Vol.2010-NL-199 No.12
2010/11/19

トピックと属性を用いたブートストラップ法に基づく語彙獲得

貞光九月[†1]　齋藤邦子[†1]
今村賢治[†1]　菊井玄一郎[†1]

本稿ではコーパスからの語彙獲得を行う際に，トピック情報と属性情報を用いる手法を提案する．語彙が用いられている文書のトピックをトピックモデルを用いて推定し，新たな語彙獲得の際の素性として用いる．また，語彙が共通して持つ属性を文書中から選択し，それを語彙選択の制約条件として用いる．提案手法を用いることでセマンティックドリフトを軽減し，語彙の獲得精度が向上したことを示す．

Entity Set Expansion based on Bootstrap Methods using the Topic Features and Attribute Constraints

Kugatsu Sadamitsu,[†1] Kuniko Saito,[†1]
Kenji Imamura[†1] and Genichiro Kikui [†1]

This paper proposes a bootstrapping entity set expansion method that utilizes the information of topics and attributes. In order to reduce the effect of semantic drifts, we introduce two distinctive features. The first is the topic of each document estimated by topic models. These topic features are used for the discriminative models. The second is the attributes occurring around entities. These attribute features are used as the constraints on candidate search of new entities. The experiments show that the accuracy of the extracted entities was improved.

1. はじめに

自然言語を用いた多様な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になる場面は多く存在する．例えば検索エンジンにおいて検索クエリのドメインを判定することで，検索精度の向上や検索結果のクラスタリングを行うことができる．しかしドメインに特化した辞書作成にはコストがかかり，多様なドメインを対象にするのであればさらに大きな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獲得したいという要求がある．

一方で，特定のドメインに対する要求のみでなく，ドメイン非依存の語彙獲得においても，より細かいクラスに分類した上で獲得する必要性が生じてきている．例えば，関根らの定義した拡張固有表現[12]は従来の固有表現のクラスが 8 クラスであったのに対し，200 もの細分化されたクラスを持つ．橋本らによって作成された関根の拡張固有表現に基づくラベル付きコーパスにより，機械学習による拡張固有表現抽出器の研究が始まっている[19)20)]が，コーパスにおいて付与された各ラベルの出現数にはばらつきがあり，極端に学習データの少ないラベルも存在する．コーパスから単純な学習により固有表現抽出器を構築した場合，これら低頻度のラベルについて正しく学習できないことが予想されるため，各クラス毎の直接的な辞書の拡充が必要とされる．

本稿で対象とするタスクは，ドメイン依存・非依存に関わらず，少量の教師データを用いて固有表現を含めた語彙 (特に実体のあるものを対象とするため，以下ではエンティティと呼ぶ) を拡充することである．本稿では学習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エンティティ獲得を行う．しかし，ブートストラップ法を用いたエンティティ獲得における課題として，獲得されるエンティティの持つ意味が，獲得対象のエンティティの持つ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，獲得精度を悪化させる要因となっている．

本稿では，セマンティックドリフトの緩和とエンティティ獲得の精度向上のために，トピックと属性という 2 種の情報を利用する．トピックとは文書全体を通じて存在する内容のジャンルを指し，統計的トピックモデルを用いて推定される．属性とは獲得対象とするエンティティにおいて共通して用いられる語と定義し，”has-a”，”is-a” に代表されるようなエンティティとの関係を持つ語とする．属性はエンティティと同様なブートストラップ法 (co-training 法) によって選択される．以下 2 節で関連研究とその課題，3 節でトピック情報と属性情報を用いた詳細な提案手法，4 節で実験結果について報告し，提案手法が少量のシードからのエンティティ獲得において効果があることを示す．

†1 NTT サイバースペース研究所
NTT Cyber Space Laboratories

Vol.2010-NL-199 No.12
2010/11/19

2. ブートストラップ法を用いた語彙獲得における課題

本節ではブートストラップ法の基本的な流れと，本手法によるエンティティ獲得の課題について述べる．本稿では識別器を適用したブートストラップ法を用いる．本手法では，初期に与えられるシードエンティティ(正例・負例を含む)とそれに付随する素性を元に識別モデルを学習し，学習された識別モデルに従って新たな正例・負例エンティティを獲得する．ここでの素性とは主に周辺文脈についてを指し，例えば X をエンティティとした場合，「X/の/株価/が/上昇」という文において，$f(surf. = \text{"の"}, position = X + 1) = 1$, $f(surf. = \text{"株価"}, position = X + 2) = 1$ といった素性関数 f で表わされる．得られたエンティティを学習データに加えて再度識別モデルを学習し，新たなエンティティを獲得する．この処理を必要なエンティティ数が得られるまで繰り返していく．

