インスタンス抽出パターンの拡張による語彙獲得

白井 尊昭　　新保 仁　　松本 裕治
奈良先端科学技術大学院大学
{takaaki-s, shimbo, matsu}@is.naist.jp

1　はじめに

本研究の目的は，与えられた複数のキーワード（語彙）に関連する他のキーワードを順序付きで列挙することである．例えば「United States」と「China」が入力として与えられた場合，国名に関する他のキーワード（「Japan」など）を上位に列挙するのが望ましい．この研究は，リスティング広告への出稿などに有用である．リスティング広告とは，検索エンジンに入力したキーワードに連動して表示される広告で，広告主が宣伝したい商品に関連する検索キーワードへ入札することにより行われる．その際，商品に関連する大量のキーワードを人手で列挙して入札するのは現実的ではない．そのため，少量のキーワードを入力した時に関連するキーワードを大量に列挙できる仕組みが必要である．

このようなタスクを語彙獲得（**Entity Set Expansion**）と呼び，様々なアプローチにより研究されてきた．中でも，ブートストラップ法による語彙獲得は広く行われている．ブートストラップ法ではまず，獲得したい意味クラス（例: 国名）のインスタンス（語彙）をシードとして与える．次に，シードの文脈をコーパスから列挙し，シードとの共起頻度が高い文脈をパターンとして獲得する．最後に，パターンとの共起頻度が高いインスタンスを新たに獲得する．これら一連の処理を反復することで，大量のインスタンスを獲得できる．

ところが，従来の研究では，1つのパターンからは限られたインスタンスしか獲得できない．例えば，パターン「＜名詞＞plays for Giants」は Giants 所属の野球選手としか共起しない．そこで，提案手法では，より多くのインスタンスと共起するようにパターンを定義した．さらに，インスタンスのスコアを計算する際，パターンとの共起頻度だけでなく,「シードと共起している項」と「インスタンス候補と共起している項」の類似度も計算に用いた．ここで，パターンは例えば「＜名詞＞plays for＜項＞」であり，Giants 所属の野球選手に限らず，スポーツ選手を表す多くのインスタンスと共起する．その際，シード「Sugano」と共起している項が「Giants」ならば，項「Dragons」の方が項「Inter」より「Giants」との類似度が高いため，インスタンス「Tanishige」のスコアが「Nagatomo」よりも高くなりやすい．これにより，大量のインスタンス候補を，精度を保ちつつ獲得することができる．語彙獲得の評価実験により,「国名」クラスについては既存の手法よりも再現率が高い傾向が見られた．

2　関連研究：ブートストラップ法による語彙獲得

Pantel ら [4] は関係抽出タスクにおいて自己相互情報量を用いてインスタンスとパターンの信頼度を定義し，信頼度を伝播させていく Espresso アルゴリズムを提案した．本研究では Espresso アルゴリズムをベースラインの 1 つとし，Espresso アルゴリズムとそれを改良した手法を組み合わせることを提案する．

McIntosh ら [2] は複数の意味クラスの語彙を同時に獲得するために，２ステップからなる Relation Guided Bootstrapping (RGB) を提案した．RGB はまず，Mutual Exclusion Bootstrapping (MEB) アルゴリズム [1] を適用する．MEB はブートストラップ法の一種で，クラスごとにシードと共起するパターンを相互排他的に抽出し，パターンから新たなインスタンスを相互排他的に抽出するアルゴリズムである．次にすべてのクラスペアについて，クラスにそれぞれ含まれるインスタンスを用いて組を構成する．最後にその組の文脈をパターンとし，インスタンス獲得と同様に MEB を適用する．RGB はインスタンスの組の文脈もパターンとして用いる点で本研究と類似しているが，本研究では獲得対象のクラス数は 1 つであり，関係のあるクラスペアを明示的に与える必要がないことが相違点となっている．

3 提案手法

3.1 データベースの準備

本研究ではまず，コーパスに対して OpenNLP Chunker[1] を用い，フレーズチャンキングを行う．次に，その結果に表 1 のテンプレートを適用し，3 つ組（インスタンス，パターン，項）または組（インスタンス，パターン）を抽出し，データベースに格納する．

テンプレート	テンプレートにマッチする例文
X VP NP	[the U.S.] [is] [global leader]
NP VP X	[Vietnam] [defeated] [the U.S.]
X VP PRT/PP NP	[the U.S.] [carried] [out] [missile test]
NP VP PRT/PP X	[Christianity] [arrived] [in] [the U.S.]
X PP NP	[the U.S.] [with] [income inequality]
NP PP X	[politics] [of] [the U.S.]
X VP ADJP	[the U.S.] [was] [independent]

