Vol.2015-NL-220 No.5
2015/1/19

文脈・語義対応の階層ベイズ推定による教師なし語義曖昧性解消

谷垣 宏一[1,2,a)]　徳本 修一[1]　撫中 達司[1]　匂坂 芳典[2]

概要：語彙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，識別対象である語義は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り，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) 類似した文脈に出現する語群の語義は，互いの語義からの外挿に従う．2) 同じ語の各出現における語義は，単語タイプ毎の事前分布に従う．これらの相補的制約を単一の階層ベイズモデルに統合し，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，教師なし学習，階層ベイズ，MCMC，ギブスサンプリング

Hierarchical Bayesian word sense disambiguation for mapping context space to sense space

Tanigaki Koichi[1,2,a)]　Tokumoto Shuichi[1]　Munaka Tatsuji[1]　Sagisaka Yoshinori[2]

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 infer a natural correspondence between unlabeled corpora 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 clearly show the advantages of our hierarchical model.

Keywords: word sense disambiguation, unsupervised learning, hierarchical Bayes, MCMC, 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，文書中に出現する全ての語を対象に語義を識別するタスクであり，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全ての語義を潜在的に識別対象とし，膨大なクラスを扱う [*1]．さらに語義の分布は品詞と較べてドメインに強く依存することが知られている [2, 3]．こうした理由から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盛んに研究されてきた．

辞書知識を利用した教師なし WSD の典型的な方法は，

[1] 三菱電機 (株) 情報技術総合研究所 / Information Technology R&D Center, Mitsubishi Electric Corporation
[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
[a)] Tanigaki.Koichi@ap.MitsubishiElectric.co.jp

[*1] 例えば SemEval の WSD タスクで用いられてきた WordNet [1] の英語 3.1 版では 11 万種類の概念 (synset) で語義を表す．

テキスト中で対象語から一定の範囲に出現している語を文脈語とし，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 [2,4]．語義を対象語毎に独立推定する換わりに，テキストの一定範囲に出現する語群を対象に，ページランク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，対象ドメインのテキストコーパスから広く対象語と類似した文脈に出現する語を参照し，手がかりとして利用するアプローチを取る．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 [7] に基づいている．これまでに我々は，文脈の類似度と語義の類似度で定義する距離空間上の分布の偏りに基づいて，各語の尤もらしい語義を同時推定する方式を提案した [8]．その中で，トークン [*2] の文脈類似度に換えて，比較的安定した統計量を利用可能な単語タイプの分布類似度 [9] を用いた評価を行い，優れた WSD 性能が得られることを示した．分布類似度を教師なし WSD に適用した例としてはこの他 [10,11] 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良い結果が得られたとの報告がある [11]．これらの先行研究は文脈窓に制約されない教師なし WSD の可能性を示しているものの，単語タイプの分布類似度を用いたものであり，厳密には WSD の問題を扱っていない．

本稿では，トークンの文脈類似度に基づく教師なし all-words WSD の新しいモデルを提案する．提案法は [8] のモデルを拡張するものであり，文脈が類似する語の間では語義が類似するという性質に加えて，新たに，単語タイプが同じトークンは同じ語義で用いられやすいという性質 [12,13] を扱い，単一の階層モデルに統合する．また，モデルパラメータの推定を [8] の最尤推定からベイズ推定として再定式化し，新たな性質を扱うため導入した潜在変数について周辺化した解を得る．このように，文脈と単語タイプそれぞれの観点による相補的制約をモデル化することで，ラベルなしコーパスのトークンと膨大な語義の間に自然な対応を推定することが可能となる．以下本稿では，まず 2 節で提案法による教師なし WSD の基本的なアイデアを述べ，3 節〜5 節で定式化する．6 節と 7 節では SemEval データセットを用いた性能評価結果を報告する．8 節では本稿のまとめと今後の課題を述べる．

[*2] 本稿では語の出現（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．

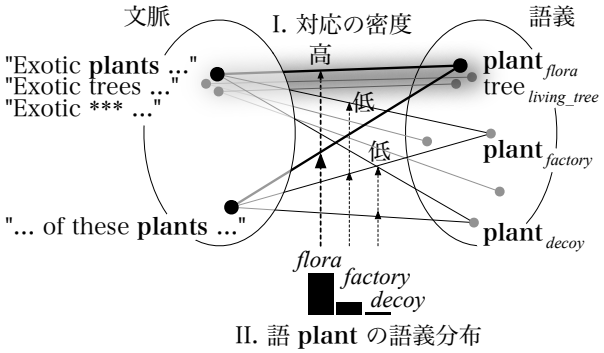

図 1 曖昧性解消のため推定する 2 つの統計量

2. 曖昧性解消の基本的なアイデア

ラベルなしコーパスで語義の曖昧性を解消するため，単一ドメインにおける語義の性質として以下 I, II に注目する．これらはトークンの語義推定に相補的な制約を与える．

性質 I. 似た文脈に現れるトークンは（単語タイプに関係なく）似た意味を表すことが多い [7].