少量のシードからブートストラップ法によってエンティティ獲得を行う際の主な課題の 1 つに，セマンティックドリフトが挙げられる．例えば獲得対象が企業名である場合に,「NTT」と「トヨタ」をシードエンティティとして与え，エンティティ獲得アルゴリズムにより「ヤクルト」が獲得できたとする．しかし「ヤクルト」には企業名以外にも，プロ野球球団名や飲料品名といった多義性が存在するため，次のイテレーションにおいて獲得されるエンティティが「巨人」等の本来獲得対象としていたエンティティ集合ではないものに変わっていく場合がある．この現象をセマンティックドリフトと呼ぶ．

先行研究では新しいエンティティを選択する際のスコア関数を独自に定義することで，セマンティックドリフトを抑えつつ精度の高いエンティティ獲得法を提案している．[14)6)11)] これらのスコア関数は，シードエンティティの特徴となるべく近い特徴を持つエンティティに高いスコアを与えるように設計されている．スコア関数はエンティティ獲得を行う上で重要であるが，本稿ではスコア関数に用いられる特徴に注目した．我々は周辺文脈のうち特にエンティティの特徴を表していると考えられる語を属性とみなし，新しいエンティティを探索する際の制約条件として用いる．また，従来用いられてきた素性はエンティティの周辺文脈を用いたものが主だったが，本稿ではトピックモデルを用いて大域的な情報を素性として取り入れることで，エンティティが文脈中で用いられている意味を明確化する．これにより局所的情報と大域的情報を効果的に併用することが可能となる．

セマンティックドリフトを抑えるためにはシードの選択も重要な課題となる．特に，少量のシードのみを手がかりに行うエンティティ獲得では，シードによる精度への影響は大きい．Pantel らは大規模な WEB に対して，比較的単純なスコアリング関数を用いて効率的なエンティティ獲得手法を提案しており[9)]，10 個程度のシードにより十分なエンティティが得られると報告されている．一方で Vyas らはシードの選択によりエンティティ獲得の結果に影響が出ることを示している[15)]．特に多義性のあるシードが混入した場合にセマンティックドリフトが生じやすく，精度の劣化は大きいと考えられるため，Vyas らは精度を落とす可能性の高いシードを除去するアルゴリズムを提案している．これに対し，本稿ではシードエンティティ集合全体において，より多くのエンティティで共通して用いられている属性やトピックを重視することで，多義性のあるエンティティであっても，獲得対象のドメインで用いられているエンティティに限って高い信頼度を付与できる．これによりセマンティックドリフトを軽減しつつ，シードの情報を有効に活用できる手法となっている．

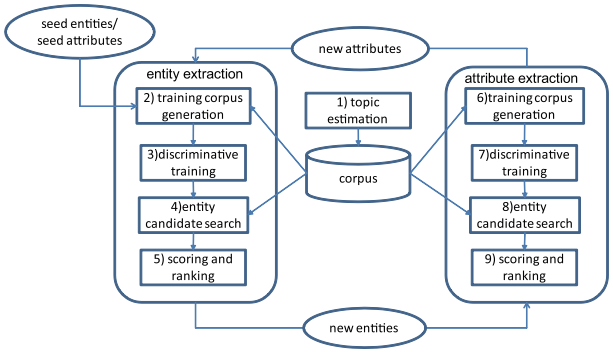

図 1 トピックと属性を用いた co-training 法
Fig. 1 The co-training methods using topics and attributes.

3. トピックと属性を用いた提案手法

本稿ではトピック情報と属性情報を用いることでエンティティ獲得精度の向上を図る．はじめに提案手法の流れを図 1 に示す．これは基本的にはエンティティと属性を交互に獲得する co-training 法である．以下では図 1 中の処理の流れを追っていく．括弧内の番号は図中の番号と一致する．

(1) [topic estimation] 文書毎のトピックをトピックモデルを用いて推定し記憶する．推定されたトピックはブートストラップ法の中で用いられる識別モデルの素性として扱われる．

(2) [training corpus generation] 正例負例それぞれについてのシードエンティティとシー

ド属性が入力として与えられた後，シードエンティティとシード属性を組み合わせをとることで，エンティティ-属性ペアを得る．正例/負例エンティティ-属性ペアを含む文を検索し，識別モデルの学習に用いる素性を抽出することでエンティティ-属性ペア毎の学習データを生成する

