2I5-OS-08b-3

幼児の学習バイアスを用いた意味の確率的表現

Probabilistic Representation of Word Senses using Learning Bias of Infant Children

高田 朋貴[*1]
Tomoki Takada

高木 友博[*1]
Tomohiro Takagi

[*1] 明治大学理工学研究科基礎理工学専攻
Computer Science Course, Graduate School of Science and Technology, Meiji University

These days computers are needed to deal with human language more exactly because the language analysis by computers is growing important. In this paper, we aim to enable computers to deal with the ambiguous word sense by proposing the combination method of the learning bias of infant children and the previous machine learning method. First, we obtain the surround information for identifying the target word sense by using LS model. Next, we input these information into LDA and generate the target word sense distribution. We use the learning bias initialization instead of the random initialization. This enables estimating the number of topics automatically and the efficient learning.

1. 序論

近年 WEB の発達により, ビッグデータと呼ばれるように大量のデータが増加し続けている. 例えばソーシャルサイトの発展により, ユーザが能動的に WEB 上にデータを生成することが容易にできるようになった. このようなデータを用い, 推薦エンジンや予測システム等の研究が盛んに行われているが, 単語の意味を考慮した言語処理は, その精度に関わる非常に重要な問題である. しかし, ユーザによるコンテンツの生成により web 上における単語や記号による表現の幅は大幅に広がることになり, 日々創出される単語を辞書に追加して計算機に処理させるだけでは不十分になりつつある. そのため, 今後は計算機が受動的に言語処理するのではなく, 能動的に単語の意味を推論し, 自然言語をいかにして計算機に理解させるかといった方法論を検討することは必要不可欠である. これを実現する為には従来の機械学習に加え, 認知言語学や脳科学, 発達心理学等の観点から人間の学習の本質的なメカニズムとは何かを考え, それらの知見を積極的に取り入れることが重要であると考えている.

本稿では, 幼児の学習バイアスとトピックモデルである LDA を組み合わせることで, 人間のように単語の意味を同定していく過程をモデル化し, 抽象的な単語の意味を確率的に表現する方法論を模索する.

2. 幼児の学習バイアス

幼児の言語獲得において, クワインが提起したガヴァガーイ問題[Quine 1960]がある. この問題は簡単に言えば,「ある事象に対して発せられた言葉が, その事象のどの部分を指示しているのかはわからない」というものである. 幼児はこの問題と同様な状況に置かれているという. 例えば, 母親が子どもに「あれはウサギだよ」と言葉を発した時, 子どもはウサギという言葉の意味を「白い動物」,「耳の長いこと」,「赤い目のこと」等のように無数の意味の候補を推測することができてしまう. この問題を解決するために, 幼児は言葉の意味を推論する際に, 全ての意味をしらみつぶしに検証するのではなく, 一種の思い込みのように意味の可能性を制限しているのではないかという考えがある. その考え方の一つとして“制約理論”[今井 2007][今井 2003]がある. 以下に二つの代表的な制約を示す.

連絡先:高田朋貴, 明治大学理工学研究科基礎理工学専攻,
214-0034 川崎市 多摩区 東三田 1-1-1,
Tel: 044-934-7483, lemone.tomoki@gmail.com

2.1 形状類似性バイアス

「形の似通った事物同士が同じラベルを持つ可能性が高い」と解釈する仮説である. 幼児は未知の事物に対して新奇な言葉が使われるのを聞くと, その言葉を特定の個体を指す固有名詞ではなく, カテゴリーを指示する普通名詞であると判断し, 形の類似性に注目して形が似た他の事物にその言葉を適用する.

ここで注目すべき点は, 幼児はある事物が同じカテゴリーであるかどうかを判定する際に, “形”の類似度に従って判定しているという点である.

2.2 相互排他性バイアス

「相異なるラベルが同じ対象事物に関連づけられることはない」と解釈する仮説である. 幼児が既に知っている事物に対して, 未知の言葉を聴いたとき, その言葉が指し示すものは, 既に名前を知っているものとは異なるものであると解釈する傾向にある.

3. Loosely Symmetric model

このモデルは, 人間の因果帰納等に存在する“対称性バイアス”および前節で述べた“相互排他性バイアス”という 2 つの非論理的な認知バイアスを緩やかに持つ確信度のモデルである[篠原 2007].