性質 II. 単語タイプが同じトークンは（文脈に関係なく）同じ意味で使われることが多い [12,13].

例えば，“Exotic plants ...”, “Exotic trees ...” などの文を含むコーパスが与えられたとする（**図 1**）．語 plant は語義として *flora*（植物），*factory*（工場），*decoy*（桜客）を持ち，曖昧性がある．同様に語 tree も語義として *living_tree*（樹木）や（樹形図）などの語義を持ち曖昧である．ここで，語 plant と tree が共に exotic（外来種の）の修飾を受けており文脈が似ていること，さらに語義 *flora* と *living_tree* が比較的近い概念であることを考慮すれば，性質 I に従いこれらのトークンをそれぞれ *flora*, *living_tree* と尤もらしく解釈できる．一方，“... of these plants ...” の plants のように有効な構文的特徴を持たない対象語や，低頻度語を文脈語とする対象語では，類似した文脈に出現する他の語が得にくいことがある．その場合，語 plant の語義として他の文脈では *flora* が尤もらしいことを考慮すれば，性質 II に従いここでも尤もらしい解釈として *flora* が得られる．

このような性質 I, II による解釈の尤もらしさは相対的なものであり，他の語の解釈に依存する．このため，文書中で I, II を決定的な規則として適用しても曖昧性を適切に解消することは難しい．そこで本稿では，I, II の性質をそれぞれ，対応の密度と語義の確率分布という統計量としてモデル化し，確率的な制約として適用する．これは以下の仮定に基づいている．

仮定 I. 辞書に規定される膨大な語義概念は本来連続な空間の量子化であり，文脈と語義の対応は単語タイプの違いを越えて滑らかな性質を持つ．

仮定 II. 個々のトークンと語義の対応は，単語タイプ毎に 1 つに定まる語義分布より確率的に生成される．

2 つの統計量は，階層ベイズ推定の枠組みで単一のモデル

として統合し，ラベルなしデータセットから同時推定する．

3. 問題の定式化

具体的なモデル化の準備として，本節では教師なし 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集合を $\mathfrak{Y} = \{Y_i | Y_i = \{y_{ij}\}_{j=1}^{M_{w_i}}\}_{i=1}^{N}$ とする．ただし Y_i はトークン x_i の語義候補集合であり，w_i は x_i の単語タイプ，M_{w_i} は語義候補の数を表す．また，任意の 2 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 $d_{\mathrm{x}}(x_i, x_{i'})$ で定義され，任意の 2 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 $X, \mathfrak{Y}$（および $d_{\mathrm{x}}, d_{\mathrm{y}}$）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 [*3]，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．

いま，トークン $x_i \in X$ の正しい語義が $y_{ij} \in Y_i$ であるとの仮説を，変数 z_i を用いて $z_i = j$ と表す．また，全トークン X の語義割り当て仮説をベクトル $\boldsymbol{z} = [z_1, \ldots, z_N]$ で表す．全モデルパラメータを $\boldsymbol{\Theta}$ とするとき，与えられたデータ $X, \mathfrak{Y}$ より $\boldsymbol{z}$ と $\boldsymbol{\Theta}$ の事後分布 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ を推定すれば，各 x_i の最適な語義 y_{ij^*} は z_i の期待値を最大化する j により次式で定まる．

$$j^*|_i \;:=\; \arg\max_j \sum_{\boldsymbol{z}: z_i = j} \int_{\boldsymbol{\Theta}} p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})\, d\boldsymbol{\Theta} \qquad (1)$$

