしたとき，t 回目に得られた z_i のサンプルを $z_i^{(t)}$ とすると，式 (2) の周辺化は $z_i^{(1)}, \ldots, z_i^{(T)}$ の平均を用いて次式で近似できる．

[*4] （半）教師あり学習にする場合は，語義の正解が与えられる語 x_i について，後述する変数 z_i を正解に固定して計算すればよい．

表 1 本論文で用いる記号
Table 1 Major symbols used in this paper.

与えられるデータ X，$\mathfrak{Y}$ に関する記号	
X	トークンの順序集合．$X = \{x_i\}_{i=1}^{N}$．
$\mathfrak{Y}$	トークンの語義候補．$\mathfrak{Y} = \{Y_i\}_{i=1}^{N}$，$Y_i = \{y_{ij}\}_{j=1}^{M_{w_i}}$．
w_i	トークン x_i の単語タイプ．$w_i \in V$．V は語彙．
$d_{\mathrm{x}}(\cdot,\cdot)$	2 つのトークンの文脈距離を与える関数．
$d_{\mathrm{y}}(\cdot,\cdot)$	2 つの語義の意味距離を与える関数．
N，M_{w_i}	データセットのトークン数，w_i の語義候補数．
潜在変数 $\boldsymbol{z}$，$\boldsymbol{\Theta}$ に関する記号	
$\boldsymbol{z}$	トークンの語義割当てベクトル．
	$\boldsymbol{z} = (z_1, \ldots, z_N)$，$z_i \in \{1, \ldots, M_{w_i}\}$．
$\boldsymbol{\Theta}$	モデルパラメータ．$\boldsymbol{\Theta} = \langle \boldsymbol{\theta}_w, \boldsymbol{\theta}_{w'}, \ldots \rangle$．
$\boldsymbol{\theta}_w$	w の語義確率分布．$\boldsymbol{\theta}_w = (\theta_{w1}, \ldots, \theta_{wM_w})$．
σ_{x}，σ_{y}	密度推定の文脈，語義平滑化ハイパーパラメータ
α	語義確率分布のディリクレ平滑化ハイパーパラメータ

$$\sum_{\boldsymbol{z}:z_i=j} \int_{\boldsymbol{\Theta}} p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})\, d\boldsymbol{\Theta} \simeq \frac{1}{T} \sum_{t=1}^{T} \delta_{z_i^{(t)} j} \tag{3}$$

式中の δ はクロネッカーのデルタであり，$z_i^{(t)} = j$ のとき $\delta_{z_i^{(t)} j} = 1$，$z_i^{(t)} \neq j$ のとき $\delta_{z_i^{(t)} j} = 0$ である．このようにして，サンプルから各トークンの語義割当て z_i の最適解 $j^*|_i$ を決定できる．以下 4 章では，同時分布 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ のモデル化について述べ，5 章では同時分布からのサンプリングについて述べる．本章〜5 章で用いる主な記号の定義を**表 1** にまとめた．

4. 語の出現と意味の対応のモデル

本章では，同時分布 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ をモデル化する．提案法は階層ベイズモデルであり，同時分布を構成する 4 種類の変数の出現を $\boldsymbol{\Theta}$，$\boldsymbol{z}$，$(X, \mathfrak{Y})$ の順にモデルの階層から生成する．生成過程における変数の依存関係をグラフィカルモデルで**図 2** に示す．図の左側がモデルの上位階層に相当する．塗りつぶしの円は観測変数（ハイパーパラメータを含む）を表し，値が外部から与えられる．中抜きの円は潜在変数を表し，観測変数より分布を推定する．矢印は変数間の依存関係を，矩形は繰返しを表し，繰返しの回数を矩形内に N，M などで示す．ただし単語タイプ $w, w', \ldots$ に関する異なり語数 $|V|$ 回の繰返しだけは矩形で省略表記せず，添字として w，w' を明示して，異なり語の間で依存関係が交差することを示す．図に示した変数のうち，実際に求めたい変数はトークンの語義割当て z であり，$\boldsymbol{\theta}$ および $\langle x, y \rangle$ は z の推定に制約を与えるための変数である．

本階層モデルのモデルパラメータ $\boldsymbol{\Theta}$ は，各単語タイプ $w, w', \ldots$（$\in V$）に対応する語義確率分布の組 $\langle \boldsymbol{\theta}_w, \boldsymbol{\theta}_{w'}, \ldots \rangle$ である．語義確率分布 $\boldsymbol{\theta}_w = (\theta_{w1}, \theta_{w2}, \ldots)$ は，単語タイプ w のトークン x_i が語義 $y_{i1}, y_{i2}, \ldots$ で用いられる確率（事前確率）を表し，$\sum_j \theta_{wj} = 1$ である．モデルの最上位階層では，各単語タイプ w に応じた $\boldsymbol{\theta}_w$ が生成される．本論

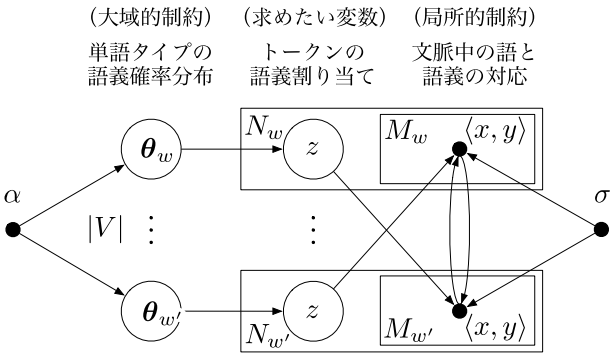


図 2 語の出現と意味の対応の確率的生成モデル
Fig. 2 Generative model of tokens and senses correspondence.

