統計的トピックモデルの動的制御による語彙獲得精度の改善

貞光 九月　　齋藤 邦子　　今村 賢治　　松尾 義博
NTT メディアインテリジェンス研究所
{sadamitsu.kugatsu,saito.kuniko,imamura.kenji,matsuo.yoshihiro}
@lab.ntt.co.jp

1　はじめに

自然言語処理技術を用いたアプリケーションにおいて，対象ドメインに特化した辞書が必要となる場面は多い．一方で辞書の作成には大きな人的コストがかかってしまうため，可能な限りコストをかけずにドメイン依存の語彙を収集したいという要求がある．本稿で対象とする語彙獲得タスクは，ターゲットとなるドメインに属する少量の語彙集合，特に固有表現集合で表される教師データを用いて，新たな固有表現集合を獲得することを目的とする．なお，本稿では獲得対象ドメインに属する固有表現をエンティティ，初期に与えられる教師信号としてのエンティティをシードエンティティと呼ぶ．語彙獲得タスクにおいては，教師データを繰り返し処理により増加させることのできる，ブートストラップ法を用いた手法が多く提案されている [Bellare et al.2006]．しかしブートストラップ法を用いた語彙獲得の課題として，獲得される語彙の持つ意味が，元来の意味から次第に外れていくセマンティックドリフトと呼ばれる現象があり，語彙獲得精度を悪化させる大きな要因となっている．

この課題に対処すべく，近年トピック情報を用いることで語彙獲得の精度向上を行う手法が提案されている．本稿におけるトピックとは，ある文書で述べられている「政治」や「スポーツ」等のジャンルを指し，統計的トピックモデル（以下トピックモデル）を用いて自動的に推定することが可能である．貞光らは識別モデルベースのブートストラップ法に，LDA によって得られる教師なしのトピック情報を利用することで，セマンティックドリフトを抑制した [貞光 et al.2012]．また，語彙獲得に近いタスクとして，LDA を拡張した生成モデルを選択制限のモデル化に用いた手法も提案されている [Ritter and Etzioni2010]．

教師なしのトピック情報はセマンティックドリフトの抑制に効果がある一方，トピックの粒度と，獲得対象ドメインの粒度が一致しない時に，十分な効果が得られないという問題がある．そこで本稿では，シードエンティティ(正例) とシステム出力結果から選択される誤った事例 (負例) を併用して適切な粒度を持ったトピックモデルへの更新を行う．これを可能とするための新しいトピックモデルとして，インタラクティブユニグラム混合モデル (Interactive Unigram Mixture: IUM) を提案し，さらに IUM を語彙獲得に組み込むことで，語彙獲得精度が改善したことを示す．

2　教師なしトピック情報を用いた語彙獲得法とその課題

本節では，Bellare らの提案した識別モデルベースの語彙獲得法に基づき，トピック情報を導入する手法を説明する [貞光 et al.2012]．はじめに N_s 個の正例シードエンティティと，エンティティの特徴を示す語として N_a 個の正例属性が入力される．識別モデルの学習データは，正例シードエンティティ集合と正例属性集合中の 1 要素対を含む文書とし，この中から要素対を中心とした文脈素性と，教師なしトピックモデルによって得られる当該文書 d のトピック z に対する事後確率 $p(z|d)$ を素性として学習を行う．学習の際に必要な負例は，正例シードエンティティ集合とトピックを異にする事例の中から平衡して選択することで，幅広い負のトピックと文脈を識別可能なモデルが得られる．さらに，正例シードエンティティが多義性を持つ故にノイズとなる事例に対しても，トピック情報を用いることで多義性を解消することができる．

新エンティティの識別においては，正例属性を含む文書中の，正例属性周辺に出現する固有表現をエンティティ候補とし，識別モデルによるスコアリングを行う．各エンティティ候補に対する最終的なスコア関数は，エンティティ候補 e - 属性 a の対に付与されたスコア $s(e,a)$ の属性集合に関する和 $score(e) = \sum_{a \in P_a} s(e,a)$ として定義され，このスコアの高い方から N_n 種類のエンティティを獲得する．

