分布類似度のための文脈素性選択

萩原 正人　小川 泰弘　外山 勝彦
名古屋大学大学院情報科学研究科
{hagiwara,yasuhiro,toyama}@kl.i.is.nagoya-u.ac.jp

1　はじめに

自然言語処理タスクにおいて，語の意味的類似性は重要な語彙知識であり，語義曖昧性解消や類義語自動獲得など，幅広い応用がある．その際，語の類似性を測る概念として「分布類似度」が提案され，広く利用されている．これは，語の持つ文脈の共通度を利用した類似度であり，類似した文脈を持つ語は意味も類似している可能性が高いという「分布仮定」[8] に基づいている．

これまでに，近接語 [3, 10] や依存関係 [9, 11]，依存パス [12] など，様々な文脈情報が提案され，分布類似度の計算に利用されてきた．しかし，分布類似度を正確に求めるためには大量の共起データが必要であり，結果として膨大な次元数を持つ意味空間を構成することになり，計算量が非実用的なものになるという問題がある．一方で，すべての文脈情報が有用というわけではないため，有効な文脈素性のみを選択し，効率的に分布類似度を計算する必要性が生ずると考えられる．

Curran and Moens [4] は，この問題に対して，元のベクトルから抽出した少数の代表的な属性である基準属性 (cannonical attributes) を各語に付与し，語の類似可能性判定に利用することを提案している．しかし，この基準属性の抽出の際に，主語，直接・間接目的語であり，重みの最も大きいもの，という条件を採用しているが，これが適切かどうかについては検討の余地がある．また，有効な文脈の比較や取捨選択を試みた研究 [3, 7] も存在するが，性能の事後的評価に留まっており，一般的な知見が得られているとは言い難い．したがって，文脈および文脈カテゴリの取捨選択に適用できる汎用的な定量的手法が必要である．

ここで，素性選択の分野に目を向けてみると，特に文書分類の分野において多くの研究が行われている [13]．これらの手法は，分類精度を保ちながら素性数を削減することに成功しているが，分布類似度の計算は文書分類問題とは本質的に異なり，同様の手法が文脈の自動選択に適用できるかどうかは改めて検討する必要がある．

そこで本研究では，既存の素性選択手法を分布類似度問題における文脈選択に適用し，その有効性を検証する．具体的には，まず 2 節にて既存の指標である DF, TS, MI, IG, CHI2 について述べ，分布類似度問題にどのように適用するかについて説明する．続いて，コーパスから依存関係を文脈として抽出し，これらの指標を用いて不要な文脈を取り除いた後，類義語の自動獲得タスクにおける性能の変化を評価する．類似度獲得手法および性能評価については 3 節と 4 節で，実験については 5 節で詳しく述べる．また，5 節では文脈カテゴリの重要度を評価できるようこれらの指標を拡張し，文脈カテゴリの評価に対しても有効であることも示す．

2　文脈選択手法

本節では，主に文書分類において提案された既存の素性選択指標について紹介し，分布類似度問題への適用方法について説明する．以下では，n と m をそれぞれ，コーパス中の異なり語数および異なり文脈数とする．また，語 w と文脈 c の共起回数を $N(w,c)$ によって表す．

2.1　文書頻度 (DF)

文書頻度 (DF) は，索引語が共起する文書数であり，情報検索における重み付けに利用される．分布類似度の場合，DF は文脈が共起する異なり語の数，すなわち，

$$df(c) = |\{w|N(w,c) > 0\}|$$

となる．DF を文脈素性選択に用いるのは，多くの語と共起する文脈は有用である可能性が高いという仮定に基づいている．

2.2 索引語強度 (TS)

索引語強度 (Term Strength; TS) は，Wilbur and Sirotkin [14] によって提案された概念であり，語が「類似した文書」に共通してどのぐらい現れやすいかを示すものである．分布類似度に対しては，TS は

$$s(c) = P(c \in C(w_2) | c \in C(w_1))$$

と定義される．ここで，(w_1, w_2) は類義語ペアであり，$C(w)$ は語 w の持つ文脈の集合，すなわち $C(w) = \{c | N(w, c) > 0\}$ とする．実際には，P_H を類義語ペアの集合として，以下のように計算する．

$$s(c) = \frac{|\{(w_1, w_2) \in P_H | c \in C(w_1) \cap C(w_2)\}|}{|\{(w_1, w_2) \in P_H | c \in C(w_1)\}|}.$$

TS は，類義語ペアからなる評価セット P_H を必要とする点で DF とは異なる．本稿では P_H として，次節において述べるクラス $s = 1$ のペアの集合を用いた．