いま原因となる事象を p, 結果となる事象を q とした時, 対称性バイアスは「$p \rightarrow q$」という情報から「$q \rightarrow p$」を導き, 相互排他性バイアスは「$p \rightarrow q$」から「$\bar{p} \rightarrow \bar{q}$」を導くことを示す. これらは論理学において逆と裏の関係にあり論理的には誤りであるが, 人間は因果帰納において度々このような推論を行う事が知られている. これらのバイアスを柔軟に扱うことにより, より人間の感覚に近い結果が得られたことが示されている.

表 1 のように a, b, c, d をそれぞれの事象の共起頻度とすれば, LS モデルは式(1)のように示すことが出来る.

$$LS(q|p) = \frac{a + \left(\frac{b}{b+d}\right)d}{a + b + \left(\frac{a}{a+c}\right)c + \left(\frac{b}{b+d}\right)d} \quad (1)$$

表 1 各事象の共起情報

	q	$\bar{q}$
p	a	b
$\bar{p}$	c	d

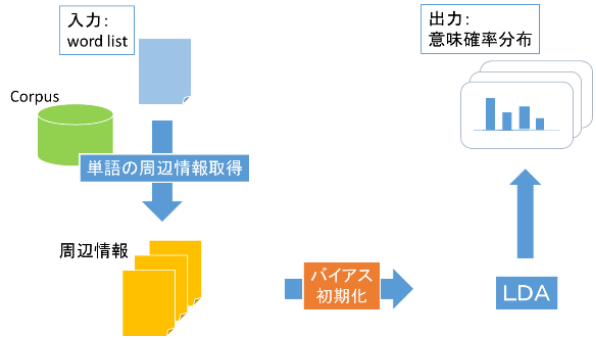

図 1　提案システムの概要図

4. 言葉の意味の確率的表現

近年，言葉の意味を確率的潜在意味分析により確率的に扱おうという研究[持橋 2002][阿部 2007]がされており，良い精度を示している．本稿ではこれらの先行研究に倣い，言葉の意味を確率的に表現することを目指す．既存手法では Probabilistic Latent Semantic Indexing（pLSI）をベースとして用いられているが，本稿では近年注目を集めているトピックモデルの一つである Latent Dirichlet Allocation(LDA)[Blei 2003]を用いる．これらの手法の大きな利点は，単語の意味を各潜在クラスへの所属確率の分布によって表現でき，言葉の意味を確率値という抽象的な表現によって計算機上で表現する事が可能な点である．

5. 提案システム

図 1 は提案システムの概要図である．まず，学習したい単語のリストを周辺情報取得モジュールに渡し，各単語の周辺情報を取得する．次に，各単語を文書，取得された周辺情報をその文書の特徴語とみなし，LDA に入力する．この時，通常の LDA のようにランダム初期化ではなく，バイアスを用いた恣意的な初期化を行う．LDA から得られた文書の所属するトピック分布は，各単語の意味を抽象的に表した確率的表現として扱う．

5.1 LS モデルを用いた周辺情報取得

本稿では，単語の意味を同定する為の周辺情報とは，入力された単語と同一文書に出現した語と定義する．この考え方は，情報検索で一般的な分布仮説を踏襲しており，幼児がある未知語を聞いた時の状況にも類似すると判断した．また，この周辺情報を今回は 1-gram の文字とした．なぜなら言語において 1-gram の文字が最も小さい断片情報であり，幼児が取得する周辺情報の一つ一つも断片的な情報であることが想定されるからである．但し，これでは非常に多くの情報を取得しすぎてしまい，どの情報が意味を同定する上で重要であるかを判断することができない．そこで LS モデルを用い，LS モデルで得られた値が閾値γ以上の文字のみを入力された単語と因果性が高い情報として取捨選択する．これらの処理を加えることで，認知バイアスにより単語の意味を同定する為の情報として有益な情報のみを取得できると考えた．