表 1: テンプレート一覧．X はインスタンス，NP は名詞句，VP は動詞句，PRT は分詞，PP は前置詞句，ADJP は形容詞句を表し，記号 [] は句のまとまりを表す．

提案手法では，同じコーパスから作成された以下のデータベース D，D'，V を用いる．

- D は表 1 の X をインスタンス，NP または ADJP を項，その間の列をパターンとした 3 つ組（インスタンス，項，パターン）の集合であり，項の類似度も用いてインスタンスの信頼度計算を行うために用いる．
- D' は表 1 の X をインスタンス，それ以外の系列をパターンとしたときのそれらの組の集合であり，Espresso による信頼度計算に用いる．
- V は表 1 の X を固有名詞に限らない NP，残りの系列をパターンとたときのそれらの組の集合である．V は項とその文脈の組を表し，項間の類似度計算に用いる．

3.2 語彙獲得のアルゴリズム

各反復において，Algorithm 1 はまず，新たに提案したアルゴリズム（Algorithm 2）と Espresso アルゴリズムのそれぞれでインスタンスの信頼度を計算する．次に，2 つの信頼度の積が高いインスタンスの上位のみをプールに蓄積する．Algorithm 2 はパターン獲得（行番号 3-6），組の獲得（行番号 8-10），インスタンスの信頼度計算（行番号 11-13）のステップに分けられる．ただし，Algorithm 2 では，記号「_」は任意の要素を表す．例えば，3 行目の $(\_,\_,p)$ は 3 つ組（任意のインスタンス，任意の項，p）を表す．以下では，各ステップについて説明する．

パターン獲得（行番号 3-6）：
シードを含む 3 つ組（インスタンス，項，パターン）の集合 D_S から，そこに含まれるパターンの集合を列挙し，その信頼度を以下の式 (1) で計算する．

$$conf_{D_S}(p) = \frac{1}{|D_S(p)|} \sum_{t \in D_S(p)} \frac{pmi(t,p)}{maxpmi} \cdot conf(t) \quad (1)$$

ここで，

$$D_S(p) = \{(s,a) | (s,a,p) \in D_S\} \quad (2)$$

であり，pmi は自己相互情報量，$maxpmi$ はデータベース D 中での最大の pmi を表す．ただし，pmi には Pantel ら [5] の提案する式 (3)

$$pmi(x,y) = \frac{C_{xy}}{C_{xy}+1} \cdot \frac{min(C_x,C_y)}{min(C_x,C_y)+1} \cdot \log_2 \left(\frac{\frac{C_{xy}}{N}}{\frac{C_x}{N} \cdot \frac{C_y}{N}} \right) \quad (3)$$

を用いた．式 (3) では，C_x はデータベース D 中での x の頻度，C_{xy} は x と y の共起頻度，N は $|D|$ を表す．式 (1) において，Espresso アルゴリズムに準じるならば，信頼度は $|D_S|$ で正規化するが，D_S 上には組（インスタンス，項）とパターンの意味的関係は複数存在し，また $|D_S|$ の値は大きいため，$|D_S(p)|$ で正規化した．最後に，信頼度が上位 a 個のパターンの集合を残し，P とする．

タプル獲得（行番号 8-10）：
パターンを含む 3 つ組（インスタンス，項，パターン）の集合から，そこに含まれる組（インスタンス，項）の集合を列挙し，その信頼度を以下の式 (4) で計算する．

$$conf_{D_p,D_S}(t) = score_{D_S,D_p}(t_a) \times \frac{1}{|D_p(t)|} \sum_{p \in D_p(t)} \frac{pmi(t,p)}{maxpmi} \cdot conf_{D_S}(p) \quad (4)$$

ここで，

$$score_{D_S,D_p}(t_a) = \frac{1}{|D_p(t)|} \sum_{p \in D_p(t)} \frac{1}{|D_S(p)|} \sum_{t' \in D_S(p)} \cos(\vec{t_a}, \vec{t'_a}) \quad (5)$$

[1] http://opennlp.apache.org/

また，式 (4) 及び式 (5) では

$$D_p(t) = \{p|(s,a,p) \in D_p, t = (s,a)\} \quad (6)$$

$$D_S(p) = \{(s,a)|(s,a,p) \in D_S\} \quad (7)$$

である．式 (5) において，t_a はタプル t の，t'_a はタプル t' の項を表し，t_a，t'_a のベクトル $\vec{t_a}$，$\vec{t'_a}$ はデータベース V においてパターンを要素とする文脈ベクトルである．また，cos はコサイン類似度を表す．