2.3 分布類似度の定式化

以下で紹介する選択手法の MI, IG, CHI2 は，クラス分類問題を対象としているという点で上記２つの手法とは本質的に異なるものである．そのためにまず，以下のようにして，分布類似度問題を分類問題として定式化する必要がある．

まず，以下では語の代わりに語のペアを分類対象として扱う．ここで，各ペアに対して，文脈 $c_1, ..., c_m$ に対応する素性 $f_1, ..., f_m$ を付与する．これらの素性は，語のペア $p = (w_1, w_2)$ が文脈 c_j を共通に持つ，すなわち $N(w_1, c_j) > 0$ かつ $N(w_2, c_j) > 0$ のとき $f_j = 1$, それ以外のとき $f_j = 0$ と定義する．さらに，語のペアが類似しているとき $s = 1$, それ以外のとき $s = 0$ であるような対象クラス s を定義する．このとき，分布類似度問題は，素性 $f_1, ..., f_m$ に基づき，語のペアに対してクラス $s \in \{0, 1\}$ を割り当てる問題として定式化される．

以下の素性選択指標を計算するために，クラス $s = 1$ に対応する類義語のペアと，クラス $s = 0$ に対応する非類義語ペアを評価セットとして準備した．これらの評価セットは，4.1 節において述べる参照セットを用いて作成した．具体的には，参照セット中から無作為に抽出した類義語ペア 5,000 個をクラス $s = 1$ の評価セットとして，また，LDV(詳細は 4.1 節で述べる) から無作為にペアを作成し，手作業で類義語ペアを取り除いた 5,000 ペアをクラス $s = 0$ の評価セットとして用いた．

2.4 相互情報量 (MI)

統計言語処理において，語の関連や重み付けによく用いられる指標に相互情報量 (MI) がある．素性 f とクラス s との相互情報量は，下式により計算される．

$$I(f, s) = \log \frac{P(f, s)}{P(f)P(s)}$$

文脈に対する重要度は，$I_{\max}(c_j) = \max_{s \in \{0,1\}} I(f_j, s)$ として両クラスの MI の値を統合したものを用いる [13].

2.5 情報利得 (IG)

情報利得 (IG) は，主に機械学習の分野において素性の重要度に対する基準として用いられる．IG は，ある事象の組に対し，一方の事象の結果を知ることによって他方の事象に関して得る情報の量として，以下のように定義される．

$$G(c_j) = \sum_{f_j \in \{0,1\}} \sum_{s \in \{0,1\}} P(f_j, s) \log \frac{P(f_j, s)}{P(f_j)P(s)}$$

2.6 χ^2 統計量 (CHI2)

χ^2 統計量は，クラスと素性の従属性を測る指標であり，一般にはコーパスからのコロケーションの発見などによく用いられる．素性 f と語 w の χ^2 統計量は，$f_j = n$ かつ $s = m$ であるようなペアの数を $F^j_{nm}(n, m \in \{0, 1\})$ とし，すべてのペアの数を N とすると，以下のように計算できる．

$$\chi^2(c_j) = \frac{N(F_{11}F_{00} - F_{01}F_{10})}{(F_{11} + F_{01})(F_{10} + F_{00})(F_{11} + F_{10})(F_{01} + F_{00})}$$

3 類義語獲得

本節では，分布類似度の重要な応用である類義語獲得について述べる．類義語獲得は，本稿で文脈選択の性能評価に用いたタスクである．まず，3.1 節でコーパスからの文脈抽出について詳しく述べ，続いて 3.2 節で，抽出された文脈からの類似度計算について述べる．

3.1 文脈の抽出

本稿では，文脈情報として依存関係 [11] を用い，依存関係の抽出には RASP Toolkit 2[1] を用いた．RASP2 は，文章を解析し，GR(grammatical relation) と呼ばれる依存関係を出力する．例えば，文:

The investigators were still looking for witnesses and the motive of the attack.