(3) [discriminative training] 学習データから新規エンティティ獲得用の識別器を学習する

(4) [entity candidate search] コーパスから正例属性を含む文を検索し，検索された文の中からエンティティ候補を抽出する．

(5) [scoring and ranking] エンティティ候補と属性を共に含む文集合から識別用の素性を抽出する．抽出された素性をエンティティ候補-属性ペア毎にまとめ，正規化することで１つの識別対象データとする．学習された識別器を用いてエンティティ候補-属性ペアに対する識別スコアを計算した後，エンティティ候補 e に対して信頼度 $C(e)$ を付与する．信頼度の高いエンティティを正例エンティティリストに加え，正例属性リストとの組み合わせをとることで，新しいエンティティ-属性ペアを得る．新規の負例については十分な量の文をランダムに抽出，同じ識別器で識別スコアを付与し，低いものを新しい負例エンティティ-属性ペアとする．なお，負例についてはエンティティと属性をそれぞれ制約条件として利用することはしないため，単純に識別対象をエンティティ-属性ペアとする．

(6) [training corpus generation] 新しく獲得されたエンティティ-属性ペアを含む文の検索と識別用素性の抽出を行い，学習データを生成する．

(7) [discriminative training] 学習データから新規属性獲得用の識別器を学習する．

(8) [entity candidate search] 属性候補を選択する．正例エンティティリスト中のエンティティを含む文を検索した後，検索された文集合の中から属性候補を選択する．選択されたエンティティと属性候補に対し識別用素性を抽出し，エンティティ-属性候補ペア毎にまとめ，正規化することで１つの識別対象データを作成する．

(9) [scoring and ranking] 学習された識別器を用いてエンティティ-属性候補ペアに対し識別スコアを計算した後，属性候補 a に対して信頼度 $C(a)$ を付与する．信頼度の高い属性を正例属性リストに加え，正例エンティティリストとの組み合わせをとることで，新しいエンティティ-属性ペアを得る．新規の負例については十分な量の文をランダムに抽出し，同じ識別器で信頼度を付与し，低いものを新しい負例エンティティ-属性ペアとする．

(10) (2)-(9) のエンティティ獲得と属性獲得の処理を必要なエンティティ数が収集できるまで繰り返し行う．

以降，(2)-(9) は 3.1 節で，(1) については 3.2 節で詳しく述べていく．

3.1 属性情報を用いた語彙獲得

本節では第一の提案手法である属性の利用方法について述べる．属性とは獲得対象とするエンティティ集合において，複数のエンティティが共通して関係する語 (”has-a” や”is-a” 等の関係) であるとする．例えばエンティティ「ヤクルト」と「巨人」の属性は,「監督」(has-a) や「球団」(is-a) 等となる．属性は，学習データの生成及び新しいエンティティ候補を選択する際の制約条件として用いられる．属性の効果について例を挙げて説明する.「ヤクルト」には曖昧性があり，それ単独では企業名とプロ野球球団名のどちらを指すかはわからない．しかし属性を付加した「ヤクルト-監督」と「ヤクルト-株価」とを比べると，これらが異なる「ヤクルト」を指すことが分かる．このため，もし企業名を欲している場合に「ヤクルト」がブートストラップ法において新たなエンティティとして獲得されたとしても，属性を付加することで，球団名として用いられている「ヤクルト」が学習データとして混入する可能性を低く抑えることができる．また次イテレーションにおけるエンティティ候補選択の場面でも同様に,「巨人」等の球団名を表すエンティティを候補とする可能性を低く抑えることができる．属性と素性の違いについてまとめると，素性は識別のための要素を指し，属性はエンティティを獲得するという最終目的に立てば，候補選択のための制約条件としての働きをするとともに，素性によって識別対象とされる．これらの点で属性は他の素性と異なる性質を持つ．

我々は属性を文書中から半自動的に獲得し，ブートストラップ法に用いることでエンティティ獲得精度を向上させる手法を提案する．本手法は図 1 中の (2)-(9) に示される．これは Bellare らの co-training 法による手法と基本的には同じアルゴリズムに基づいている[1)]．

以下，各処理の詳細について述べる．2. のシード属性の付与に関しては，PMI や χ^2 等の統計量を用いて自動的に獲得することもできるが，本稿ではシード属性の獲得については範囲とせず，人手で与えることとした．負例となるエンティティ-属性ペアについてはランダムにエンティティ-属性ペアを選択した後，人手で確認した後に与える．ここでもシードエンティティ及び属性に含まれないようなランダムなエンティティ-属性ペアを選択することで初期負例とした．

2,4.(及び属性抽出においては 6,8.) における訓練コーパス生成とエンティティ候補探索においては，本節の主旨である属性を以下のような制約条件として用いる．まず，2. において学習データに含まれる文は，エンティティと属性が N 単語以内に存在しなければならな