インスタンスの信頼度計算（行番号 11-13）：
最後に，インスタンス s の信頼度を以下の式 (8) で計算する．

$$conf_{D_p,D_S}(s) = \frac{1}{|D_p(s)|} \sum_{(a,p)\in D_p(s)} conf_{D_p,D_S}(s,a) \quad (8)$$

ただし，式 (8) はインスタンス s を含むタプルの信頼度を平均した値である．また，

$$D_p(s) = \{(a,p)|(a,p,s) \in D_p\} \quad (9)$$

とする．$D_p(s)$ は s と共起する組（項，パターン）の集合を表す．

4 語彙獲得の評価実験と考察

4.1 実験設定

本研究における提案手法の有効性を検証するための実験を行う．実験で用いるデータベースは，英語版 Wikipedia から本文を抽出してチャンキングし，表 1 のテンプレートを適用して作成した．ただし，テンプレートにより獲得した 3 つ組，2 つ組のうち 10%をランダムサンプリングし，頻度 5 以上のインスタンスを含むものをデータベースに登録した．

提案手法（proposed）と比較するベースラインは 2 つである．1 つ目は，データベース D' においてシードとのコサイン類似度の平均が高いインスタンスの順に列挙する単純な手法（cos）で，2 つ目は，データベース D' に対して Espresso アルゴリズムを適用する手法（Espresso）である．

評価尺度として Precision 式と Recall 式を用いたが，Recall は token ではなく type 単位で計算した．理由として，単語によっては極端に使用頻度の少ないと考えられる表記が存在し，それらが recall を大幅に下げるためである．また，実験では Pantel ら [3] が公開しているデータを正解インスタンスとし，頻度の高いインスタンスを初期シードとした．

Algorithm 1 提案手法

INPUT: 初期シード集合 *Seeds*，データベース D, D', V，獲得するインスタンスの個数 *num*，各反復でパターン獲得に用いるシードの個数 n，各反復でインスタンス獲得に用いるパターンの個数 a, b

OUTPUT: 蓄積したインスタンスの集合 *Pool*

1: **for** $s \in Seeds$ **do**
2: $conf(s) = 1$
3: **end for**
4: $Pool \leftarrow Seeds$
5: **while** $|Pool| < num$ **do**
6: $S \leftarrow$ (*Pool* に含まれるインスタンスの中で，信頼度 *conf* がトップ n 個の集合)
7: $A \leftarrow$ 提案手法によるインスタンスと信頼度の獲得 $(Algorithm2)(S, D, V, a)$
8: $B \leftarrow$ *Espresso* によるインスタンスと信頼度の獲得 (S, D', b)
9: $I = \emptyset$
10: **for** $(s, conf_A) \in A$ **do**
11: $conf = conf_A \times (B$ における s の信頼度)
12: $I \leftarrow I \cup \{(s, conf)\}$
13: **end for**
14: $Pool \leftarrow Pool \cup (I$ で信頼度がトップ n 個のインスタンス s の集合)
15: **end while**

Algorithm 2 提案手法によるインスタンスの信頼度計算

INPUT: シード集合 S，データベース D, V，インスタンス獲得に用いるパターンの個数 a

OUTPUT: インスタンスとその信頼度の組の集合 I'

1: $I' = \emptyset$
2: $D_S = \{(s,a,p)|s \in S, (s,a,p) \in D\}$
3: **for** パターン候補 $p \in \{p|(\_,\_,p) \in D_S\}$ **do**
4: $conf_{D_S}(p) =$ 式 (1) により p の信頼度を計算
5: **end for**
6: パターン集合 $P \leftarrow conf_{D_S}(p)$ がトップ a の p の集合
7: $D_p = \{(s,a,p)|p \in P, (s,a,p) \notin D \setminus D_S\}$
8: **for** インスタンスと項の組 $t \in \{(t_s,t_a)|(t_s,t_a,\_) \in D_p\}$ **do**
9: $conf_{D_p,D_S}(t) =$ 式 (4) により t の信頼度を計算
10: **end for**
11: **for** $s \in \{s|(s,\_,\_) \in D_p\}$ **do**
12: $I' \leftarrow I' \cup \{(s,$式 (8) による s の信頼度 $conf_{D_p,D_S}(s))\}$
13: **end for**