に対しては，以下のような GR の組が出力される．

```
(ncsubj look investigator _)
(ncmod _ look still)
(aux look be)
(iobj look for)
(dobj for and)
(conj and witness)
(conj and motive)
(det motive the)
(iobj motive of)
(dobj of attack)
(det attack the)
(det investigator the)
```

RASP2 の出力する GR は一般に n 項組であるが，ここでは，以下のようにして名詞を対象語として取り出し，語とその文脈との共起に変換する．

```
(語)          -  (文脈)
investigator  -  (ncsubj look * _)
investigator  -  (det * the)
witness       -  (conj and *)
motive        -  (conj and *)
motive        -  (det * the)
motive        -  (iobj * of)
attack        -  (dobj of *)
attack        -  (det * the)
```

3.2 類似度の計算

続いて，求められた語と文脈の共起頻度を基に，語の類似度を計算する．共起頻度をそのまま用いることも可能であるが，予備実験の結果を受け，本稿では以下のように重みとして相互情報量を用いる．

$$\mathrm{wgt}(w,c) = \log \frac{P(w,c)}{P(w)P(c)}$$

ここで，負の重みは性能を悪化させる原因となる可能性があるため [4]，$\mathrm{wgt}(w,c) \geq 0$ となるように修正した．最後に，語 w_1 と語 w_2 との類似度は，以下のように Jaccard 係数を用いて求めた．

$$\frac{\sum_{c \in C(w_1) \cap C(w_2)} \min(\mathrm{wgt}(w_1,c), \mathrm{wgt}(w_2,c))}{\sum_{c \in C(w_1) \cup C(w_2)} \max(\mathrm{wgt}(w_1,c), \mathrm{wgt}(w_2,c))}$$

4 性能評価

本節では，前節で述べた手法を用いて自動獲得された類義語に対する性能評価指標として，平均精度 (AP) と相関係数 (CC) について述べる．

4.1 平均精度 (AP)

平均精度 (AP) は，情報検索において広く用いられている性能評価指標であり，検索された文書 (本稿では，獲得された類義語) に含まれる正解の割合である．この平均精度を計算するために，まず，精度を求める際にクエリとして用いるクエリ語集合を準備する．ここでは，Longman Defining Vocabulary [1] をクエリ語の候補として用いた．LDV 中の各語を，Roget's Thesaurus [5], Collins COBUILD Thesaurus [2] および WordNet Release 3.0 [6] の 3 つの既存のシソーラスを用いて調べ，記載されている類義語の和集合を参照セットすなわち正解とした．ただし，名詞としての類義語のみを用い，“idiom”, “informal”, “slang” のラベルのついているもの，および 2 語以上からなるものは除外した．この結果残った LDV 中の 771 語に対して類義語を自動獲得し，再現率が 0%, 10%, ..., 100%の各点における精度を平均し，AP の値とした．

4.2 相関係数 (CC)

相関係数 (CC) は，求められた類似度と，基準類似度，すなわち正解との相関係数である．基準類似度は，WordNet の木構造中の節点の「近さ」に基づき計算する．この CC の値が大きいほど，得られた結果は WordNet に類似しており，性能が高いと判断できる．具体的には，語義 $w_1, ..., w_{m_1}$ を持つ語 w と，語義 $v_1, ..., v_{m_2}$ を持つ語 v との基準類似度は，w_i と v_j に対応する節点の深さをそれぞれ d_i, d_j とし，両者の共通祖先の深さの最大値 d_{dca} とすると，

$$sim(w,v) = \max_{i,j} sim(w_i, v_j) = \max_{i,j} \frac{2 \cdot d_{\mathrm{dca}}}{d_i + d_j},$$

として計算する．この基準類似度を用いて，CC の値は，基準類似度の列 $\boldsymbol{r} = (r_1, r_2, ..., r_n)$ と，求められた類似度の列 $\boldsymbol{s} = (s_1, s_2, ..., s_n)$ との相関係数として計算する．なお，これらの類似度は，LDV から無作為に抽出した語のペア 4,000 個から，基準類似度の低いもの 2,000 個を除外した残りの 2,000 個に対して計算する．