このように教師なしトピック情報を用いることは，セマンティックドリフトを抑えるのに効果的であるが，必ずしも十分ではない．例えばドメインが「車」の場合，教師なしトピックと文脈情報だけでは「車」と

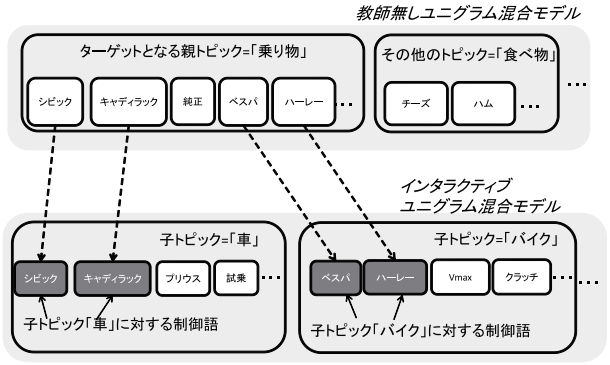

図 1: IUM の概念図. 色つき四角中の単語は制御語，色抜き四角中の単語は自動的に獲得されたトピックにおいて特徴的な単語

「バイク」を判別できない場合がある．これはモデル中のトピック粒度が，ターゲットとなるドメインに対し合致していないことが原因として挙げられ，トピック粒度の制御が課題となる．

これに対し，近年トピックモデルのトピック粒度を制御するため，人のインタラクションを許容する新たなモデルが提案されている [Hu and Boyd-graber2011]. Hu らは ディリクレ森事前分布 に基づいた インタラクティブトピックモデル (ITM) を提案し，ユーザが任意の語を機械に与えることで，トピックモデルを更新することを可能とした．本稿ではユーザから与えられる更新の手掛かりとなる語を「制御語」と呼ぶ．しかし，ITM はギブスサンプリング法に基づいてモデル更新を行うため，計算コストが大きいという問題がある．ユーザのインタラクションとトピックモデルの更新がシステム内部に組み込まれる本稿の語彙獲得のようなアプリケーションの場合，特に再学習の高速性が重要となる．そこで次節において高速な制御が可能となるトピックモデルと，それを組み込んだ語彙獲得法を提案する．

3 トピックモデルの動的制御による語彙獲得法

3.1 インタラクティブユニグラム混合モデル

本節では，トピックモデルの動的制御を高速に可能とするインタラクティブユニグラム混合モデル (IUM) を提案する．IUM はもっとも単純なトピックモデルとして知られる ユニグラム混合モデル (Unigram Mixtures: UM) [Nigam et al.2000] を拡張したものである. IUM は EM アルゴリズムに基づいてモデル更新を行うため，ギブスサンプリング法よりも高速で，Map-Reduce 等の並列分散処理にも UM 同様に適したモデルである．

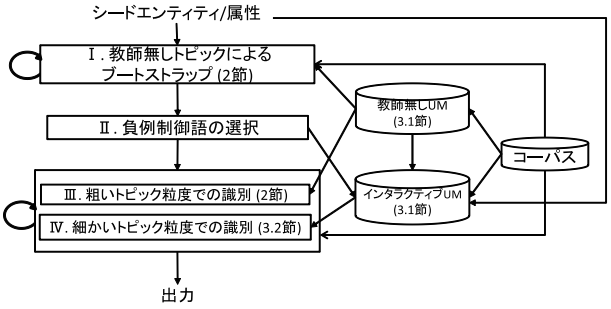

図 2: 提案手法におけるシステム構成図

IUM のモデル更新では，与えられた制御語により，ターゲットとなるトピックが複数の子トピックに細分化される．例えば，ターゲットとなる親トピック「乗り物」に対し，制御語が { シビック, キャディラック } と { ハーレー, ベスパ } という語の集合で与えられると，新たな子トピックとして「車」と「バイク」とが生成される (図 1).