文では教師なし学習を考えるため，$\boldsymbol{\theta}_w$ が特定の語義に偏ることは事前に仮定せず，すべての語義候補が同様に出現しやすいとする．このとき，$\boldsymbol{\theta}_w$ の生起は次式の対称ディリクレ分布に従う．

$$\boldsymbol{\theta}_w \sim \mathrm{Dir}(\,\cdot\,|\,\alpha/M_w, \ldots, \alpha/M_w\,) \tag{4}$$

ここで記号 $\sim$ は左辺の生起が右辺の分布に従うことを表す．Dir() はディリクレ分布の確率密度関数を表し，M_w は w の語義候補数を表す．α はあらかじめ与える正の実数定数であり，一様分布の重みをコントロールするハイパーパラメータである．α に大きな値を設定するほど，得られる $\boldsymbol{\theta}_w$ は一様分布に偏る．

モデルの第 2 層は 2 章で述べた仮定 (I) を反映したものであり，単語タイプ w_i に固有の語義確率分布 $\boldsymbol{\theta}_{w_i}$ より，各トークン x_i の語義割当て z_i が生成される．z_i は，試行回数 1 回の多項分布（確率関数を Mul1() で表す）において，各要素が 1 となる確率の分布に従う．

$$z_i \sim \mathrm{Mul1}(\,\cdot\,|\,\boldsymbol{\theta}_{w_i}\,) \tag{5}$$

$\boldsymbol{\theta}_{w_i}$ は w_i の異なりに依存するが i 自体には依存しない．このため $\boldsymbol{\theta}_{w_i}$ は，単語タイプが等しいトークンの語義割当て $z_i, z_{i'}, \ldots$ に対し，共通の大域的制約として作用する．

モデルの第 3 層では，語義割当て z から文脈におけるトークンと語義の対応 $\langle x, y \rangle$ が生成される．このとき，ラベルなしコーパスを用いた学習では，z の正解が与えられない点に注意が必要である．このため，z_i から $\langle x_i, y_{ij} \rangle$ を直接生成するようなモデル化をしても，正しい割当てを学習できない．そこで，x_i とは単語タイプが異なるトークンの集合 $X_{\setminus w_i}$ とクロスバリデーションすることで，z から $\langle x, y \rangle$ の生成をモデル化する．すなわち，あるトークン x_i と語義の対応 $\langle x_i, y_{ij} \rangle$ は，トークン $x_{i'}, x_{i''}, \cdots \in X_{\setminus w_i}$ とそれらトークンの語義割当て $z_{w_{i'}}, z_{w_{i''}}, \ldots$ によって定まる対応 $\langle x_{i'}, y_{i'z_{i'}} \rangle, \langle x_{i''}, y_{i''z_{i''}} \rangle, \ldots$ より生成されるとしてモデル化する．このような単語タイプによるクロスバリデーションが有効な理由は，all-words WSD では語義の候補が単語タイプによって異なるからであり，このため互い

に類似した文脈に出現するトークンの間では，あるトークンの語義候補を，単語タイプが異なる他のトークンの語義に関する緩い教師情報（類似した語義ほど尤もらしい）として利用できるからである．単語タイプによるクロスバリデーションを，図 2 では上段と下段にまたがる依存関係の交差で示した．

この第 3 層は 2 章で述べた仮定 (II) を反映したものであり，カーネル密度推定 [13] を適用して得られる平滑な密度分布より，トークンと語義の対応 $\langle x, y\rangle$ が生成される．

$$\langle x_i, y_{ij}\rangle \sim \mathrm{Kdens}\big(\cdot,\cdot \,\big|\, X_{\setminus w_i}, \mathfrak{Y}_{\setminus w_i}, \boldsymbol{z}_{\setminus w_i}, \sigma_\mathrm{x}, \sigma_\mathrm{y}\big) \quad (6)$$

Kdens() は，$X_{\setminus w_i}$ に含まれる $N_{\setminus w_i}$ 個のトークンと語義の対応 $\langle x_{i'}, y_{i'z_{i'}}\rangle$ に基づいて $\langle x_i, y_{ij}\rangle$ を外挿する確率密度関数であり，$N_{\setminus w_i}$ 個のカーネルを用いた密度推定により，次式で定義する．

$$\begin{aligned}&\mathrm{Kdens}\big(x_i, y_{ij} \,\big|\, X_{\setminus w_i}, \mathfrak{Y}_{\setminus w_i}, \boldsymbol{z}_{\setminus w_i}, \sigma_\mathrm{x}, \sigma_\mathrm{y}\big) \quad (7)\\ &\quad\equiv \frac{1}{N_{\setminus w_i}} \sum_{i':w_{i'}\neq w_i} k\big(x_i, y_{ij}, x_{i'}, y_{i'z_{i'}} | \sigma_\mathrm{x}, \sigma_\mathrm{y}\big)\end{aligned}$$

カーネル関数としてガウスカーネルを用いると，上式の $k()$ はトークンの文脈距離 $d_\mathrm{x}(x_i, x_{i'})$ と語義距離 $d_\mathrm{y}(y_{ij}, y_{i'j'})$ を用いて次式で定義できる．式中，σ_x，σ_y はカーネルによる平滑化の強さを決定するハイパーパラメータである．

$$\begin{aligned}&k\big(x_i, y_{ij}, x_{i'}, y_{i'j'} | \sigma_\mathrm{x}, \sigma_\mathrm{y}\big) \quad (8)\\ &\quad\equiv \frac{1}{2\pi\sigma_\mathrm{x}\sigma_\mathrm{y}} \exp\left[-\frac{{d_\mathrm{x}}^2(x_i, x_{i'})}{2{\sigma_\mathrm{x}}^2} - \frac{{d_\mathrm{y}}^2(y_{ij}, y_{i'j'})}{2{\sigma_\mathrm{y}}^2}\right]\end{aligned}$$

この密度分布からは，他の語の語義割当て $\boldsymbol{z}_{\setminus w_i}$ が，文脈的にも意味的にも近傍[*5]に多数存在するほど，対応 $\langle x_i, y_{ij}\rangle$ が生成されやすくなる．いい換えれば，モデルの第 3 層では，求めようとする変数 z_i の推定に対し，近傍にある他の語と語義の対応が局所的制約として作用する．