本稿では事後分布 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ の推定にギブスサンプリングを用いる．ギブスサンプリングはマルコフ連鎖モンテカルロ法の一つであり，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，求めたい事後分布からのサンプルを得る方法である [14]．$\boldsymbol{z}$ および $\boldsymbol{\Theta}$ を構成する各変数を十分な回数 T 回ずつサンプリングし，t 回目に得られた z_i のサンプルを $z_i^{(t)}$ とすると，式 (1) の積分は $z_i^{(t)}$ の平均によって次式で近似できる．

$$\sum_{\boldsymbol{z}: z_i = j} \int_{\boldsymbol{\Theta}} p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})\, d\boldsymbol{\Theta} \;\simeq\; \frac{1}{T} \sum_{t=1}^{T} \delta_{z_i^{(t)} j} \qquad (2)$$

式中の δ_{ij} はクロネッカーのデルタであり，$i = j$ のとき $\delta_{ij} = 1$，$i \neq j$ のとき $\delta_{ij} = 0$ である．

このように，本方式ではモデルパラメータ $\boldsymbol{\Theta}$ について周辺化した解を得る．$\boldsymbol{\Theta}$ には後述するように元の WSD の問題には含まれない潜在変数を導入し，語義の推定に緩い制約を与えるため利用するが，最終的には導入した潜在変数による条件付けを消去し，WSD として合目的的な解を得ることができる．以下 4 節では，事後分布 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ のモデル化について述べ，5 節では事後分布からのサンプリングについて述べる．**表 1** に 3 節〜5 節で用いる記号を纏めた．

[*3] （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について後述する変数 z_i を正解に固定すればよい．

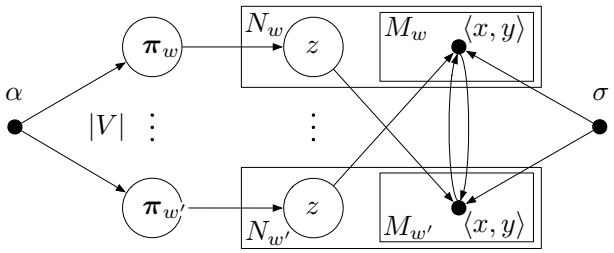

図 2 文脈と語義の対応の確率的生成モデル

表 1 本稿で用いる記号

与えられるデータ $X, \mathfrak{Y}$ に関する記号	
X	トークンの集合．$X = \{x_i\}_{i=1}^{N}$．
$\mathfrak{Y}$	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．
$d_{\mathrm{x}}(\cdot,\cdot)$	トークン $x_i, x_{i'}$ 間の文脈距離を与える関数．
$d_{\mathrm{y}}(\cdot,\cdot)$	語義 $s_{ij}, s_{i'j'}$ 間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	
$\boldsymbol{z}$	トークンの語義割り当てベクトル．
	$\boldsymbol{z} = [z_1, \ldots, z_N]$, $z_i \in \{1, \ldots, M_{w_i}\}$．
$\boldsymbol{\Theta}$	全モデルパラメータ．$\boldsymbol{\Theta} = \langle \boldsymbol{\pi}_w, \ldots, \boldsymbol{\pi}_{w'}, \sigma_{\mathrm{x}}, \sigma_{\mathrm{y}}, \alpha \rangle$．
$\boldsymbol{\pi}_w$	w の語義確率分布．$\boldsymbol{\pi}_w = [\pi_{w1}, \ldots, \pi_{wM_w}]$．
$\sigma_{\mathrm{x}}, \sigma_{\mathrm{y}}$	カーネル密度推定の文脈，語義平滑化パラメータ．
α	語義確率分布のディリクレ平滑化パラメータ．

4. 文脈と語義の対応の確率的生成モデル

本節では，事後分布 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ をモデル化する．本モデルをプレート表記を用いて**図 2** に示す．塗りつぶしの円は観測変数を表し，値が外部から与えられる．中抜きの円は潜在変数を表し，値は観測変数を元に推定する．矢印は変数間の依存関係を，矩形は繰り返しを表す．ただし $w, w', \cdots \in V$ に関する繰り返しは矩形で省略せず，添字として w, w' を明示して依存関係が単語タイプ間で交差することを示す（V は語彙，$|V|$ は異なり語数を表す）．変数の依存関係における本モデルの特徴は，観測として得られるトークンと語義（候補）の対応 $\langle x, y \rangle$ が生成される確率過程を階層化し，各トークンの語義割り当て z の推定に 2 つの緩やかな制約を与える点である．すなわち：