5 実験

本節では，2 節で述べた文脈選択手法の評価実験の手順および結果について述べる．

[1] http://www.cs.utexas.edu/users/kbarker/working_notes/ldoce-vocab.html

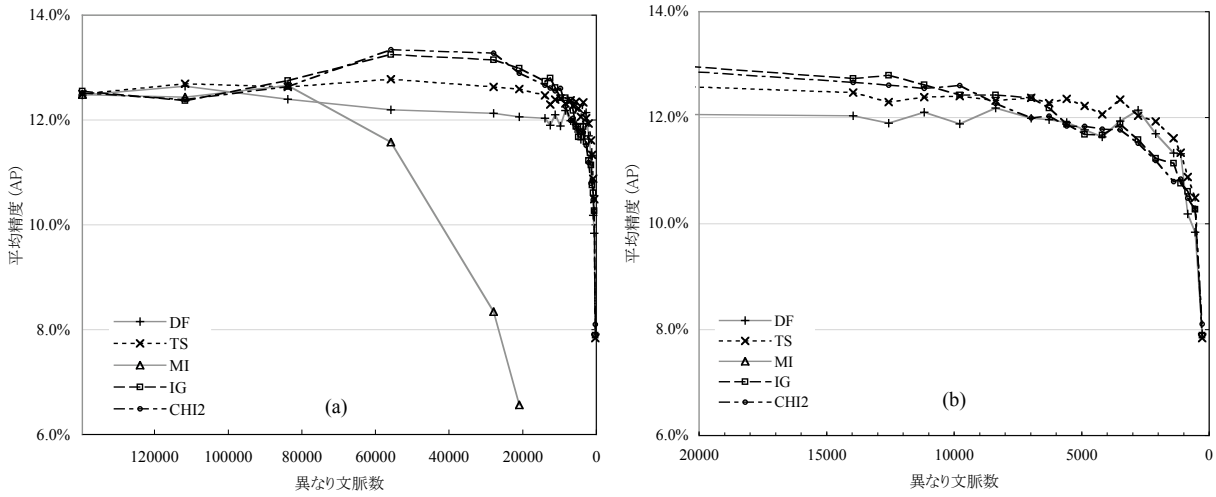

図 1: 文脈選択後の類義語自動獲得の性能
(a) 全体図 (b) 異なり文脈数 0 から 20,000 までの拡大図

5.1 条件

コーパスには，English Gigaword[2]の New York Times の記事 (130 万文書，9.14 億語) を用いた．また，低頻度語を取り除き，計算量を軽減するために，出現頻度に関する閾値 θ_f を設け，$\sum_c N(w,c) < \theta_f$ となる語 w と，$\sum_w N(w,c) < \theta_f$ となる文脈 c を共起から除去した．閾値 θ_f は 40 に設定した．

なお，本実験では，類義語獲得の対象を名詞に限定し，対応する文脈のみを抽出した．具体的には，RASP によって付与された POS タグが，`APP`, `ND`, `NN`, `NP`, `PN`, `PP` のいずれかであるものを名詞と判定した．この結果，異なり語数 40,461 および異なり文脈数 139,618 からなる共起行列が作成された．

5.2 文脈の削減

最初の実験では，2 節で述べた文脈選択手法の有効性を検証する．各文脈に対して DF, TS, IM, IG, CHI2 の 5 個の指標を計算し，値の小さい文脈を取り除いた後の性能 (AP と CC) の変化を求めた．この操作は，文脈を全て用いる場合 (139,618 個) から始め，文脈数が 0.2%(279 個) になるまで続けた．その結果が図 1 である．

この結果から，性能を元の水準もしくはそれ以上に保ちながら，文脈選択によって 80%(約 110,000 個) 以上の文脈を削減することができることが分かる．全体としては，TS と IG が良い性能を示す一方で，MI の性能は文脈の削減に伴って急速に低下した．紙面の都合で結果は省略するが，CC も同様の傾向を示した．なお，タスクは異なるが，文書分類タスク [13] と一貫性のある結果を示したことから，2.3 節の定式化は，今回の目的に照らして適切であったことが分かる．

5.3 文脈カテゴリの選択

次の実験では，分布類似度の計算に対して有効な文脈の構成に注目し，文脈カテゴリに対する素性選択手法を検証する．まず，性能の観点から見て有効な文脈に対象を絞るために，DF, TS, IG, CHI2 によって選ばれた上位 10%(13,961 個) の文脈の共通集合を求め，これを「エリート文脈」の集合とした．その結果，6,440 個がエリート文脈として獲得された．

次に，文脈が対象語に対して持つ文法的関係に基づき，文脈をカテゴリへと分割した．さらに，このカテゴリが性能に対して与える影響を測るために,「カテゴリ重要度」をカテゴリに含まれる全ての文脈の IG の和として定義した．ここで IG の和を採用したのは，(a) 前述の実験において比較的良い性能を上げていたこと，および，(b) 文脈の独立性を仮定した場合，カテゴリの IG の値は含まれる文脈の IG の総和になることが理由である．

RASP の GR に基づいて定義した頻出カテゴリ:`ncsubj`, `dobj`, `obj`, `obj2`, `ncmod`, `xmod`, `cmod`, `ccomp`,