次に特徴量の値の問題がある．一般的な LDA は文書を対象としたモデルである為，特徴量は整数で扱う．特徴語に重み付けを行うことで LDA を実行する方法[Wilson 2010]もあるが，LS モデルは値の分散値が非常に小さい為，重み付けをしてもほとんど効果がないと考えられる．その為，本稿では 2 種類の特徴量の方法論を検討する．一つ目は単純に文字の存在の有無で特徴量を 0 か 1 で扱う方法である．しかし先に述べたように，LDA は文書に対するモデルであり，文書の特徴量はおおよそ冪乗（べきじょう）則に従うことが知られている．そこで，二つ目として冪乗分布になるような特徴量に変換する為に LS モデルで求めた値を Zipf の法則に当てはめる方法を考えた．LS モデルの値はあまり反映させずにその順序に着目し，式(2)の Zipf の法則に当てはめることで，冪乗化しようというものである．ただし単純に Zipf の法則に当てはめてしまうと，1-gram の文字と関連する単語がたくさんあった場合，つまり順位付けする文字がたくさんあった場合に特徴量が非常に大きくなってしまう．その為，パラメータ s の値に各文書の LS モデルの最大値を代入することで，値が大きくなりすぎることを抑えることにした．なお，冪乗化する際に値が少数になってしまう事があるが，その場合は四捨五入をすることで，整数化を行った．

アルゴリズム：バイアスを用いた初期化アルゴリズム

0. 初期のトピックは 1 つだけであり，一番初めに学習する入力語はそのトピックに所属させる
1. 次の入力語を読み込み，トピック数が 1 の時は，類似度が閾値θ以上であれば，初期のトピックと同じトピックに振り分け，θ未満であれば新しいトピックを生成し，その新規トピックに振り分ける．トピック数が 2 以上になった時，手順 2 に移行し，そうでなければ，手順 1 を繰り返す
2. 入力語と初期化済みの全ての語との類似度を計算する
3. 初期化済みの語のトピック分布 A と入力語との類似度を変換公式に代入し，得られた結果から分布 B を生成し，分布 A の時点で所属確率が最大であったトピックのみに注目し，分布 B からそのトピックの値のみを取得する
4. 全ての初期化済みの単語に対して手順 3 が終わったら，各トピックの最大値を取得する
5. 各トピックの値の総和が 1 になるように正規化した分布 C を生成する
6. 分布 C が一様分布か否かにより，以下のように分岐する
   a) 一様分布であれば，新たなトピックを生成し，入力語を新たなトピックに割り当てる
   b) 一様分布でなければ，分布 C に基づきトピックの初期化を行う
7. 全ての入力語を初期化し終われば終了し，そうでなければ手順 2 に戻る

$$f(k;s,N)=\frac{1/k^s}{\sum_{n=1}^{N}1/n^s} \qquad (2)$$

(s:パラメータ，N:全要素の数，k:順位)

5.2 バイアスによる LDA の初期化

一般に，LDA には 2 つの問題点が存在する．一つ目は，システムの設計者によるトピック数の設定である．二つ目は，初期化による精度依存問題である．前者はあらかじめ設計者がトピック数を事前に設定しなければならず，その数も経験的に設定しなければならない．その解決策として，HDP-LDA[Teh 2006]のように確率的にトピック数を求める手法が提案されているが，トピック数の大きさにバラつきが生じることや，出力の中に人間が解釈できない出力が得られてしまうことがある等の問題点がある．後者は，最初の初期化によっては，局所解から容易に抜け出す事ができず，精度に影響を与えてしまうという問題である．上記 2 点の問題を解決する為に，提案手法では上記に示すアルゴリズムのような学習バイアスを考慮した初期化手法を提案する．

手順 2 の類似度は，式(3)の cosine 尺度を用いた．但し，特徴語の値は使わず，存在の有無のバイナリベクトルとして扱う．従って，各文書のベクトルの値は 1 か 0 のどちらかである．

$$\cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|} \quad (3)$$

手順 3 で用いる変換公式は式(4)である. この式は類似度が高ければ高いほど, 現在注目している初期化済みの単語のトピック分布と同じような分布となり, 低いほど異なるトピック分布となる. 幼児は形状類似性バイアスにより類似している事物に対して同じラベルを適用するという仮説から, 形の類似の度合により同一のカテゴリーかどうかを判断していると考えられる. したがって本稿では形の類似度合を周辺情報同士の類似度と見なし, 類似度が高ければ同一のトピックである可能性が高く, 類似度が低ければその事物の所属するトピックとは異なるトピックに所属する確率が高いと解釈することにした. また, 式(6)はシグモイド関数であり, 値域が[0,1]に収まるように調整がしてある.