本モデルで $\boldsymbol{\theta}_w$ としてモデル化する大域的制約は，先行研究においては “one sense per discourse” と呼ばれるヒューリスティクス [11] として手続き的な手法によって実現され，self-training で訓練データを拡張する際の補助的手段として用いられた [12]．本モデルでは，これを階層ベイズモデルの事前分布として扱う．このため，単語タイプに応じて強さがそれぞれ異なる大域的制約を，ラベルなしコーパスから同時推定して曖昧性解消に利用できる．階層ベイズモデルを語義に適用した先行研究としては，潜在的ディリクレ配分法（LDA）や階層ディリクレ過程（HDP）を用いた研究がある [14], [15], [16]．これらは WSD ではなく Word Sense Induction（WSI）を目的としたものであり，クラスタに相当する語義から文脈語を直接生成するモデルを用いてトークンをクラスタリングする．このようなモデル化は，対象とする語義の間で文脈語の分布がオーバラップしないことを前提としたものであり，語彙を限定した WSI に適している．これに対し提案法は，種々の語の多様な語義を識別する all-words WSD を目的としている．モデルの生成過程を異なり語で交差させることで，文脈に関する分布のオーバラップを教師情報として利用する点が特徴である．

[*5] 本論文で「近傍」ないし「近傍語」とは，距離関数 d_x，d_y で規定される距離が近いことを表す．特に断らない限り，実際のテキストや文書中で出現位置が近いことは意味しない．

5. サンプリングによる同時分布の推定

3 章では，同時分布 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ のサンプルが何らかの方法で得られたときに，それらサンプルから各トークンの最適な語義を決定する方法を述べた．本章では Gibbs サンプリング [17] を適用して，実際に分布のサンプルを得る方法を述べる．Gibbs サンプリングはマルコフ連鎖モンテカルロ法の 1 つであり，注目する変数のサンプリングと，サンプリングする変数の入れ替えとを繰り返して，求めたい分布からサンプルを得る方法である．ある変数のサンプルは，残りの変数をすべて固定したときの条件付き分布より得られる．提案する階層モデル（図 2）でサンプリングの対象となる変数とは，$\boldsymbol{z}$ と $\boldsymbol{\Theta}$ の各要素，すなわち，各トークンの語義割当て z_i（$i = 1, \ldots, N$）と，各単語タイプの語義確率分布 $\boldsymbol{\theta}_w$（$w \in V$）である．そこで以下では，z_i をサンプリングする条件付き分布 $P(z_i | X, \mathfrak{Y}, \boldsymbol{z}_{\setminus i}, \boldsymbol{\Theta})$ と，$\boldsymbol{\theta}_w$ をサンプリングする条件付き分布 $p(\boldsymbol{\theta}_w | X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta}_{\setminus \boldsymbol{\theta}_w})$ をそれぞれ導出する．

モデルの階層順に従って，まず，条件付き分布 $p(\boldsymbol{\theta}_w | X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta}_{\setminus \boldsymbol{\theta}_w})$ を導出する．本モデルで $\boldsymbol{\theta}_w$ は X，$\mathfrak{Y}$，$\boldsymbol{\Theta}_{\setminus \boldsymbol{\theta}_w}$ とは条件付き独立であり，$\boldsymbol{z}$ のうち，単語タイプが w に等しい z_i のみに依存する．いま $\boldsymbol{\theta}_w$ 以外の変数はすべて固定して考えるため，$\boldsymbol{\theta}_w$ の条件付き分布は，$\boldsymbol{\theta}_w$ 以外の変数との同時分布に比例することに注意すると，

$$\begin{aligned}&p\big(\boldsymbol{\theta}_w \,\big|\, X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta}_{\setminus \boldsymbol{\theta}_w}\big)\\ &\quad\propto p(\boldsymbol{\theta}_w) \prod_{i:w_i=w} P(z_i | \boldsymbol{\theta}_w) \quad (9)\\ &\quad\propto \mathrm{Dir}(\boldsymbol{\theta}_w \,|\, \alpha/M_w, \ldots, \alpha/M_w) \prod_{i:w_i=w} \mathrm{Mul1}(z_i \,|\, \boldsymbol{\theta}_w)\\ &\quad\propto \prod_j {\theta_{wj}}^{\alpha/M_w - 1} \prod_j {\theta_{wj}}^{\mathcal{N}_{wj}} \quad (10)\\ &\quad\propto \mathrm{Dir}\left(\boldsymbol{\theta}_w \,\middle|\, \frac{\alpha}{M_w} + \mathcal{N}_{w1}, \ldots, \frac{\alpha}{M_w} + \mathcal{N}_{wM_w}\right) \quad (11)\end{aligned}$$

と書ける．よって式 (11) のディリクレ分布より $\boldsymbol{\theta}_w$ のサンプルが得られる．なお，式 (9) 中の $p(\boldsymbol{\theta}_w)$，$p(z_i | \boldsymbol{\theta}_w)$ の展開は式 (4)，(5) で仮定したディリクレ分布（事前分布）と多項分布を適用したものである．以下，式 (11) のディリクレ分布までの変形は，LDA などでも用いられるディリクレ分布と多項分布の共役性を利用している．式 (10)，(11) 中の $\mathcal{N}_{wj}$ は単語タイプが w であるトークン $\{x_i | w_i = w\} \subseteq X$ のうち，語義割当て z_i が j となっているトークンの数を表

す．式 (10) の導出には以下の変形を用いた．

$$\prod_{i:w_i=w} \mathrm{Mul1}(z_i \,|\, \boldsymbol{\theta}_w) = \prod_{i:w_i=w} \theta_{wz_i} = \prod_{i:w_i=w} \prod_j \theta_{wj}{}^{\delta_{z_i j}} = \prod_j \theta_{wj}{}^{\mathcal{N}_{wj}}$$

次に，条件付き分布 $P(z_i|X, \mathfrak{Y}, \boldsymbol{z}_{\backslash i}, \boldsymbol{\Theta})$ を導出する．本モデルにおいて，トークン x_i の語義割当て z_i が依存する変数とは，単語タイプ w_i の語義確率分布 $\boldsymbol{\theta}_{w_i}$，および，$x_i$ とは単語タイプが異なるトークン $X_{\backslash w_i}$ とその語義候補 $\mathfrak{Y}_{\backslash w_i}$，語義割当て $\boldsymbol{z}_{\backslash w_i}$ である．式 (9) と同様に，依存変数との同時分布を用いると，求める条件付き分布は，

$$\begin{aligned} &P\big(z_i = j \,\big|\, X, \mathfrak{Y}, \boldsymbol{z}_{\backslash i}, \boldsymbol{\Theta}\big) \\ &\quad \propto \mathrm{Mul1}(\, j \,|\, \boldsymbol{\theta}_{w_i}) \prod_{i':w_{i'} \neq w_i} \mathrm{Kdens}\big(x_{i'}, y_{i'z_{i'}} \big| \\ &\qquad X_{\backslash w_{i'}}, \mathfrak{Y}_{\backslash w_{i'}}, \boldsymbol{z}_{\backslash w_{i'}}, \sigma_{\mathrm{x}}, \sigma_{\mathrm{y}}\big) \\ &\quad \propto \mathrm{Mul1}(\, j \,|\boldsymbol{\mu}_i) \end{aligned} \tag{12}$$