依存関係 I. 文脈と語義の対応が単語タイプの違いを越えて滑らかな性質を持つとの仮定 I に基づいて，z から $\langle x, y \rangle$ の生成に単語タイプ間のクロスバリデーションを適用する．これにより，或るトークン x_i の語義割り当て z_i は，（文脈の近傍において）単語タイプが異なる他のトークン $x_{i'}$ と語義の対応 $\langle x_{i'}, y_{i'z_{i'}} \rangle$ を良く外挿するほど尤もらしいとして，z_i に制約を与える．

依存関係 II. 各トークンの語義割り当て z が単語タイプ

毎の語義分布より生成されるとの仮定 II に基づいて，単語タイプ毎の語義分布 $\boldsymbol{\pi}_w$, $\boldsymbol{\pi}_{w'}$, ... をモデルパラメータ $\boldsymbol{\Theta}$ の要素として導入し，z の事前分布とする．ここで $\boldsymbol{\pi}_w$ は，単語タイプ w の語が語義 $y_{i1}, y_{i2}, \ldots$ で用いられる確率ベクトル $\boldsymbol{\pi}_w = [\pi_{w1}, \pi_{w2}, \ldots]$ である ($\sum_j \pi_{wj} = 1$)．或る単語タイプ w に関するトークン $x_i, x_{i'}, \ldots$ に対し $z_i, z_{i'}, \ldots$ を推定する際，これら $z_i, z_{i'}, \ldots$ が $\boldsymbol{\pi}_w$ を期待値とする共通の事前分布から生成されるとして緩やかな制約を与え，文脈上の手がかりが十分得られないトークンで安定した推定を可能とする．

II で導入する単語タイプ毎の語義確率 $\boldsymbol{\pi}_w$ 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として手続き的な手法で実現され，半教師あり WSD に用いられた [12,13]．提案法は，これを単一モデルの事前確率として扱い，階層ベイズ推定の枠組みで実現する．また先行研究でベイズ推定を語義に適用した例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 [15–17]．これらは word sense induction タスクで出現文脈によるトークンのクラスタリングを狙ったもので，単語や単語素性を語義から直接生成するモデルとなっている．このようなモデル化は，各語義（クラス）について語の出現が十分割り当てられることを前提としたものであり，語義の種類が膨大で，個々の語義について十分なデータが得られない all-words WSD に適用することは難しい．提案法では，スパースな語義の揺わりに平滑化した密度分布を介して生成過程をモデル化する．

モデルの各変数は以下の過程より生成されると仮定する．まず，一様分布を事前分布として，単語タイプ $w \in V$ に応じた語義確率 $\boldsymbol{\pi}_w$ が生成される．$\boldsymbol{\pi}_w$ の生起確率は次式の対称ディリクレ分布に従う．

$$\boldsymbol{\pi}_w \;\sim\; \mathrm{Dir}(\boldsymbol{\pi}_w | \,\alpha/M_w, \ldots, \alpha/M_w\,) \tag{3}$$

ここで記号 $\sim$ は左辺の生起確率が右辺の確率分布に従うことを表す．M_w は w の語義候補数を表す．α は予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 $\boldsymbol{\pi}_w$ の分布は一様分布に偏る．

各トークン x_i の語義割り当て z_i は，単語タイプ w_i の語義確率分布 $\boldsymbol{\pi}_{w_i}$ より生成される．z_i は多項分布に従う．

$$z_i \;\sim\; \mathrm{Mul}(z_i \,|\boldsymbol{\pi}_{w_i}\,) \tag{4}$$

各トークンと語義の対応 $\langle x_i, y_{ij}\rangle$ の出現確率は，クロスバリデーションにより，データセット中で x_i とは単語タイプが異なるトークン X_{-w_i} の語義候補 $\mathfrak{Y}_{-w_i}$，および，語義割り当て $\boldsymbol{z}_{-w_i}$ との関係よりモデル化する．文脈と語義の関係が連続との仮定に基づいて，カーネル密度推定 [18] を適用して得られる密度分布より $\langle x_i, y_{ij}\rangle$ が生成される．

$$\begin{aligned}&\langle x_i, y_{ij}\rangle \\ &\;\sim\; \mathrm{Kdens}(x_i, y_{ij} \,|\, X_{-w_i}, \mathfrak{Y}_{-w_i}, \boldsymbol{z}_{-w_i}, \sigma_\mathrm{x}, \sigma_\mathrm{y}\,)\end{aligned} \tag{5}$$