い．また，4. において候補となるエンティティは，正例属性リスト中の属性を含む文中から選択され，かつ正例属性の前後 N 単語以内に存在する固有表現または名詞のみをエンティティ候補とする．例えば「監督」が正例属性であるとして,「ヤクルト/の/新しい/監督/が/就任」という文が与えられた場合，N=3 ならば「ヤクルト」と「就任」がエンティティ候補となる．

5. 及び 9. で識別器によって付与される信頼度 $C(e), C(a)$ は以下のように定義する．

$$C(e) = \sum_{a \in A} S(e, a, +1)$$

$$C(a) = \sum_{e \in E} S(e, a, +1)$$

ここで e, a はある 1 つのエンティティ及び属性を表し，A, E は正例属性/エンティティリストを示す．$S(e, a, +1)$ は e と a が共起する場合の正例としての識別スコアを表している．本稿では識別器から出力される識別スコアを直接用いる．実験では SVM を識別器に用い，識別スコアには SVM から出力される識別平面からの距離を利用した．識別スコアから全ての属性 (/エンティティ) についての和をとることで，共通した属性 (/エンティティ) と共起しやすいエンティティ(/属性) に高い信頼度が付与される．

3.2 トピック情報を用いた語彙獲得

第二の提案手法はトピック情報を用いた手法である．トピック情報を用いることで，周辺文脈より広い文書全体からの大域情報を反映することができる．例えばエンティティ「ヤクルト」は企業名としても球団名としてもありうる．さらに，属性として「広報」が与えられた場合でもなお曖昧性は残る．この時，文書が次のように与えられているとする.「18 日の夜，ヤクルトの広報担当者が取材に対してコメントを発表した．18 日の試合で途中退場した Y 選手は，診断の結果軽いねんざと診断された，とコメントは伝えている.」文書全体を読めば，このエンティティが「球団名」を指していることが明らかである．我々は，文書全体を通して存在するトピックを，エンティティ/属性識別の際の素性として用いる．本稿で用いるエンティティ・属性の信頼度は，識別器が出力するスコアに基づいているため，柔軟に素性を加えることができる利点を利用する．

文書の背景にあるトピックを利用する場合，文書に対して明示的にトピックラベルが付与されているデータであらば，そのラベルを直接トピック情報として用いることができるが，全ての文書にトピックラベルが付与されているわけではない．トピック情報の取得をラベル無しの文書からも自動的に行うため，本稿では文書のトピックと単語との関係をモデル化するトピックモデルを用いる．トピックモデルは，文書のトピックと関連の強い単語に高い確率を付与することで，文書をより緻密に表現できるモデルである．例えばある文書のトピックがスポーツであるならば,「サッカー」といったスポーツに関する単語が出現しやすく,「国会」といった単語が出現しにくい，といった大域的情報を扱うことができる．

トピックモデルの種類としては，Latent Dirichlet Allocation(LDA)[2] をはじめとして多くのトピックモデルが提案されているが，本稿では文書のトピックの扱いを単純にするため，文書に対して 1 つのトピックを仮定するユニトピックモデルの一種である混合ディリクレモデルを用いる[16]．混合ディリクレモデルとは，$P_{Mul}(d|\mathbf{p})$ を文書 d を評価する多項分布とした時に，多項分布パラメータ $\mathbf{p}$ についての事前分布としてディリクレ分布の線形和，$P_{DM}(\mathbf{p}; \lambda, \alpha) = \sum_{\mathbf{z}} \lambda_{\mathbf{z}} \mathbf{P_{Dir}}(\mathbf{p}; \alpha_{\mathbf{z}})$ をおいた合成分布を指す．ここで $p(d)$ は文書 d の確率，$z \in 1, \ldots, Z$ は隠れ変数で，1 つの z が 1 つのトピックを示す．λ_z は隠れ変数 z に対する事前確率を表し，$\sum_z \lambda_z = 1$，また，$\alpha = \alpha_{\mathbf{1}}, \ldots, \alpha_{\mathbf{Z}}$ は隠れ変数 z における $\mathbf{p}$ 上のディリクレ分布のパラメータを表す．混合ディリクレモデルを学習する際には，繰り返し最適化手法の 1 種である EM アルゴリズムを用いて学習する．

次にトピックモデルの利用方法について述べる．コーパスから学習された混合ディリクレモデルは，パラメータとして α, λ を保持しているため，学習データに含まれる文書と学習データに含まれない文書の両方に対して，隠れ変数 z に対する事後確率 $p(z|d) \propto p(d, z)$ を計算することが可能となる．この事後確率は各文書におけるトピックについての重みであるとみなせるため，エンティティ/属性に対する識別及び識別モデル学習において，直接素性として用いることとする．