と書ける．よって式 (12) の多項分布より z_i のサンプルが得られる．ただし，式中の $\boldsymbol{\mu}_i$ は確率ベクトル $(\mu_{i1}, \ldots, \mu_{iM_{w_i}})$ であり，

$$\mu_{ij} \propto \theta_{w_i j} \prod_{i':w_{i'} \neq w_i} \mathrm{Kdens}\big(x_{i'}, y_{i'z_{i'}} \big| \\ X_{\backslash w_{i'}}, \mathfrak{Y}_{\backslash w_{i'}}, \boldsymbol{z}_{\backslash w_{i'}: z_i = j}, \sigma_{\mathrm{x}}, \sigma_{\mathrm{y}}\big) \tag{13}$$

とする．

6. 語義曖昧性解消実験

6.1 実験条件

モデル階層化の効果を評価するため，all-words WSD の実験を行った．本実験では，階層モデルの性能を 2 つの手法と比較する．比較対象の 1 つは，階層モデルから上位階層を取り除いた非階層モデルである．もう 1 つは，取り除いた上位階層の代わりに，“one sense per discourse” の手続き的実現法 [12] を単純化して適用する方法である．これらの方式の詳細を以下に示す．

- **階層モデル（提案法）：**本論文が提案する階層モデル（図 2）を用いて語義を推定する．このモデルでは，単語タイプが同じトークンにはできるだけ同じ語義が割り当てられるように，各トークンの語義割当てに対し，単語タイプごとの大域的制約がかかる．ただし，この制約は確率的で緩いため，別の語義の方がよりもっともらしいという手がかりが文脈の近傍語から十分得られれば，この制約に従わないこともできる．
- **非階層モデル：**上述の階層モデルから変数 α と $\boldsymbol{\theta}$ を取り除いたモデルを用いる．このモデルでは単語タイプで語義を一致させる制約がいっさい課されず，文脈近傍語の語義のみを手がかりとして語義が推定される．
- **事後拘束法：**上述の非階層モデルを用いて各トークンの語義を暫定的に推定した後，それらトークンの語義推定結果の最頻語義を単語タイプごとの代表語義として採択する．この代表語義で元の各トークンの語義推定結果を置き換えて最終的な出力とする．代表語義が一意に定まらない場合は複数の代表語義を等スコアで出力する．この方法は，単語タイプが同じトークンには必ず同じ語義が出力されるように厳格な制約を課すことになるため，文脈によって複数の語義で使い分けられている語には対応できない．

実験には，SemEval-2 英語 all-words WSD タスクのデータセット [18] を用いた．このデータセットは，評価型国際ワークショップ Senseval/SemEval が公開しているデータセットのうち，英語 all-words WSD 用の最新のものである．このタスクでは，テストセットのほかに，テストセットと同一ドメイン[*6]のラベルなしコーパスが提供されている．テストセットは 3 文書，5,348 語であり，うち 1,398 語（名詞 1,032 語，動詞 366 語）が WSD の対象語として指定されている．ラベルなしコーパスは 270 万語からなる（語数はいずれも延べ）．本実験では，ラベルなしコーパスは後述する文脈距離の実装で分布類似度を計算するためだけに用いた．配布されるデータセットには辞書引き（語義候補の取得）に必要な単語の品詞と基本形が含まれていないため，RASP parser [19] を適用して品詞と基本形を得た．辞書はタスク規定の WordNet 3.0 を用いた．

ところで，比較する 3 つの手法で実際に性能差が観測されるかは，テストセットの語義に，単語タイプによる関連付けを利用して解消される曖昧性がどの程度あるかに依存する．そこで事前分析としてテストセット 1,398 語の曖昧性を調べた．曖昧性の尺度としては，正解語義のパープレキシティおよび候補語義のパープレキシティを用いる[*7]．正解語義のパープレキシティは，1 つの単語タイプについて平均何種類の語義がテストセット中で正解としてラベル付けされているか（文脈によって使い分けられているか）を表す．一方，候補語義のパープレキシティは，辞書で規定されている語義の候補が平均何種類あるかを表す．パープレキシティの分析結果を**図 3** に示す．2 色に色分けした

[*6] SemEval-2 の all-words WSD はドメイン適応をテーマに実施され，単一ドメイン（環境ドメイン）のデータで性能が評価された [18].

[*7] 語義のパープレキシティは，単語タイプが与えられたときの語義のエントロピーによって $\mathrm{PP} = 2^{H(Y|W)}$ で定義する．ただし Y，W は語義および単語タイプを表す確率変数とする．エントロピーは $H(Y|W) = -\sum_{i,j} P(W = w_i)P(Y = y_{ij}|W = w_i) \log_2 P(Y = y_{ij}|W = w_i)$ で与えられる．
ここで，$P(Y = y_{ij}|W = w_i)$ は単語タイプが与えられたときの語義の確率であり，正解語義のパープレキシティを計算する際には，テストセット全体でその単語タイプのインスタンス（トークン）に付与されている正解語義の比率を用いる．これに対し，候補語義のパープレキシティを計算する際には，$P(Y = y_{ij}|W = w_i)$ として候補語義の数の逆数 $1/M_{w_i}$ を一律に用いる．