以下に具体的な IUM のモデル更新方法を述べる．はじめに制御語を含む文書 d_s に対する擬似的な事後確率 $p_s(z|d_s)$ を導入する．擬似事後確率は，$n_{d_s}(z)$ を文書 d_s 中におけるトピック z に属する制御語の出現数として，$p_s(z|d_s) = n_{d_s}(z)/\sum_z n_{d_s}(z)$ という単純な式で表される．擬似事後確率は，単語レベルの教師情報を，文書レベルにまで伝播させたものと見なせる．IUM の学習において，制御語を含む文書に関し，通常の事後確率の代わりに擬似事後確率を用いるのみで，一般の EM アルゴリズムと同様の学習が可能となる．ただし，擬似事後確率は非常に粗い推定値であるため，E ステップで求められる事後確率との線形補間を施して用いる．

3.2 インタラクティブユニグラム混合モデルを用いた語彙獲得法

本節では IUM の語彙獲得への適用法について述べる．なお，本節の内容をシステム構成図として図 2 に示している．はじめに，通常の教師なしトピックモデルを用いた語彙獲得 (図中 I) の結果出力された新エンティティ集合の中から，少量の負例エンティティ集合 E_{IN} を選択する (例:「車」ドメインに対し「ハーレー, ベスパ」)(図中 Ⅱ). この負例選択は自動的に行う手法 [McIntosh and Curran2009]，もしくは人手で選択する手法のいずれでもよいが，本稿はトピックモデルの制御に注目しているため，人手によって数個の負例エンティティと，負例属性＝負例トピック名 “C_{IN}” (例：「車」トピックに対し 「バイク」トピック) を与えることとした．

次に IUM を用いたトピックモデルの動的制御を行う．負例エンティティと負例属性は，正例エンティティ・属性と共に，ターゲットとなる親トピック z_p に対する各子トピックの制御語として用いられる．なお z_p はシードエンティティの帰属するトピックの度合いによって自動的に決定される．また，IUM モデル更新中の E-step において，各文書に対する新たな子トピックの事後確率が付与されるため，これを識別モデルの素性としてそのまま用いることができる．

更新されたトピック情報をより効果的に用いるため，2 段階での識別を行う．第一段階では，元の語彙獲得と同様に，教師なしトピックモデルによる識別を行い，粗い粒度の語彙選択を行う (図中 Ⅲ)．第二段階では，IUM から得られた更新後のトピック情報を用いて，第一段階の識別で正例に選別されたものに対し，さらに細かい粒度での識別を行う (図中 Ⅳ). 予備実験において，2 種の粒度のトピックモデルから獲得される素性を同時に用いて識別した場合，細かい粒度の識別に偏ってしまい，粗い粒度の識別が適切に行われない傾向が見られたのに対し，2 段階処理にすることでこの悪影響は軽減されることが確認されたため，本手法を採用することとした．

本節の最後に，先行研究との比較を行う．我々に最も近い手法として，McIntosh の手法が挙げられる [McIntosh2010]．彼女は自動選別された負例エンティティ自体を分布類似度に基づいてクラスタリングし，語彙獲得の精度改善のために用いている．また，Vyas らはエンティティと素性の類似度に関する独自の尺度に基づいてエンティティを修正し，獲得精度を向上させている [Vyas and Pantel2009]. しかし我々の知る限り，提案手法はトピックモデルを動的制御することで語彙獲得を行うはじめての研究である．洗練されたトピック情報を用いることで，教師なしトピック情報を用いる場合よりも精度を改善することができ，負例エンティティのみを用いる場合と比しても，単語そのものに加えトピックを捉えることによるロバストな獲得モデルの更新と，精度改善が期待できる．

4 実験

4.1 実験条件

本稿での実験用コーパスとして，2008 年 5 月にクロールした日本語ブログ 3000 万記事を用いた. その他の詳細な実験条件は [貞光 et al.2012] に準ずる．獲得対象ドメインには比較的粒度の細かい 「車名」「ドラマ名」「サッカーチーム名」 の計 3 ドメインを選択した (ドメインとトピックと区別するため，各ドメインには「名」を付した)．また，各種調整パラメータは $N_e = 10, N_a = 10, N_n = 100$, 負例制御語のトピック数=2, 負例側制御語数はトピック毎に 2 単語 とした．ただし,「サッカーチーム名」に対してはより難しい状況を仮定し， $N_a = 4$，負例制御トピック数 $= 1$ とした．いずれのドメインにおいても 10 回のブートストラップイテレーションの後，計 1000 個の新エンティティを得た．識別モデルとしては SVM の 2 次多項式カーネルを用い，教師なし UM と LDA は，それぞれ 100 混合で学習した．IUM の学習における子トピックの総数は，負例制御トピック数よりも余剰を持たせ 5 とした．