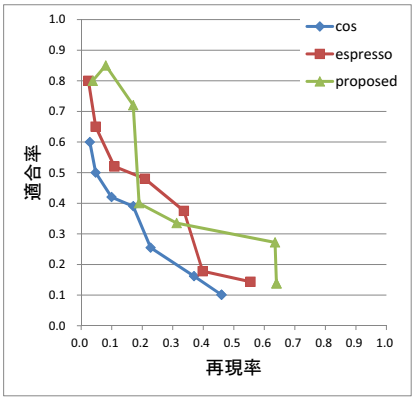

図 1: クラス「国名」の実験結果

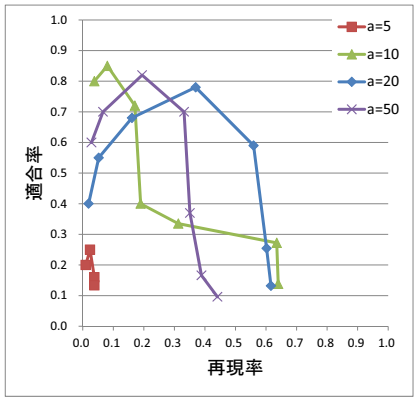

図 2: パラメータ依存性の調査

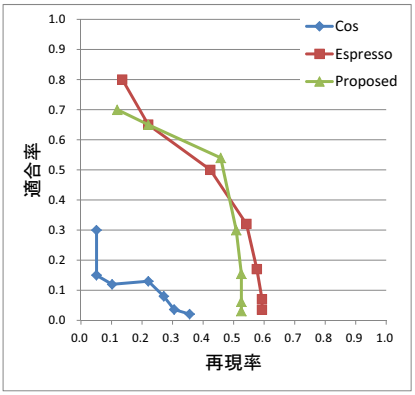

図 3: クラス「元素」の実験結果

4.2 実験結果

実験結果を図 1～図 3 に示す．各図において，同色の各点は 1 手法の同じ設定で行われた実験結果であり，再現率の低い方から順に，獲得順位が 10，20，50，100，200，500，1000 位以内のインスタンスに関する再現率と適合率を表す．図 1 は「国名」クラスの実験結果で，各手法は最も良い性能を出すパラメータに設定した場合の結果である．パラメータはシードの個数を 5，10，20，パターンの個数を 5，10，20 で実験したところ，手法 cos はシードの個数が 5 個の場合，Espresso はシードが 5 個，パターンは 20 個の場合が最も良かった．提案手法では，Espresso のパラメータは同一で，Algorithm 2 で用いる，組（インスタンス，項）を獲得するためのパターン数（a）は 10 個の場合が最も良かった．図 2 は「国名」クラスを提案手法により獲得した場合のパラメータ依存性をみるための実験結果で，シードを 5 個，Espresso のパターンを 20 個に固定し，パラメータ a を 5，10，20，50 個に変化させた場合の実験結果を表す．図 3 ではシードは 10 個，Espresso のパターンは 10 個，a は 20 に設定した．

4.3 実験結果に対する考察

「国名」クラスについては，図 1 より，特に上位 500 個のインスタンスを出力した場合の再現率が 0.64 となっており，２つのベースライン（再現率 0.40）を大きく上回っている．これは，提案手法のパターンがベースラインと比較してより多くのインスタンスと共起することで，相対的に多くのインスタンスを候補とできるためと考えられる．実際，提案手法では例えば，データベース D 上では「＜インスタンス＞ withdraw from ＜項＞」など，国名のインスタンスと良く共起するようなパターンの信頼度が高かった．一方，Espresso ではデータベース D' を用いるため，項がない．よって，D' 上での似たようなパターンとしては「＜インスタンス＞ withdraw from Iraq」があるが，このようなパターンは限られたインスタンスとのみ共起するため，信頼度が低く，インスタンス抽出には使われない[2]．また，国名に関しては図 2 より，パラメータ a と性能の間には相関が見られなかったが，これは，パターンが少ないと十分なインスタンスが得られず，多いと信頼度の低いパターンが獲得されてしまうためと考えられる．「元素」クラスについては，図 3 より，提案手法と Espresso の間に大きな性能差は見られなかった．これは，データベース D' に Espresso を適用した場合，「＜インスタンス＞ is element」や「＜インスタンス＞ with acid 」など，「元素」を意味するインスタンスと高頻度で共起するパターンが多く獲得され，提案手法による改善の余地が少ないためと考えられる．

参考文献

[1] T. McIntosh. and J. R. Curran. Weighted mutual exclusion bootstrapping for domain independent lexicon and template acquisition. In *ALTA*, 2008.

[2] T. McIntosh, L. Yencken, J. R. Curran, and T. Baldwin. Relation guided bootstrapping of semantic lexicons. In *ACL*, 2011.

[3] P. Pantel, E. Crestan, A. Borkovsky, AM. Popescu, and V. Vyas. Web-scale distributional similarity and entity set expansion. In *EMNLP*, 2009.

[4] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *ACL*, 2006.

[5] P. Pantel and D. Ravichandran. Automatically labeling semantic classes. In *HLT-NAACL*, 2004.

[2] ただし，D' 上では他に「countries such as ＜インスタンス＞」のように，国名に特徴的なパターンは存在する．