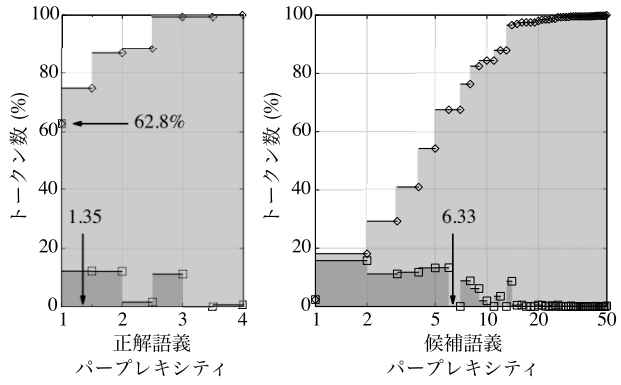


図 3 SemEval-2 テストセットにおける語義の曖昧性：正解語義パープレキシティ（左図）は最小値 1 の点（全インスタンスで正解語義が同じ）に過半の 62.8%が集中するが，パープレキシティが 1 より大きいトークンの数と区別するため．ヒストグラムは左開右閉の半開区間 $\{x \mid a < x \leq b\}$ の度数でプロットする．このため左端の集中点には柱が描画されない

Fig. 3 Sense ambiguity of the SemEval-2 test set. 68% of tokens is concentrated at the minimum of proper sense perplexity, where bins $\{x \mid a < x \leq b\}$ are not drawn. This is because we intend to distinguish the frequency of tokens whose perplexity equals to one and those whose perplexity is greater than one.

ヒストグラムのうち，濃い網掛けは区間ごとの頻度分布を表し，薄い網掛けは累積頻度分布を表す．ヒストグラムの各階級は右端を含み左端を含まない．図中下向きの矢印でテストセット全体のパープレキシティを示した．図 3 左のグラフから，正解語義パープレキシティは最小値 1 の点（全インスタンスの正解語義が同じ）に過半の 62.8%が集まることが分かる．ただしこの左端位置は柱としては描画されない．正解語義のパープレキシティはテストセット全体で 1.35 であり，候補語義のパープレキシティ 6.33 と比べて 1/4〜1/5 に絞られる．したがって同じ語の語義をできるだけ一致させようとする制約は本データセットでも有効に作用する可能性があり，階層モデルおよび事後拘束法では，非階層モデルと比べて優れた性能を期待できる．ただし，図 3 左のグラフの左端位置に左向き矢印で示すように，正解語義のパープレキシティが 1，すなわちテストセット中で同じ語のすべてのインスタンスに同じ語義が割り当てられるトークンは全体の 62.8%にとどまる．残る 4 割弱のトークンでは語義が文脈に応じて使い分けられていることから，そうした使い分けを識別できない事後拘束法と比べて，階層モデル（提案法）では優れた性能を期待できる．

提案法および 2 つの比較手法はいずれもデータセットと語義候補に加えて，文脈の距離関数 $d_{\mathrm{x}}(\cdot,\cdot)$ と語義の距離関数 $d_{\mathrm{y}}(\cdot,\cdot)$ が与えられることを仮定している．実験では，これらの距離関数を以下のように実装して用いた．まず，文脈距離は，構文的依存関係を分布類似度で平滑化して単語ベクトルを構成し，ベクトルの余弦を類似度とする Thater らの方法 [20] をベースとして用いた．ただし，類似度は余弦ではなく内積を用いることで，まったく異なる単語ペア間で類似度を比較する際に，一致する特徴が多いペアほど類似度が高くなるようにした．また，距離関数とするため類似度の逆数をとり，さらに特徴の数が増えたときの感度を抑えるため対数をとって用いた．サンプリングの計算効率化と，ノイズの影響を低減するため，文脈距離は一方のトークンが他方の k 最近傍である場合のみ計算に用いた．k の値は，予備実験の結果より $k > 10$ としても正解率が大きく変化しないことが分かっており，本実験では $k = 10$ とした．一方，語義距離には，先行研究 [7], [9] で WSD に適用し優れた性能が報告されている Jiang & Conrath の距離 [21]（WordNet 階層のエントロピーに基づいて定義する距離）を用いた．実際の計算には Pedersen らの提供するライブラリ [22] を利用した．これらの文脈距離と語義距離は，いずれも品詞が同じ語の間でのみ定義されるため，WSD の計算はテストセットの名詞と動詞に分けて別々に行った[*8]．

曖昧性解消結果の評価には SemEval が提供する評価ツールを用いた．このツールによって算出される評価スコアは再現率と適合率であるが，再現率・適合率の定義は，情報検索などにおける一般的な定義とは異なり，個々のターゲットについて出力の総和が 1 となるように正規化した確率カウントが用いられるため，再現率は候補を複数出力しても有利にはならない．そこで，以下では基本的に適合率の表記を省略し，タスクで正式なランキング基準にもなっている再現率のみを「正解率」と表記して示す．なお，本実験ではすべての対象語に対して必ず語義を出力するため，適合率もこの正解率に一致する．Gibbs サンプリングの回数は，すべての条件で正解率にほぼ収束が見られた 10 万回の場合を一律に示す．burn-in 期間は設定せず，得られた全サンプルを用いた．ハイパーパラメータ α は $0.001 \leq \alpha \leq 1$ の範囲で $\alpha = 1, 2, \ldots, 9 \times 10^n$（$n$ は整数）の場合を評価した．σ_{x}，σ_{y} は固定とし，データセット中の文脈距離，語義距離の平均 2 乗距離をそれぞれ設定した．

6.2 実験結果

実験結果を**図 4** に示す．折れ線グラフは階層モデルの正解率を，ハイパーパラメータ α を変化させてプロットした

[*8] 別の距離を実装する場合であれば，品詞ごとに計算を分ける必要はない．たとえば，注目する語の前後一定範囲に出現する語で構成する 1 次文脈ベクトルや，これを大規模コーパスにおける語の共起で展開して得られる 2 次文脈ベクトル [23]，あるいは，各語の分散表現を結合して構成されるパラグラフベクトル [24] などの類似度は，品詞が異なる語の間でも計算可能であり，本論文の文脈距離として利用することもできる．また，意味的類似度の導出に利用する概念辞書が品詞ごとの体系になっているため Jian & Conrath の類似度では不可能だが，語釈文の類似度で計算する Lesk 類似度であれば，品詞が異なる語の語義間でも計算できる．これらの文脈距離と意味距離を用いることで，本実験においても名詞と動詞の語の曖昧性を同時に解消することが可能であり，少なくとも一定の名詞・動詞の関係を効果的に利用できる可能性がある．