3.3 関 連 研 究

属性は関係抽出のタスクでは早くから着目されている概念であり，代表的なものに PMI を用いてスコア関数を定義した Pantel らの研究が挙げられる[10]．Bellare らは Espresso を変形した，エンティティと属性による co-training 法を提案している[1]．我々はこれらの研究で獲得対象とされる属性を語彙獲得においても用いることで，獲得精度を高めることができるのではないかと考えた．本稿での属性の利用方法は Bellare らのものに近いが，Bellare らがエンティティと属性の両方の獲得を目的としていたのに対し，我々は属性の獲得は目的とせず，あくまでエンティティ獲得に目的を絞っている．そのため属性を網羅的に獲得する必要がなく，十分に信頼できる少量の属性のみを用いれば良いという利点がある．

また，3.2 節では属性と同時にトピック情報を用いることでセマンティックドリフトを軽減することを提案した．ここで用いるトピックモデルは一種のクラスタリングモデルであ

る．エンティティ獲得にクラスタ情報を用いた先行研究として，Paşca らの研究が挙げられる[8]．彼らの手法では出現したクエリの周辺パターンをクラスタリングしているのに対し，我々は文書全体からトピックを推定する点で，より広域な情報を取り入れることができる．

文書以外のリソースとして，クエリログを使ったエンティティ獲得の研究も進められている．小町らはクエリログ中に共起する単語をエンティティ及び属性とみなし，ブートストラップ法に基づくエンティティ獲得法の提案を行っている[17]．クエリログを使った他の手法としては Sekine らの研究[13] や Paşca らの研究[8] が挙げられる．しかし，クエリログ単独ではトピックのような大域的な文脈を考慮することが不可能であり，また，非公開で一般的に入手が困難なリソースである．我々はこれらの観点から文書をリソースとして用いることとした．

4. 実　　験

本節では提案手法の有効性を示すために，少量のシードエンティティからの新規エンティティ獲得精度を手法毎に比較し，その結果についての考察を行う．

4.1 実験条件

コーパスは毎日新聞 2001 年から 2007 年版から，計 713,012 記事を用いた．単語及び固有表現を処理単位としており (以後簡単のため固有表現を含めて単語と呼ぶ)，形態素解析には JTAG[5] を，IREX 定義に基づく固有表現抽出器には最小誤り分類基準に基づく CRF を用いた[21]．識別器には SVM^{light} を，カーネルには 2 次の多項式カーネルを用いた[⋆1]．トピックモデルである混合ディリクレモデルは同毎日新聞 7 年分のうち，5 回以上出現した単語から 10 トピックの混合ディリクレモデルをモデル化した．モデル学習に用いた文書全てについて各トピックの事後確率を保持し，各文書においてトピックの事後確率の高い順に，その和が 0.5 を超えるまでのトピックを素性として用いる．これは，全ての z についての $p(z|d)$ を出力するよりも適切な閾値以上の値の z についてのみ素性化した方が効果を得やすいことが予備実験の結果分かったためである．

本稿では「企業」(company) と「書籍」(book) の 2 ドメインについて実験を行った．実験で用いたエンティティ及び属性のシードを表 1 に示す．セマンティックドリフトの影響をいかに緩和できるか評価するため，book と company のいずれに対しても，セマンティックドリフトを生じやすいと考えられるエンティティ(「ヤクルト」(プロ野球球団名),「ハリー・ポッターと賢者の石」(映画名)) を含んでいる．さらに，属性としては company1 で比較的セマンティックドリフトを起こしにくい属性 (「決算」) を選択し，company2 では比較的セマンティックドリフトを起こしやすい属性 (「好調」) を選択している．これは実際の適用場面において付与される属性の影響を知るためのものである．また，シードエンティティ数を 3 個と少なく設定しているのは，セマンティックドリフトの影響を拡大して見るためであり，シードエンティティ数を増やすことで，本実験以上の獲得精度が得られることが期待される．エンティティと属性を用いた co-training 法を行う場合には，シードエンティティとシード属性の全ての組み合わせをシードとする．全ての実験において 1 イテレーション毎に獲得するエンティティ数を 10，属性の獲得数を 1 とし，10 イテレーション繰り返して 100 エンティティを獲得するまでを処理を行った．また正例とのデータ量を平衡に保つようにランダムにサンプリングした．[⋆2]

以上の条件において，以下の 3 種類の手法についての実験を行った．

- BaseLine:エンティティのみによる self-training 法
- BL+attr.:属性を加えたエンティティとの co-training 法
- BL+attr.+topic:属性とトピックを加えた co-training 法

ベースラインで用いる self-training 法とは，獲得対象に対する識別器の学習と，識別器によって新たに得られた識別結果を学習データとして取り入れることで，繰り返し学習を行うブートストラップ法である．最終的な獲得対象であるエンティティにのみ注目して処理を行っていく点が co-training 法とは異なっている．なお，トピックのみをベースラインに加えた実験は条件に含んでいない．これは「ヤクルト」がコーパス中で野球のトピックにおいて多く使われているために，シード全体として見ても野球側に大きな重みを与えてしまうため，効果は少ないと考えられるためである．

表 1　シードエンティティ及びシード属性
Table 1　The initial seed sets of entities and attributes.