Kdens() は，X_{-w_i} に含まれる N_{-w_i} 個のトークンの語義割り当て $\langle x_{i'}, y_{i'z_{i'}}\rangle$ より，カーネル関数 $k()$ で $\langle x_i, y_{ij}\rangle$ を外挿する密度の平均であり，次式で定義する．

$$\begin{aligned}&\mathrm{Kdens}(x_i, y_{ij} \,|\, X_{-w_i}, \mathfrak{Y}_{-w_i}, \boldsymbol{z}_{-w_i}, \sigma_\mathrm{x}, \sigma_\mathrm{y}\,) \\ &\quad:=\quad \frac{1}{N_{-w_i}} \sum_{i':w_{i'}\neq w_i} k\big(x_i, y_{ij}, x_{i'}, y_{i'z_{i'}} | \sigma_\mathrm{x}, \sigma_\mathrm{y}\big)\end{aligned} \tag{6}$$

本稿で用いる $k()$ はガウスカーネルであり，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する．式中，$\sigma_\mathrm{x}, \sigma_\mathrm{y}$ 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．

$$\begin{aligned}&k\big(x_i, y_{ij}, x_{i'}, y_{i'j'} | \sigma_\mathrm{x}, \sigma_\mathrm{y}\big) \\ &\quad:=\quad \frac{1}{2\pi\sigma_\mathrm{x}\sigma_\mathrm{y}} \exp\left[-\frac{{d_\mathrm{x}}^2(x_i, x_{i'})}{2{\sigma_\mathrm{x}}^2} - \frac{{d_\mathrm{y}}^2(y_{ij}, y_{i'j'})}{2{\sigma_\mathrm{y}}^2}\right]\end{aligned} \tag{7}$$

すなわち式 (5) では，文脈的にも意味的にも近い他の語の語義割り当てが数多く存在するほど，対応 $\langle x_i, y_{ij}\rangle$ が生成されやすくなる．

5. サンプリングによる事後分布の推定

3 節では，事後分布 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ から得られたサンプルをもとに最適な語義を決定する方法を述べた．本節ではギブスサンプリングを適用して実際に変数のサンプルを得る方法を述べる．サンプリングの対象は，各単語タイプの語義確率 $\boldsymbol{\pi}_w$ と各トークンの語義割り当て z_i である．

いま，単語タイプの語義確率 $\boldsymbol{\pi}_w$ に注目し，その他の変数 $\boldsymbol{z}, \boldsymbol{\Theta}_{-\boldsymbol{\pi}_w}$ を固定すると，事後確率 $p(X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta})$ は条件付き確率 $p(\boldsymbol{\pi}_w | X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta}_{-\boldsymbol{\pi}_w})$ に比例する．ここで，

$$\begin{aligned}&p(\boldsymbol{\pi}_w \,|\, X, \mathfrak{Y}, \boldsymbol{z}, \boldsymbol{\Theta}_{-\boldsymbol{\pi}_w}) \\ &\quad\propto\quad \mathrm{Dir}(\boldsymbol{\pi}_w \,|\, \alpha/M_w, \ldots, \alpha/M_w\,) \prod_{i:w_i=w} \mathrm{Mul}(z_i \,|\, \boldsymbol{\pi}_w\,) \\ &\quad\propto\quad \prod_j {\pi_{wj}}^{\alpha_w/M_w - 1} \prod_j {\pi_{wj}}^{\mathcal{N}_{wj}} \\ &\quad\propto\quad \mathrm{Dir}\left(\boldsymbol{\pi}_w \,\middle|\, \frac{\alpha_w}{M_w} + \mathcal{N}_{w1}, \ldots, \frac{\alpha_w}{M_w} + \mathcal{N}_{wM_w}\right)\end{aligned} \tag{8}$$

であるから，式 (8) のディリクレ分布より $\boldsymbol{\pi}_w$ の新たなサンプルを得る．式中の $\mathcal{N}_{wj}$ 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，語義割り当て z_i が j となっているトークンの数である．

同様にして，トークンの語義割り当て z_i のサンプルは式 (9) のディリクレ分布より得る．

$$\begin{aligned}&P(z_i = j \,|\, X, \mathfrak{Y}, \boldsymbol{z}_{-i}, \boldsymbol{\Theta}\,) \\ &\quad\propto\quad \mathrm{Mul}(z_i = j \,|\, \boldsymbol{\pi}_{w_i}\,) \prod_{i':w_{i'}\neq w_i} \mathrm{Kdens}\big(x_{i'}, y_{i'z_{i'}} \big| \\ &\qquad X_{-w_{i'}}, \mathfrak{Y}_{-w_{i'}}, \boldsymbol{z}_{-w_{i'}}, z_i = j, \sigma_\mathrm{x}, \sigma_\mathrm{y}\big)\end{aligned} \tag{9}$$