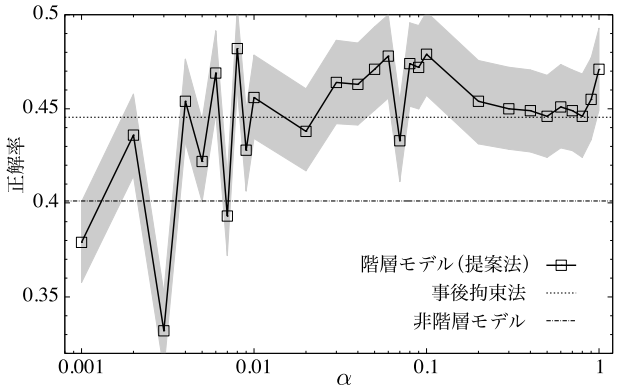


図 4 モデル階層化の効果

Fig. 4 Performances of the proposed hierarchical model compared with non-hierarchical models.

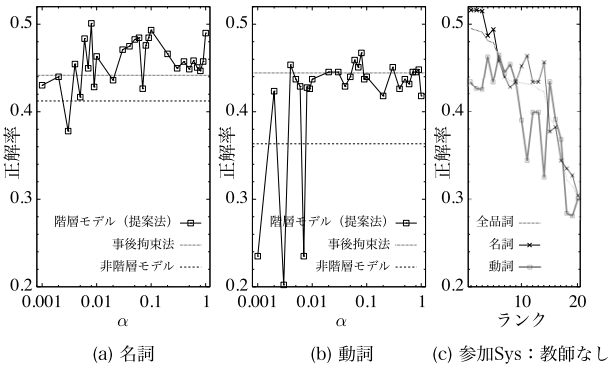


図 5 品詞による曖昧性解消性能の違い

Fig. 5 WSD performances depending on part-of-speech.

ものであり，網掛けの領域は 95%信頼区間を表している．階層モデルと条件を揃えるため，パラメータ α を持たない非階層モデルおよび事後拘束法については，サンプリングに使う乱数の種を変えることで階層モデルと同様に 28 回の実験を行い，得られた正解率の中央値をそれぞれ一点破線と点線で示した．

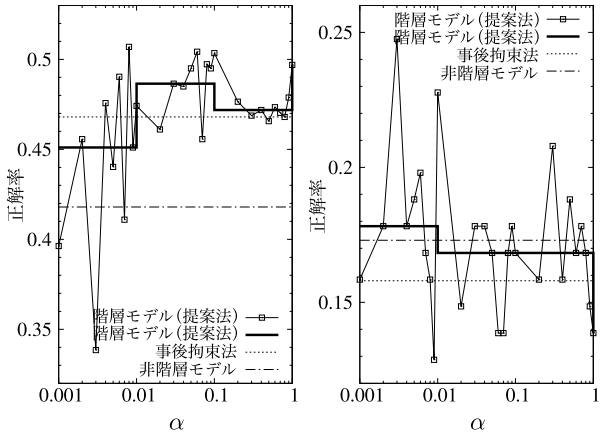
図 4 から，3 つの方式を比べると全般に，階層モデル，事後拘束法，非階層モデルの順で良い性能が得られたことが分かる．階層モデルでは，極端に α を小さく設定したときこそ性能のばらつきが大きくなる傾向が見られたものの，$\alpha \geq 0.01$ では安定して他の 2 つの方式を上回った．特に非階層モデルに対しては $\alpha \geq 0.01$ でつねに有意な性能の向上が得られた．非階層モデルと比べて，提案法（階層モデル）および事後拘束法で良い性能が得られたことから，後者で利用した単語タイプごとの語義制約が，本データセットのように単一ドメインの all-words WSD に対して有効であることが確認できた．また，提案法（階層モデル）では事後拘束法よりも良い性能が得られたことから，単語タイプごとの語義制約を実現する方法として，提案法による事前分布としてのモデル化が効果的であったといえる．

本実験において，階層モデルによる正解率の最大値は 0.482（$\alpha = 0.08$）であった．ただし上述のように $\alpha < 0.01$ における性能は不安定であるため，比較的安定したピークの性能は $0.05 \leq \alpha \leq 0.1$ における 0.47〜0.48 と見ることができる．なお，本テストセット中の最頻語義による正解率は 0.903，辞書の第 1 語義による正解率は 0.485 であり（いずれも教師ありベースライン），階層モデルはこれを下回る．最頻語義による正解率の高さは，単一ドメインからなる本テストセットの語義の偏りを反映している．ところで，SemEval-2 タスクに参加したシステムのうち，教師なし方式で最良のシステム [25] の正解率は 0.495 である [18]．実験条件が異なるため単純な比較はできないが，本実験結果はこれを下回った．SemEval-2 タスクはドメイン適応をテーマに実施されたため，上位システムでは，本実験のように与えられたリソース（データセットと辞書）をそのまま用いるのではなく，各種ドメイン適応手法を適用しており，辞書をあらかじめドメインに合わせてプルーニングしたり [25]，Web を検索して大量に取得した対象ドメインのテキストを利用したりする [26] などしている．特に後者のように，ラベルなしコーパスを大量に利用することで提案法の性能が改善される可能性については，次章で考察する．

品詞別の正解率を**図 5** (a)，(b) に示す．階層モデルと非階層モデルでは，動詞と比べて名詞で高い正解率が得られた．これに対し，事後拘束法では名詞・動詞ともほぼ変わらない性能が得られた．参考のため，SemEval-2 の all-words タスクに参加した教師なしシステム（全 20 エントリ）の品詞別正解率 [18] を転記して図 5 (c) に示す．これらのシステムでも，動詞と比べて名詞で正解率が高くなる傾向が観察される．よって，この結果（名詞の方が正解率が高くなる）は提案法に依存するものではなく，本テストセットと辞書を用いたときの一般的な傾向といえる．