[2]http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05

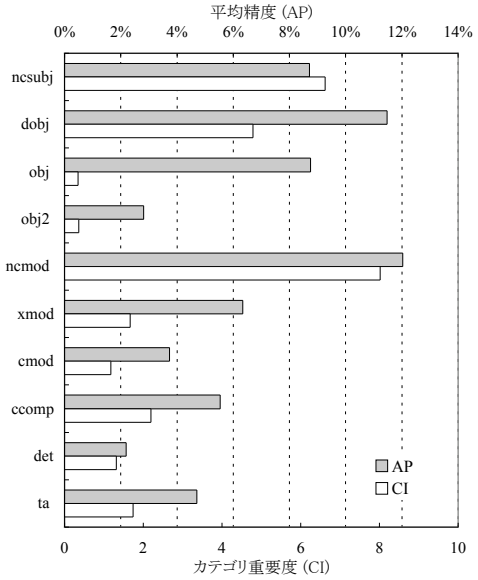

図 2: 類義語獲得の性能とカテゴリ重要度との比較

det, ta に対して，エリート文脈集合におけるカテゴリ重要度を計算した結果を示したのが図 2 である．図には，各カテゴリを文脈として単体で用いて類義語を獲得した際の性能も併せて示してある．結果から，カテゴリ重要度と性能との間に強い相関 ($r = 0.766$) が存在することが分かり，エリート文脈という限定された文脈集合におけるカテゴリ重要度によって，最終的な類義語獲得の性能を予測できる可能性が示唆される．

なお，カテゴリの持つ文法的関係に注目すると，修飾関係 (ncmod) が比較的有効であることが分かる．この結果は，これまでに広く用いられてきた subj や obj と比較して mod がより有効であるという文献 [7] の結果ともよく一致している．ただ，mod が有効であるのは，単にそれが最大のカテゴリ (エリート文脈中 2,515 個) であるからという可能性もあり，文脈カテゴリのサイズと性能との関係について，引き続き検討する必要がある．

6 おわりに

本稿では，分布類似度問題を分類問題として定式化することにより，これまで主に文書分類の分野において提案されてきた素性選択手法を適用し，その有効性を検証した．類義語獲得タスクにおける性能評価の結果，性能を元の水準もしくはそれ以上に保ちながら，文脈選択によって約 90%もの文脈を削減することができることを示した．また，文脈選択手法を拡張することにより，文脈カテゴリに対する重要度を定義し，有効な文脈カテゴリを事前に予測できる可能性を示した．近接語や依存パスなど，他の種類の文脈における検証は今後の課題である．

参考文献

[1] Ted Briscoe, John Carroll and Rebecca Watson: The Second Release of the RASP System. *Proc. of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 77 - 80, 2006.

[2] Collins Cobuild Major New Edition CD-ROM, HarperCollins Publishers, 2002.

[3] James R. Curran and Marc Moens: Scaling Context Space, *Proc. of the 40th Anual meeting of the ACL*, pp. 231 - 238, 2002.

[4] James R. Curran and Marc Moens: Improvements in automatic thesaurus extraction. In Workshop on Unsupervised Lexical Acquisition. *Proc. of ACL SIGLEX*, pp. 231–238, 2002.

[5] Editors of the American Heritage Dictionary. Roget's II: The New Thesaurus, 3rd ed. *Houghton Mifflin*, 1995.

[6] Christiane Fellbaum: WordNet: An Electronic Lexical Database. *MIT Press*, 1998.

[7] Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama: Selection of Effective Contextual Information for Automatic Synonym Acquisition. *Proc. of COLING/ACL 2006*, pp. 353–360, 2006.

[8] Zellig Harris: Distributional Structure. Katz, J. (ed.) *The Philosophy of Linguistics*. Oxford University Press. pp. 26 - 47, 1985.

[9] Donald Hindle. Noun classification from predicate-argument structures. *Proc. of the 28th Annual Meeting of the ACL*, pp. 268–275, 1990.

[10] Will Lowe and Scott McDonald: The direct route: Mediated priming in semantic space. *Proc. of the 22nd Annual Conference of the Cognitive Science Society*, pp. 675–680, 2000.

[11] Dekang Lin: Automatic retrieval and clustering of similar words, *Proc. of the COLING/ACL 1998*, pp. 786 - 774, 1998.

[12] Seastian Pado and Mirella Lapata. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, Volume 33, Issue 2, pp. 161–199, 2007.

[13] Yiming Yang and Jan O. Pedersen: A Comparative Study on Feature Selection in Text Categorization. *Proc. of ICML 97*, pp. 412–420, 1997.

[14] John Wilbur and Karl Sirotkin: The automatic identification of stop words. *Journal of Information Science*, pp. 45–55, 1992.