	Entities	Attributes
company1	東芝，ヤクルト，ファンケル	株価，社長，決算
company2	東芝，ヤクルト，ファンケル	株価，社長，好調
book	国家の品格，バカの壁，ハリー・ポッターと賢者の石	新書，実用書，ベストセラー

⋆1 http://svmlight.joachims.org/

⋆2 エンティティが出現する文書数をあらかじめ記憶しておき，正例エンティティが入力された後に，そのエンティティの持つ文書数と同程度のものを負例エンティティとして選択した．co-training 法の場合には，負例エンティティは上記の基準で選択し，負例属性はランダムに決めた．

Vol.2010-NL-199 No.12
2010/11/19

エンティティ候補として選択される単語は，既に属性リスト中に存在する属性の出現位置から 1〜2 単語を挟んで出現している名詞及び固有表現に限った (属性候補についても同様)．素性にはエンティティと属性の前後 2 単語までの各単語の出現位置及び表層形と品詞の組み合わせを用いた．エンティティ自体については品詞のみを素性として用い，属性については表層と品詞の両方を素性に用いた．ただしいずれの素性も出現数 5 回未満の素性はカットオフしている．

評価は 1 名による評価を行った．いずれのドメインについても獲得されたエンティティをクエリとして WEB を検索し，ヒットした上位 20 件中に各ドメインの文脈においてエンティティが使用されていれば正解であるとした．[⋆1]

表 2 獲得されたエンティティの精度
Table 2 The accuracy of extracted entities.

	company1	company2	book
BaseLine	26%	26%	77 %
BL+attr	63%	26%	87 %
BL+attr+topic	66%	70%	91 %

表 3 獲得エンティティと属性 (company1)
Table 3 The extracted entities and attributes (company1).

BL+attr.		BL+attr.+topic	
ent.	attr.	ent.	attr.
サムスン	会議	**ＩＢＭ**	１０年
住友不動産マンション事業本部 (-1)	出発	**三菱地所**	発注
コロニー・キャピタル	握手	**カネボウ化粧品**	中間決算
経済産業省原子力安全・保安院 (-1)	力投	**ＴＤＫ**	将来
みずほ証券	署名	**東京ガス**	２年目
富士通	退任	**ヤマダ社**	パンフレット
日本郵政公社 (-1)	辞任	**グ社**	生損保
パ・リーグ (-1)	さん	**ＮＥＣソリューションズ**	不良
国民新党 (-1)	ちゃん	旧長銀 (-1)	事件
ＡＰ (-1)	元首相	**南証券**	部下

⋆1 本実験における企業の定義は，営利目的の法人とした．愛称や略称等も正解としているが，特殊法人や非営利組織は不正解とした．書籍については雑誌等のように定期的に刊行されるものは不正解としている．

4.2 実験結果と考察

各ドメインについて 3 手法によって獲得されたエンティティについての精度を表 2 に，実際に獲得されたエンティティと属性を表 3,4,5 に示す．表 3〜5 では属性を入れた”BL+attr.” とさらにトピックを入れた”BL+attr.+topic” の 2 手法について比較している．エンティティと属性は獲得された順番に上から列挙している．各イテレーションにおいて獲得される 10 エンティティのうち，信頼度の最も高いと判断されたエンティティのみを示している．各エンティティの後に (-) が付いているものが誤りとされたエンティティであり，太字で記

表 4 獲得エンティティと属性 (company2)
Table 4 The extracted entities and attributes (company2).

BL+attr.		BL+attr.+topic	
ent.	attr.	ent.	attr.
ＮＴＴドコモ	さん	**三菱**	社名
オリックス	委員長	**カシオ**	カード
毎日新聞	会長	衆院通信委員会 (-)	１９７０年
慶大 (-)	議員	極東 (-)	説明
中大 (-)	議長	**トモテック**	将来
公明 (-)	局長	**四谷ラウンド**	不明朗
公明党 (-)	長官	**Ｚ社**	来月初め
通産 (-)	部長	通産省 (-)	準備
日銀 (-)	時代	ＯＰＥＣ事務局 (-)	１１月
最高裁 (-)	たち	**大王製紙**	後手

表 5 獲得エンティティと属性 (book)
Table 5 The extracted entities and attributes (book).