$$g(x_i, sim) = \frac{f(x_i, sim)}{\sum_j f(x_j, sim)} \quad (4)$$

$$f(x, sim) = sig(sim) * x + \big(1 - sig(sim)\big) * (1 - x) \quad (5)$$

$$sig(x) = \frac{1}{1 + e^{-gain(x-0.5)}} \quad (6)$$

(x_i:トピック i の所属確率, sim:類似度)

手順 6 では一様分布か否かにより, 新規にトピックが生成できるかを判定する. 一様分布は各トピックに所属する確率が等確率であり, 裏を返せばどのトピックにも明確に所属しないということから, 新たなトピックに所属するのではという仮説に基づいている. 一様分布であるかどうかの判定には, 分布 C と一様分布との Jensen-Shannon Divergence を用いた. ここでは, 類似度のように 1 が最も類似していると計算するために, 式(7)のように変形して用いている. この値が閾値δ以上であるときに, 分布 C は一様分布であると判定する.

$$Sim_{dist}(P, Q) = e^{-D_{JS}(P||Q)} \quad (7)$$

$$D_{JS}(P||Q) = \frac{1}{2}\big(D_{KL}(P||R) + D_{KL}(Q||R)\big) \quad (8)$$

$$D_{KL}(P||Q) = \sum_x P(X = x) \log \frac{P(X = x)}{Q(X = x)} \quad (9)$$

6. 実験

6.1 実験設定

実験のコーパスとして, 読売新聞(1989,1990,1994)を用いた. 1989, 1990 年を学習データ, 1994 年をテストデータとして, Perplexity の測定に用いた. 各閾値は, γ=0.5, θ=0.5, δ=0.999 とし, 式(6)のシグモイド関数のパラメータは gain=10 とした. また, LDA の推論には Collapsed Gibbs Sampler を用い, ハイパーパラメータはα=1, β=1 とし, イテレーション回数は 100 回とした. また, 学習データ中に出現した普通名詞 52,703 単語を今回学習させる単語とした.

6.2 実験結果

まず初めに LS モデルを用いて取得した周辺情報について述べる. 表 2, 3 はそれぞれ国会とソ連に関する周辺情報を示しており, 変換値は Zipf の法則に当てはめて変換した時の値を示している. 5.1 節で述べたように LS 値の分散値は非常に小さい事が見て取れ, 変換値は冪乗分布に従う値に変換されている. また, 各単語と上位に関連している文字は一文字ではあるが, それぞれに関連している単語を想起することができる. (国会であるならば政党や議員, ソ連であるならばゴルバチョフ等)

次に, バイアスを用いた初期化アルゴリズムを用いた時のトピックの増加具合を示すグラフを図 2 に示す. 特徴量により多少の増加具合は異なっているが指数関数的には増加せず, ほぼ単語数に対して線形に増加している.

図 3, 4 はランダムに初期化した場合とバイアスを用いて初期化した場合の Perplexity の違いを示したものである. イテレーション回数を重ねるとランダムによる初期化との差異は現れないが, 回数が少ない場合ではより早く Perplexity が減少した.