6. 評価実験

6.1 実験条件

提案する階層モデルの有効性を確認するため，次の 3 種類の条件で語義の曖昧性解消を行い，性能を比較する．これら 3 つの方法は，単語タイプが等しいトークンの語義を推定するときの制約の強さが異なる．

(1) **階層モデル：** 本稿で提案する階層モデル（図 2）を用いて語義を推定する．本モデルでは，単語タイプが同じトークン群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り当てに対し，単語タイプ毎の語義分布による制約がかかる．この制約は緩やかなため，別の語義の方がより尤もらしいという手がかりが文脈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．

(2) **非階層モデル：** 図 2 より変数 α と $\boldsymbol{\pi}$ の階層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一切かからず，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．

(3) **非階層モデル最頻語義：** (2) の非階層モデルを用いて各トークンの語義を推定した後，それらトークンの語義推定結果より単語タイプ毎の最頻語義を決定する．この最頻語義で元のトークンの語義推定結果を置き換えて出力する．最頻語義が一意に定まらない場合は複数の語義を等確率で出力する．この方法では，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固い制約がかかる．このため，同じ語が文脈によって複数の語義で使い分けられることは考慮されない．

評価用データには，SemEval-2 英語 all-words WSD タスクのデータセット [19] を用いた．本データセットでは，テストセットとラベルなしコーパスが，単一のドメイン（環境ドメイン）に閉じて提供される．テストセットは 3 文書，5,348 語からなり，うち 1,398 語が WSD の対象語である．対象語は名詞 1,032 語，動詞 366 語からなる．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）．ラベルなしコーパスは，後述の文脈距離の実装において，分布類似度を計算するためだけに用いた．語義を規定する辞書は WordNet 3.0 である．

ところで，比較する 3 つの方式でどの程度効果が得られるかは，テストセットの語義にどの程度曖昧性があるかに依存する．そこで事前分析として，テストセット 1,398 語における語義の曖昧性を調べた．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 [*4]．正解語義のパープレキシティは，ひと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使われているか（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，辞書の規定で語義候補が平均何種類あるかを表す．分析の結果を**図 3** に示す．濃い網掛けは区間毎の頻度分布を表し，薄い網掛けは累積分布を表す．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．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 と較べて 1/4 ∼ 1/5 に絞られることがわかる．したがって同語の語義が一致するよう制約を与える (1) 階層モデルおよび (3) 非階層モデル最頻語義では，(2) 非階層モデルと較べて良い性能を期待できる．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語の全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%に留まる．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 (3) 非階層モデル最頻語義と較べて (1) 階層モデルでは優れた性能が得られると期待できる．

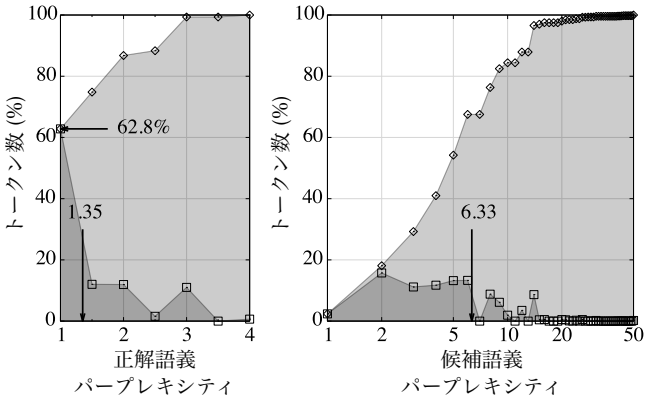

図 3 SemEval-2 テストセットにおける語義の曖昧性

提案法は計算に際し，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．評価用には，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 et al. 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，全く異なる単語ペアどうし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する．距離関数とするため類似度の逆数を取り，さらに特徴の数に対する感度を抑えるため対数を