BL+attr.		BL+attr.+topic	
ent.	attr.	ent.	attr.
清貧の思想	著書	**頭の体操**	著書
下流社会	０５年	**ガルブレイス　わが人生を語る**	旧作
プラハの春	童話	**生命の歌**	佐藤俊樹
資本蓄積論	一方	**防人と衛士**	著者
秋の叙勲 (-)	０３年	**戦後日本共産党私記**	８０年
美しい夏キリシマ (-)	０１年	**二つのコリア**	前作
煙か土か食い物	０６年	**谷川の２１世紀定跡２**	近著
春の叙勲 (-)	０４年	**霊長類と哲学者**	精神
蹴りたい背中	０２年	**ハンセン病療養所**	１３日
肩ごしの恋人	今年	**ナザレのイエス**	５日

されたものが正解のエンティティである．

まず company のドメインから結果を見ていく．company1,2 のシードエンティティに含まれる「ヤクルト」は，企業名としてよりも球団名として出現しやすいため，セマンティックドリフトを非常に起こしやすい条件となっている．そのためベースラインとなるエンティティ単独での self-training を用いた場合，早いイテレーションの段階でセマンティックドリフトが生じており，結果獲得されたエンティティの精度も低かった (26%)．company1 の場合，属性を入れることで (BL+attr.) ベースラインに比べ精度が大幅に改善している．これは球団としての文脈で用いられた「ヤクルト」をある程度除外し，セマンティックドリフトを軽減できたためと考えられる．表 3 に挙げた実際のエンティティの精度は悪いが，企業の中の部署を指していたり (住友不動産マンション事業本部)，本実験では不正解とした国営企業 (日本郵政公社) や非営利組織 (AP 通信) が含まれているため，獲得すべき対象としては大きく間違ってはおらず，各イテレーションで得られた他のエンティティの精度は高かった．一方，属性単独によってトピックを絞ることと似た働きをしたために，”BL+attr.” と”BL+attr.+topic”の間に差はなかった．

シード属性の１つを「決算」から「好調」に変更した company2 の場合，球団の文脈で用いられる「ヤクルト」が学習データ及びエンティティ候補として出現する可能性が高くなる．そのため，属性を加えた”BL+attr.”では，トピックを限定する効果のない属性（＝ジェネリックな属性．例えば，人称接尾辞として使われる名詞の「さん」）を獲得してしまい，獲得すべき対象がシフトし,「議員」「長官」といった無関係な属性を獲得するに至っている．結果，4 イテレーション以降全てにおいて誤ったエンティティを抽出している (26%)．これに対し，トピックを加えた”BL+attr.+topic”はトピックを限定することで，ジェネリックな属性の獲得数を減らし，精度を改善することができた (70%)．“BL+attr.+topic” では企業名と無関係な単語を属性とすることはなく，これが”BL+attr.” に比べて精度が高い理由と考えられる．

次に book のドメインについて見ると，company と比べ精度が全体的に高くなっている．しかし，ベースラインにおいて獲得されたエンティティを個別に見ると,「千と千尋の神隠し」のように，書籍として実際に存在はするものの，映画等他のドメインにおいてより多く使われやすいものが多く含まれていた．属性を用いた場合 (“BL+attr.”) はベースラインに比べ精度向上しているが (87%),「０１年」や「０７年」といったジェネリックな属性が獲得されたため,「秋の叙勲」や「春の叙勲」といったエンティティへと獲得対象がシフトしている．また，表に挙げたもの以外の正解とされたエンティティの中にも，ベースライン同様に「千と千尋の神隠し」のようなエンティティが獲得されていた．これに対しトピックを加えた場合 (“BL+attr.+topic”)，書籍としての特徴を表す属性や，特定のエンティティのみと共起する属性が獲得されており，セマンティックドリフトの影響が減少した (91%)．しかし，特定のエンティティにのみ結びつくような属性の獲得は，広範なエンティティを獲得する上では適していない場合もあるという点や，イテレーションが進んでいくと「13 日」や「5 日」といったセマンティックドリフトを起こしやすい属性が獲得されることからも，一般的に出現しやすい属性候補に対してペナルティを課すようなスコア関数に洗練する必要があると考える．

5. まとめと今後の課題

本稿ではトピック情報とエンティティ周辺に現れる属性情報を手がかりとしたエンティティ獲得法を提案した．属性は学習データ生成及び識別候補選択の制約条件として用い，トピック情報は事後確率による素性として与えることで，エンティティ獲得の精度を改善させることができた．また，トピックを素性として導入したことでシード属性の揺れにも頑健になった．今後はより安定して高精度にエンティティを獲得できるアルゴリズムに洗練していくことが課題である．そのためには，よりトピックと属性に適したスコア関数を設計する必要がある．さらにシード属性や負例の扱いも改善点として挙げられる．シード属性については PMI や χ^2 値によって取得する方法が考えられる．負例について本稿ではランダムに選択していたが，適切に負例を選択することは識別器の精度に直接影響する．Liu ら[7] や Denis ら[4] は，正例のみが存在する学習データにおいて負例を自動生成する手法について研究しており，これらの手法の適用を考えていきたい．