続いて, 生成された各語の意味確率分布について考察する. 図 5, 6 は“国会”の意味確率分布であり, 図 7, 8 は“ソ連”の意味確率分布である. 図 5 と図 6, 図 7 と図 8 をそれぞれ比較すると, Zipf の法則による特徴量付け(ZIPF と表示)の方が, 全ての特徴量を 1 として取得した時(ALL1 と表示)よりも尖った分布となっており, 曖昧性が少ないと考えられる分布であった. これは, 特徴量を全て 1 とした場合, 特徴量による文書間の違いが鮮明でなくなるため, LDA が明確にトピック分割できなかったことが想定され, 分布に散らばりが起きてしまったと考えられる. 次に, ランダムによる初期化とバイアス初期化による比較だが, 特徴量を 1 とした場合のトピック－単語分布において, 「、」や「。」のようなストップワードが上位の確率値を持つトピックがランダム初期化の場合は 10 トピックあったのに対し, バイアス初期化では 6 トピックであった. これはバイアス初期化による精度向上と考えることが出来る. しかし, Zipf の法則に従い特徴量を付与した場合にはほとんど違いが見られなかった. 更に, Zipf の法則での特徴づけを用いた場合の方で, 表 3, 4 のように式(7)を用いて意味確率分布同士を比較しても, 大きな違いは見られなかった. これは, 特徴量を全て 1 とした場合は特徴量による違いは大きくない為, バイアスを用いた初期化が結果に大きな影響を与えることになったと考えられるが, Zipf に基づく特徴量では, 既に特徴量の時点で大きく差異が出ている為, 最終的な LDA の結果では大きな差が生じなかったと考えられる.

7. 結論

LS モデルを用い, 単語の意味を同定するための周辺情報を取得し, LDA の初期化に幼児のバイアスに基づく手法を導入した. 周辺情報に対する特徴量づけについては, 文字の存在の有無を表したバイナリベクトルを用いた方法よりも, Zipf の法則に基づき変換させた値を用いた手法の方が, 生成された意味確率分布は曖昧性が少ない分布となった. また, 幼児の学習バイアスを考慮した初期化手法により, LDA のトピック数を確率的な手法に依らずに自動的に決定することができ, 学習初期の効率化の効果を認めることが出来たが, 初期化による最終的な意味確率分布の精度向上の優位性までを示すことができなかった. これについては, より幼児の学習バイアスを考慮した手法を検討していく必要性があると考えている.

本手法により, 単語の曖昧な意味を確率的に表現することができれば, よりきめの細かい単語間の類似度を測定でき, 更に, 一意的な単語の意味を捉えるだけでなく, 文脈を考慮して意味確率分布を変化させ, 単語間の類似度も評価する事ができると推察している. これらにより, 文脈に依存した単語集合を生成することができ, 人間のようなアドホックな概念表現も可能になると期待している. 最終的には人間の概念生成を計算機上で実現することが本研究の目標である.