テストセットのトークン 1,398 語のうち，テストセット中で最頻の正解語義がそのトークンでも正解となっているケースは 1,297 語（うち 63 語には最頻語義が 2 つ以上ある）であり，残り 101 語は，最頻語義以外の語義（以下，マイナー語義と呼ぶ）が正解のトークンであった．それぞれの語群に対する正解率を**図 6** に示す．階層モデル，非階層モデル，事後拘束法の性能をそれぞれ実線，点線，一点破線で示す．性能が横軸 α に依存する階層モデルのプロットは，各 α での性能をプロットした折れ線グラフと，区間ごとの中央値を太線で示す．これに対し，性能が α に依存しない非階層モデルおよび事後拘束法の性能は，全 28 回の実験結果の中央値を示す．

正解が最頻語義のトークンに対する性能は，もともとそれらのトークンが本テストセットでは全体の 9 割強を占めているため，傾向としてもテストセット全体の傾向（図 4）とほぼ一致し，$\alpha \geq 0.01$ で提案法，事後拘束法，非階層モデルの順に良い性能が得られる（図 6 (a)）．一方，正解がマイナー語義のトークンに対する結果では順位が入れ替わ



(a) 正解が最頻語義のトークン (b) 正解がマイナー語義のトークン

図 6 テストセット中の最頻語義/マイナー語義に対する正解率
Fig. 6 Performances on most/non-most frequent word senses in the data set.

り，階層モデル（提案法）と非階層モデルがほぼ同等の性能であったのに対し，事後拘束法はそれらより低い性能となった（図 6 (b)）．正解がマイナー語義のトークンはデータ量が少ないため有意な差はないものの，提案法は事後拘束法と比べて，マイナー語義の曖昧性解消性能を大きく損なわずに，過半を占める最頻語義に対する性能を改善できることを示唆する結果が得られた．

7. 考察

本章では，提案法を適用するラベルなしデータセットのサイズと曖昧性解消性能の関係について考察する．本論文の目的は 1 章で述べたように，語の出現と意味の関係を一般化することであり，それによって，手がかりが限定される直接的な文脈語に代えて，大量に入手可能なラベルなしコーパスの語を曖昧性解消に利用することである．本論文では文脈に依存しない大域的制約を導入する方法を提案し，その有効性を前章で示した．そこで本章では元の目的に立ち返って，ラベルなしデータの拡充により曖昧性解消性能が改善される可能性を検証する．データを大量に利用するほど類似した文脈に出現する語が得やすくなり，曖昧性の解消性能が向上すると期待される．

そこで，データセットのサイズを変えて曖昧性解消実験を行い，提案法（階層モデル）の正解率の変化を調べた．実験に用いるデータセットは前章と同様 SemEval-2 のテストセットであるが，本章では語の並びを変えずにテストセット全体を N 個の連続する区間に等分して（$N = 1, 2, \ldots, 10$），計算対象の語数が $1/N$ のサイズとなったデータセットでそれぞれ曖昧性を解消した後，テストセット全体で集計して正解率を求めた．α の値は 0.01，0.1，1 の 3 種類の場合で実験した．

実験結果を**図 7** に示す．このグラフから，横軸のデータセット・サイズに対する正解率は，局所的には上下の変動

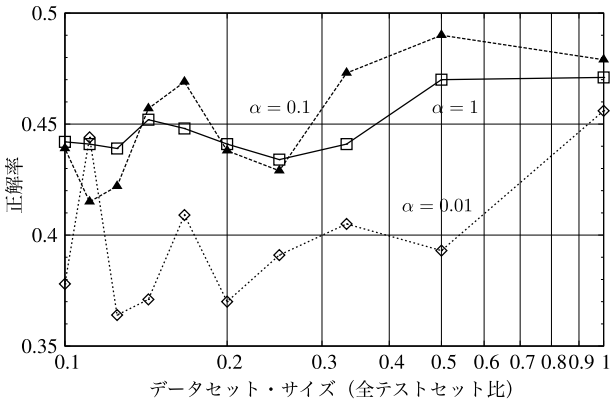


図 7 ラベルなしデータセットの大きさによる性能変化
Fig. 7 Performance changes to the size of unlabeled data set.

があるものの，全体的には α の大小にかかわらず右上がりの傾向にあると見ることができる．すなわち，本論文が目的とする，大量に入手可能なラベルなしコーパスの語を曖昧性解消に利用することで，性能の向上を実現できることを示唆する結果が得られた．また，今回用いたテストセットに対しても，ラベルなしコーパスから語を追加して提案法を適用することで，テストセットの正解率が改善される可能性がある．

8. おわりに

all-words WSD のための教師なし学習モデルを提案した．提案法は，ラベルなしコーパスの語と膨大な語義の間に自然な対応を推定するため，1) 同じ語のトークンで語義が一致しやすい性質（大域的制約），および，2) 文脈が類似した語の間で語義が類似しやすい性質（局所的制約），を扱い，単一の階層ベイズモデルとして統合する．SemEval-2 データセットを用いた実験では，階層化により曖昧性解消性能が有意に改善され，提案する階層モデルの有効性が確認された．また，適用するデータセットのサイズを拡大することで性能が改善される傾向が見られることを報告し，提案法の，低コストなラベルなしコーパスを利用する教師なし all-words WSD 方式としての可能性を示唆した．

今後，実際に大規模なラベルなしコーパスに提案法を適用して性能を評価する必要がある．また，本論文では，Thater らの構文的依存関係に基づく類似度 [20] をテストセットとともに提供されたコーパスで計算して文脈距離のベースとしたが，提案法の性能は文脈距離および意味距離に依存する．このため，今後，他の距離関数を用いたときの性能についても評価する必要がある．

本論文では，英語の単一ドメインで構成されるテストコレクションに提案法を適用して有効性を議論した．英語以外への適用に関しては，もともと提案法でモデル化している 2 つの性質（大域的制約/局所的制約）が，日本語を含む他の言語においても一般に期待できるものであることから，他の言語や辞書に対しても，距離関数を本論文と同様

に実装して適用可能と見ることができ，同様の効果を期待できる．一方，ドメインが複数混在するテストコレクションへの適用に関しては，大域的制約の仮定はむしろ性能の劣化を招く恐れがある．したがって，混合ドメインに適用する場合には，あらかじめクラスタリングなどを適用してデータセットを分離するか，あるいは，トピックモデルのようにモデルパラメータの方を分離するしくみを導入する必要がある．これら，他言語や混合ドメインのテストコレクションへの適用は今後の課題である．