[*4] 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を用いる．

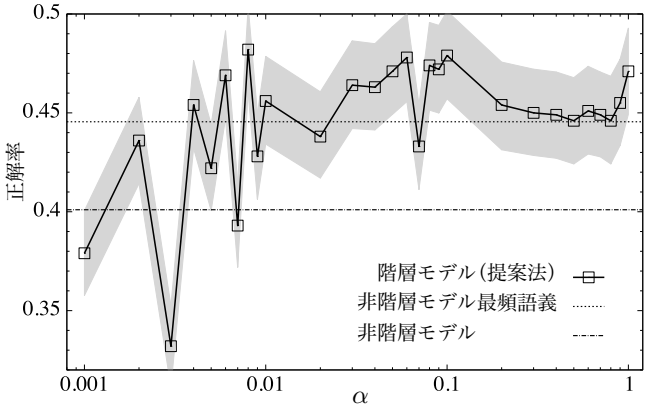

図 4 モデル階層化の効果

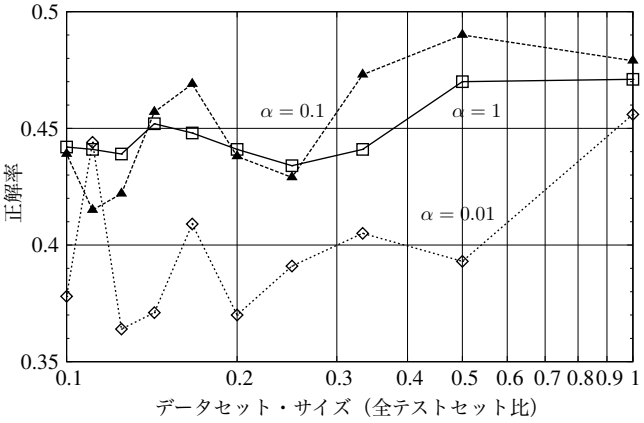

図 5 ラベルなしデータセットの大きさによる性能変化

取って用いた．また，サンプリングの計算効率化とノイズ除去のため，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る．k の値は，事前に行った評価より $k > 10$ 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，本評価では $k = 10$ とした．一方，語義距離には，先行研究 [8,10] 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）を用いる．実際の計算には Pedersen et al. の提供するライブラリ [22] を用いた．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う．

WSD 性能の評価には SemEval が提供する評価ツール scorer2 を用いた．scorer2 による評価スコアは再現率と適合率であるが，システム出力をカウントする際，単語毎にシステムの総出力数で正規化した確率カウントを用いており，一般的な再現率・適合率の定義とは異なる．したがって以下では scorer2 で算出した再現率を「正解率」と表記して示す．なお，本評価では全対象語に対し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．ギブスサンプリングの回数は，全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 $1, 2, \ldots, 9 \times 10^N$ 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．実線と四角の印によるグラフは α を変えたときの階層モデル（提案法）の正解率をプロットしたものであり，網掛けの領域は 95%信頼区間を表している．非階層モデルおよび非階層モデル最頻語義については，乱数の種を換えて階層モデルと同様に 28 回の評価を行い，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．

この図から，3 つの方式を較べると全般に，階層モデル，非階層モデル最頻語義，非階層モデルの順で良い性能が得られたことがわかる．階層モデルでは，α を極端に小さく取ったときには性能が不安定になる傾向が見られるものの，$\alpha \geq 0.01$ では安定し，他の 2 つの方式を有意に上回る性能が得られた．非階層モデルと較べて階層モデルおよび非階層モデル最頻語義で良い性能が得られることは，後者で利用する同語トークン間の語義制約が語義曖昧性解消に有効であることを示している．さらに非階層モデル最頻語義と較べて階層モデルの性能が上回ることは，同語トークン間の語義制約を実現する方式として，提案法による緩やかな制約としてのモデル化が有効であることを示している．すなわち，本稿で提案する階層モデルが語義の曖昧性解消において有効であることが確認できた．

本評価において，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，SemEval-2 タスクに参加したシステムのうち，教師なし方式で最も良いシステムの正解率は 0.495 である [19]．評価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る．SemEval-2 タスクの上位システムでは性能を上げるため，本評価のようにリソース（データセットと辞書）をそのまま用いるのではなく，辞書語義のプルーニングを適用したり [23]，Web から大量に取得したテキストを利用する [24] 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を次節で考察する．

7. 考察

本節では，提案法を適用するラベルなしデータセットのサイズと正解率の関係について考察する．2 節・4 節で述べたように，提案法の特徴は語義の曖昧性を解消するために 2 つの制約を適用する点である．そのひとつが前節で効果を検証した同語トークンの語義に対する事前分布の適用