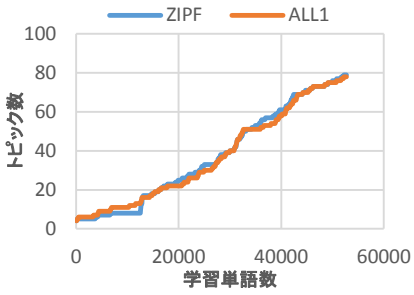

図 2　トピック数の推移

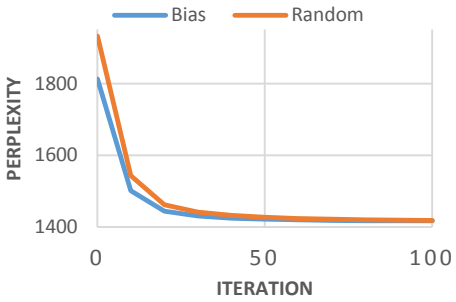

図 3　特徴量：ZIPF

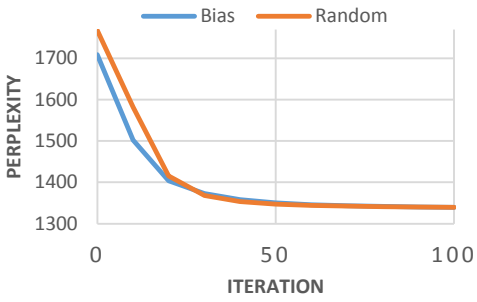

図 4　特徴量：ALL1

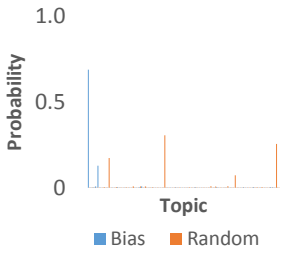

図 5　国会：ZIPF

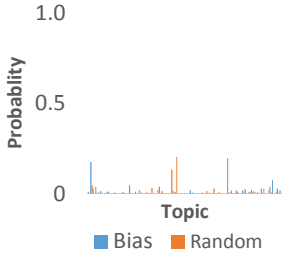

図 6　国会：ALL1

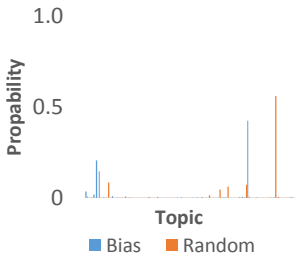

図 7　ソ連：ZIPF

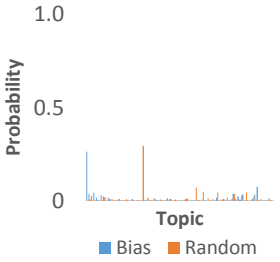

図 8　ソ連：ALL1

表 2　LS モデルから取得された周辺情報

国会

	LS 値	変換値
党	0.555	19
議	0.537	13
民	0.531	10
政	0.529	9
衆	0.529	8
国	0.527	7
案	0.525	7
会	0.524	6
院	0.522	6
員	0.520	5

ソ連

	LS 値	変換値
ソ	0.628	29
連	0.561	19
ワ	0.530	15
領	0.528	12
フ	0.527	11
国	0.527	9
ゴ	0.525	9
ル	0.523	8
ス	0.522	7
チ	0.521	7

表 3　バイアス初期化による単語類似度

国会

	類似度
国会	1.000
自民党	0.985
懇談	0.975
参院	0.970
審議	0.968
可決	0.966
辞任	0.964
党首	0.962
議員	0.960
会派	0.951

ソ連

	類似度
ソ連	1.000
東ドイツ	0.931
モスクワ	0.928
西独	0.916
ボン	0.916
共和	0.915
ジュネーブ	0.914
ホワイトハウス	0.909
日越	0.893
東独	0.892

表 4　ランダム初期化による単語類似度

国会

	類似度
国会	1.000
会派	0.960
否決	0.954
両氏	0.942
両院	0.939
議決	0.937
党首	0.933
野党	0.927
各党	0.926
会期	0.925

ソ連

	類似度
ソ連	1.000
共和	0.939
東ドイツ	0.912
クウェート	0.911
サウジアラビア	0.911
布施	0.903
チェコスロバキア	0.903
ビリニュス	0.900
ワシントン	0.899
ラトビア	0.895

参考文献

[Blei 2003] David M. Blei, Andrew. Y. Ng and Michael I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, pp.993-1022, 2003.

[Quine 1960] W. V. Quine: Word and Object, Cambridge, MA. : MIT Press, 1960.

[Teh 2006] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei: Hierarchical Dirichlet processes. Journal of the American Statistical Association, Vol. 101, No. 476, pp.1566-1581, 2006.

[Wilson 2010] Andrew T. Wilson and Peter A. Chew: Term weighting schemes for Latent Dirichlet Allocation, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, vol. 10, pp. 465–473, 2010.

[阿部 2007] 阿部慶賀, 中川正宣: 言語統計解析を用いた確率的言語知識の構築とその心理学的妥当性の検証, 認知科学, Vol.14, No.1, pp.91-117, 2007.

[今井 2007] 今井むつみ, 針生悦子: レキシコンの構築 子どもはどのように語と概念を学んでいくのか, 岩波書店, 2007.

[今井 2003] 今井むつみ, 野島久雄: 人が学ぶということ－認知学習論からの観点, 北樹出版, 2003.

[小林 1999] 小林郁夫, 古川庸一, 今井むつみ, 尾崎知伸: 機能論理プログラミングによる幼児の名詞語彙獲得のモデル化, 電子情報通信学会技術研究報告 言語理解とコミュニケーション研究会(NLC), Vol. 99, No. 387, pp.29-36, 1999.

[篠原 2007] 篠原修二, 田口亮, 桂田浩一, 新田垣雄: 因果性に基づく信念形成モデルと N 本腕バンディット問題への応用, 人工知能学会論文誌 22 巻 1 号 G, pp.58-68, 2007.

[持橋 2002] 持橋大地, 松本裕治: 意味の確率的表現, 情報処理学会研究報告, 自然言語処理研究会, 2002-NL-147, 77-84.