参考文献

[1] Navigli, R.: Word sense disambiguation: A survey, *ACM Computing Surveys* (*CSUR*), Vol.41, No.2, p.10 (2009).
[2] Miller, G.A.: WordNet: A lexical database for English, *Comm. ACM*, Vol.38, No.11, pp.39–41 (1995).
[3] Agirre, E. and Edmonds, P.: *Word sense disambiguation: Algorithms and applications*, Vol.33 (2006).
[4] Resnik, P. and Yarowsky, D.: A perspective on word sense disambiguation methods and their evaluation, *Proc. ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pp.79–86 (1997).
[5] Agirre, E. and Soroa, A.: Personalizing PageRank for word sense disambiguation, *Proc. 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp.33–41 (2009).
[6] Reddy, S. and Inumella, A.: WSD as a distributed constraint optimization problem, *Proc. ACL 2010 Student Research Workshop*, pp.13–18 (2010).
[7] Tanigaki, K., Shiba, M., Munaka, T. and Sagisaka, Y.: Density maximization in context-sense metric space for all-words WSD, *Proc. 51st Annual Meeting of the Association for Computational Linguistics*, Vol.1, pp.884–893 (2013).
[8] Harris, Z.S.: Distributional structure, *Word* (1954).
[9] McCarthy, D., Koeling, R., Weeds, J. and Carroll, J.: Unsupervised acquisition of predominant word senses, *Computational Linguistics*, Vol.33, No.4, pp.553–590 (2007).
[10] Agirre, E., De Lacalle, O.L., Soroa, A. and Fakultatea, I.: Knowledge-based WSD on specific domains: Performing better than generic supervised WSD, *Proc. 21st International Jont Conference on Artifical Intelligence*, pp.1501–1506 (2009).
[11] Gale, W.A., Church, K.W. and Yarowsky, D.: One sense per discourse, *Proc. Workshop on Speech and Natural Language*, pp.233–237 (1992).
[12] Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods, *Proc. 33rd Annual Meeting on Association for Computational Linguistics*, pp.189–196 (1995).
[13] Parzen, E.: On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, Vol.33, No.3, pp.1065–1076 (1962).
[14] Lau, J.H., Cook, P., McCarthy, D., Newman, D. and Baldwin, T.: Word sense induction for novel sense detection, *Proc. 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp.591–601 (2012).
[15] Yao, X. and Van Durme, B.: Nonparametric Bayesian word sense induction, *Proc. TextGraphs-6: Graph-based Methods for Natural Language Processing*, pp.10–14 (2011).
[16] Brody, S. and Lapata, M.: Bayesian word sense induction, *Proc. 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp.103–111 (2009).
[17] Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Analysis and Machine Intelligence*, No.6, pp.721–741 (1984).
[18] Agirre, E., de Lacalle, O.L., Fellbaum, C., Hsieh, S.-K., Tesconi, M., Monachini, M., Vossen, P. and Segers, R.: Semeval-2010 task 17: All-words word sense disambiguation on a specific domain, *Proc. 5th International Workshop on Semantic Evaluation*, pp.75–80 (2010).
[19] Briscoe, T., Carroll, J. and Watson, R.: The second release of the RASP system, *Proc. COLING/ACL on Interactive Presentation Sessions*, pp.77–80 (2006).
[20] Thater, S., Fürstenau, H. and Pinkal, M.: Word Meaning in Context: A Simple and Effective Vector Model, *Proc. 5th International Joint Conference on Natural Language Processing* (*IJCNLP*), pp.1134–1143 (2011).
[21] Jiang, J.J. and Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy, arXiv preprint cmp-lg/9709008 (1997).
[22] Pedersen, T., Patwardhan, S. and Michelizzi, J.: WordNet::Similarity: Measuring the relatedness of concepts, *Demonstration Papers at HLT-NAACL 2004*, pp.38–41 (2004).
[23] Purandare, A. and Pedersen, T.: Word sense discrimination by clustering contexts in vector and similarity spaces, *Proc. Conference on Computational Natural Language Learning*, pp.41–48 (2004).
[24] Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents, *Proc. 31st International Conference on Machine Learning* (*ICML-14*), pp.1188–1196 (2014).
[25] Kulkarni, A., Khapra, M.M., Sohoney, S. and Bhattacharyya, P.: CFILT: Resource Conscious Approaches for All-Words Domain Specific, *Proc. 5th International Workshop on Semantic Evaluation*, pp.421–426 (2010).
[26] Tran, A., Bowes, C., Brown, D., Chen, P., Choly, M. and Ding, W.: TreeMatch: A Fully Unsupervised WSD System Using Dependency Knowledge on a Specific Domain, *Proc. 5th International Workshop on Semantic Evaluation*, pp.396–401 (2010).

谷垣 宏一 （正会員）

1995 年東北大学大学院情報工学科博士前期課程修了．同年三菱電機株式会社入社．1997 年～2000 年 ATR 音声翻訳通信研究所．現在，三菱電機情報技術総合研究所．2012 年より早稲田大学国際情報通信研究センター招聘研究員．自然言語処理，音声言語処理の研究に従事．博士（国際情報通信学）．言語処理学会会員．

撫中 達司 （正会員）

1986 年三菱電機入社．オペレーティングシステム（OS），ネットワークシステム，ネットワークセキュリティの研究開発に従事．2015 年より東海大学情報通信学部組込みソフトウエア工学科教授．製造業等の産業用システムへの IoT 活用に関する研究開発に従事．第 21 回電気通信普及財団テレコムシステム技術賞受賞（2006 年），第 62 回電機工業技術功労表彰奨励賞受賞（2012 年）．博士（工学）．IEEE Senior member，電子情報通信学会会員．本会シニア会員．

匂坂 芳典

1973 年早稲田大学理工学部物理学科卒業．1975 年同大学大学院修士課程修了．同年日本電信電話公社（現，NTT）武蔵野電気通信研究所入社．1986 年より国際電気通信基礎技術研究所（ATR）に出向．2001 年より早稲田大学教授，現在の所属は早稲田大学基幹理工学部応用数理学科．工学博士．音声合成・音声認識を中心とした音声情報処理，言語情報処理の研究に従事．日本音響学会，電子情報通信学会，IEEE，米国音響学会各会員．