であり，もうひとつが，以下でその効果を検証する文脈と語義の対応の一般化である．文脈と語義の対応を類似度により一般化してモデル化することで，テキスト中の語の出現位置近傍に限定することなく，データセット中の文脈が類似した語を参照し，語義曖昧性解消の手がかりとすることができる．このとき，データセットのサイズが大きい程，文脈が類似した語が得やすくなり，性能も向上するものと期待できる．

そこでデータセットのサイズを変えて正解率の変化を調べた．評価に用いるデータセットは前節と同様 SemEval-2 のテストセットであるが，本節では語の出現順序を保持したままテストセットを品詞毎に N 等分 ($N = 1, 2, \ldots, 10$) して，$1/N$ のサイズとなったデータセットでそれぞれ WSD を行なった後，テストセット全体の正解率を評価する．α は 0.01, 0.1, 1 の 3 種類の場合で実験した．実験結果を**図 5** に示す．グラフは局所的には上下の変動があるものの，全体的には α の値に関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，提案法ではラベルなしデータを増やすことで WSD 性能の向上が得られる．したがって今後，テストセットにラベルなしコーパスを追加して大規模化したデータセットに提案法を適用することにより，テストセットの正解率が向上する可能性がある．

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) 文脈近傍語の間で語義が類似しやすい性質，および，2) 同語のトークンで語義の一致しやすい性質，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評価では，同語のトークンで語義の制約を用いない方法や，同語のトークンで語義を厳密に一致させる手法と比較して正解率の有意な向上が得られ，提案する階層モデルの有効性を確認した．また，計算対象のデータセットを大きくすることで，正解率が改善される傾向があることを示した．今後，大量のラベルなしコーパスを利用したときの性能を評価すると共に，文脈距離の改良についても検討したい．

参考文献

[1] Miller, G. A.: WordNet: a lexical database for English, *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41 (1995).
[2] Agirre, E. and Edmonds, P.: *Word sense disambiguation: Algorithms and applications*, Vol. 33 (2006).
[3] 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*, pp. 79–86 (1997).
[4] Navigli, R.: Word sense disambiguation: A survey, *ACM Computing Surveys (CSUR)*, Vol. 41, No. 2, p. 10 (2009).
[5] Agirre, E. and Soroa, A.: Personalizing pagerank for word sense disambiguation, *Proc. EACL 2009*, 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] Harris, Z. S.: Distributional structure., *Word* (1954).
[8] Tanigaki, K., Shiba, M., Munaka, T. and Sagisaka, Y.: Density Maximization in Context-Sense Metric Space for All-words WSD, *Proc. ACL 2013*, pp. 884–893 (2013).
[9] Lin, D.: Automatic retrieval and clustering of similar words, *Proc. COLING 1998*, pp. 768–774 (1998).
[10] 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).
[11] Agirre, E., De Lacalle, O. L., Soroa, A. and Fakultatea, I.: Knowledge-based WSD on specific domains: performing better than generic supervised WSD, *Proc. IJCAI 2009*, pp. 1501–1506 (2009).
[12] Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods, *Proc. ACL 1995*, pp. 189–196 (1995).
[13] Abney, S.: Understanding the yarowsky algorithm, *Computational Linguistics*, Vol. 30, No. 3, pp. 365–395 (2004).
[14] Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, No. 6, pp. 721–741 (1984).
[15] Lau, J. H., Cook, P., McCarthy, D., Newman, D. and Baldwin, T.: Word sense induction for novel sense detection, *Proc. EACL 2012*, pp. 591–601 (2012).
[16] Yao, X. and Van Durme, B.: Nonparametric bayesian word sense induction, *Proc. TextGraphs-6*, pp. 10–14 (2011).
[17] Brody, S. and Lapata, M.: Bayesian word sense induction, *Proc. EACL 2009*, pp. 103–111 (2009).
[18] Parzen, E.: On estimation of a probability density function and mode, *The annals of mathematical statistics*, Vol. 33, No. 3, pp. 1065–1076 (1962).
[19] 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. SemEval-2010*, pp. 75–80 (2010).
[20] Thater, S., Fürstenau, H. and Pinkal, M.: Word Meaning in Context: A Simple and Effective Vector Model., *Proc. IJCNLP 2011*, 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] Kulkarni, A., Khapra, M. M., Sohoney, S. and Bhattacharyya, P.: CFILT: Resource Conscious Approaches for All-Words Domain Specific, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 421–426 (2010).
[24] 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 Do-

main, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 396–401 (2010).