表 1: 語彙獲得精度の比較．太字は提案手法とベースライン手法の差が，二項検定において $P < 0.05$ で有意であったもの，斜字は $P < 0.1$ で有意であったもの．

手法	車名	ドラマ名	サッカーチーム名
1. LDA(BL)	0.531	0.524	0.528
2. UM(BL)	0.387	0.676	0.542
3. 文脈のみ更新 + 2. (BL)	0.397	0.658	0.522 (503/963)
4. トピック更新 + 3. (提案法)	**0.636** (517/813)	**0.713**	*0.564* (558/989)

実験は以下の 4 つの実験設定で行った．はじめの 3 手法がベースラインに相当する．1 つ目は教師なし LDA を用いた結果，2 つ目は前述の LDA の代わりに教師なし UM を用いたもの，3 つ目は教師なし UM に加え，制御語を用い，文脈情報に関してのみ識別モデルを再訓練したもので，これは次の IUM を使う場合の純粋な効果を見るための比較手法である．4 つ目が提案手法で，3 つ目で用いた制御語から IUM を学習し，識別モデルに利用した場合の結果である．獲得された各エンティティは 2 名の評価者により検索エンジンを参照しながら 正解 または 不正解 でラベル付けされた．このうち 1231 個のエンティティについては二重チェックを行い，その κ 値は 0.843 であった．

4.2 実験結果

表 1 は各手法毎の語彙獲得精度を表している．なお，スコア値の閾値 ($score(e) > 0$) を満たすエンティティが 1000 語に満たない場合については，獲得できたエンティティ総数を括弧内に示している．はじめに 2 種類の教師なしトピックモデルを用いた場合のベースライン手法を比較する．UM を用いた場合，LDA よりも “ドラマ名” では改善しているものの，“車名” では悪化している．また，UM を用いた場合は，全体的に LDA よりも精度のばらつきが大きいことが読み取れ，これは UM が LDA よりも過適応しやすいトピックモデルであることが原因の 1 つと考えられる．次に 3 番目の制御語を文脈についての学習にのみ用いた場合，効果はほとんど見られなかった．これは細かい粒度でのエンティティ分類においては，正例と負例の文脈が

表 2: 提案手法により獲得されたエンティティとモデル化されたトピック

獲得エンティティ (上:車名，下:サッカーチーム名)	シード及び制御語のトピック	制御語 (UM では括弧内にシードエンティティを記述)	各トピックの特徴語
ベースライン (UM): シルビア, ハーレー, カブ 90, E700	正例トピック: z_p=車	(シードエンティティ =シビック，カローラ，etc.)	塗装, 接着, プラグ, 純正
提案手法 (IUM): シルビア, 117 クーペ, nubi250 (カーナビ)	正例制御トピック: C_P=車	シードエンティティと同じ	ターボ, 車, タイヤ, 試乗
	負例制御トピック 1: C_{IN_1}=バイク	ハーレー, CB400	バイク, プラグ, ボルト, クラッチ
	負例制御トピック 2: C_{IN_2}=列車	E700, E531 (列車名)	稼働, ガンプラ, プラモ, バンダイ
ベースライン (UM): 中日, A. マドリード, ジャイアンツ	正例トピック: z_p=サッカー	(シードエンティティ (浦和レッズ, ローマ,etc.)	チェルシー, 投手,安打, 失点
提案手法 (IUM): A. マドリード, マンチェスター C,Red Wings	正例制御トピック: C_P=サッカー	シードエンティティと同じ	マン U,DF, FW, FC 東京
	負例制御トピック: C_{IN}=野球	ジャイアンツ, タイガース	安打, 投手, 回表, 四球