別の方向性としては，ブートストラップ法のグラフ理論的な解釈がある．小町らはエンティティ獲得のアルゴリズムをグラフ理論に基づいて解釈し，グラフカーネルの一種であるラプラシアンカーネルを導入することで，その性能が改善したと述べている[18]．トピックモデルを扱えるグラフ理論に基づく枠組みとしては，Cohn ら提案した PHITS があり[3]，彼らの考えを導入することができれば，より高い精度のエンティティ獲得法を構築できると考える．また，実験に用いた新聞コーパスはドメインが限られてしまうため，獲得できるエンティティの種類には限界があるため，多様なエンティティを含む WEB のような大規模データに対する本手法の適用を行っていきたい．

参 考 文 献

1) Bellare, K., Talukdar, P., Kumaran, G., Pereira, F., Liberman, M., McCallum, A. and Dredze, M.: Lightly-supervised attribute extraction, *Proceedings of the Advances in Neural Information Processing Systems Workshop on Machine Learning for Web Search* (2006).
2) Blei, D., Ng, A. and Jordan, M.: Latent dirichlet allocation, *The Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
3) Cohn, D. and Chang, H.: Learning to probabilistically identify authoritative documents, *Proceedings of the 17th International Conference on Machine Learning*, pp. 167–174 (2000).
4) Denis, F., Gilleron, R. and Tommasi, M.: Text classification from positive and unlabeled examples, *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (2002).
5) Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer using Word Co-occurrence-JTAG, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Proceedings of the Conference*, pp.409–413 (1998).
6) Ghahramani, Z. and Heller, K.: Bayesian sets, *Proceedings of the Advances in Neural Information Processing Systems* (2005).
7) Liu, B., Lee, W., Yu, P. and Li, X.: Partially supervised classification of text documents, *Proceedings of the 19th International Conference on Machine Learning*, pp. 387–394 (2002).
8) Paşca, M. and Van Durme, B.: Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pp.19–27 (2008).
9) Pantel, P., Crestan, E., Borkovsky, A., Popescu, A. and Vyas, V.: Web-scale distributional similarity and entity set expansion, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp.938–947 (2009).
10) Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp.113–120 (2006).
11) Sarmento, L., Jijkuon, V., de Rijke, M. and Oliveira, E.: More like these: growing entity classes from seeds, *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pp.959–962 (2007).
12) Sekine, S.: Extended named entity ontology with attribute information, *Proceedings of the 6th International Language Resources and Evaluation* (2008).
13) Sekine, S. and Suzuki, H.: Acquiring ontological knowledge from query logs, *Proceedings of the 16th international conference on World Wide Web*, pp.1223–1224 (2007).
14) Thelen, M. and Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern contexts, *Proceedings of the 2002 conference on Empirical methods in natural language processing*, pp.214–221 (2002).
15) Vyas, V., Pantel, P. and Crestan, E.: Helping editors choose better seed sets for entity set expansion, *Proceeding of the 18th ACM conference on Information and Knowledge Management*, pp.225–234 (2009).
16) 貞光九月，三品拓也，山本幹雄：混合ディリクレ分布を用いたトピックに基づく言語モデル，電子情報通信学会 D-II，Vol.J88-D-II, No.9, pp.1771–1779 (2005).
17) 小町　守，鈴木久美：検索ログからの半教師あり意味知識獲得の改善，人工知能学会論文誌，Vol.23, No.3, pp.217–225 (2008).
18) 小町　守，工藤　拓，新保　仁，松本裕治： Espresso 型ブートストラッピング法における意味ドリフトのグラフ理論に基づく分析，人工知能学会論文誌，Vol.25, No.2, pp.233–242 (2010).
19) 橋本泰一，乾　孝司，村上浩司：拡張固有表現タグ付きコーパスの構築，情報処理学会自然言語処理研究会，No.113, pp.113–120 (2008).
20) 橋本泰一，中村俊一：拡張固有表現タグ付きコーパスの構築 - 白書，書籍，Yahoo!知恵袋コアデータ，言語処理学会第 16 回年次大会，pp.916–919 (2010).
21) 齋藤邦子，鈴木　潤，今村賢治：CRF を用いたブログからの固有表現抽出，言語処理学会第 13 回年次大会，pp.107–110 (2007).