似てしまうことが原因と考えられる．一方提案手法では，全てのドメインにおいて有意に精度を改善できていることが確認できた．この精度改善は，LDA と UM 間のトピックモデルの選択よりも大きな効果をもたらしており，IUM が UM をベースとしつつも，精度の揺れの少ないロバストな手法と言える．

最後に提案手法が有効に機能していることを定性的に評価する．表 2 は，UM と IUM を用いた手法に関し，1 列目に各手法で獲得されたエンティティ，2 列目にシード語と負例制御語のトピック，3 列目に各トピック毎のエンティティ, 4 列目に UM と IUM がモデル化した各トピックの特徴語を示している．ここでの特徴語とはトピック z の特徴を表す単語を指し，$p(v|z)/p_{uni}(v)$ の高いものから取得している．また下線の付与されたエンティティと特徴語は，意図したドメイン/トピックに対し誤りと判断されたものである．

まず，表の 1 列目のエンティティ獲得結果から，提案手法では「カブ９０」等，負例制御トピック中の負例の混入が減っていることが読み取れる．ただし，制御語として与えられないトピックに属する「nubi250」等の誤りは残されている．また，表の 4 列目のトピック特徴語が，与えられた制御語に近い語であることから，IUM が意図通りに学習されたことが読み取れる．ただし,「車名」ドメイン中,「車」及び「バイク」トピックに関する特徴語は適切に抽出されているのに対し,「列車」トピックにおいては想定していない「プラモデル」に関する特徴語が得られた．これはユーザが「プラモデル」トピックの解釈を誤ったものと解釈できる．このようにユーザの想定外のモデリングが行われた場合でも，IUM は「E700」のような誤ったエンティティを排除することに成功し，エンティティ獲得精度を向上できていることから，提案手法のロバスト性を示す一つの事例となっていると考える．

5 おわりに

本稿では，任意の単語による制御を介してトピックモデルを動的に更新することで，語彙獲得精度を向上させる手法について述べた．本手法を可能とするために，少量の制御語を加えるだけで適切にモデル更新ができる，インタラクティブユニグラム混合モデルを提案した．実験において，教師なしのトピックモデルを使う場合に比べ，提案手法が語彙獲得精度を改善することを示した．

今後の課題として，自動的に負例を発見する手法との結合を行うことが挙げられる．また，ユニグラム混合モデル以外のトピックモデルに対しても，アプリケーションの目的に沿う，高速な動的制御を可能としていきたい．

参考文献

[Bellare et al.2006] Kedar Bellare, Partha P. Talukdar, Giridhar Kumaran, Fernando Pereira, Mark Liberman, Andrew McCallum, and Mark Dredze. 2006. Lightly-supervised attribute extraction. In *Proc. of NIPS Workshop*.

[Hu and Boyd-graber2011] Yuening Hu and Jordan Boyd-graber. 2011. Interactive Topic Modeling. In *Proc. of the ACL-HLT 2011*, pages 248–257.

[McIntosh and Curran2009] Tara McIntosh and James R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proc. of the ACL-AFNLP 2009* , pages 396–404.

[McIntosh2010] Tara McIntosh. 2010. Unsupervised discovery of negative categories in lexicon bootstrapping. In *Proc. of the EMNLP 2010*, pages 356–365.

[Nigam et al.2000] Kamal Nigam, Andrew K Mccallum, Sebastian Thrun, and Tom Mitchell. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103–134.

[Ritter and Etzioni2010] Alan Ritter and Oren Etzioni. 2010. A Latent Dirichlet Allocation method for Selectional Preferences. In *Proc. of the ACL 2010*, pages 424–434.

[Vyas and Pantel2009] Vishnu Vyas and Patrick Pantel. 2009. Semi-automatic entity set refinement. In *Proc. of HLT-NAACL 2009*, pages 290–298.

[貞光 et al.2012] 貞光 九月, 齋藤 邦子, 今村 賢治, 松尾 義博, and 菊井 玄一郎. 2012. トピック情報を用いたブートストラップ法に基づく語彙獲得. **言語処理学会論文誌**, 19(2):89–106.