perplexityを用いた類義語獲得の自動評価

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@kl.i.is.nagoya-u.ac.jp

1　はじめに

語彙に関する知識は，自然言語処理において基本的かつ重要な知識である．特に類義語に関する知識関係は，シソーラスの構築や情報検索など，応用も幅広く，その自動獲得に関する研究が盛んに行われてきた．

類義語の自動獲得のためには，コーパス中の語に関する文脈情報から，ベクトル空間モデルや確率モデルなどの言語モデルを生成し，そこから単語間の類似度を求める手法が一般的である．しかし，その評価に関しては一般化された手法が確立されておらず，人手もしくは既存の正解セットに頼っているのが現状である．人手により評価する場合，評価の信頼性を確保するためには，大量の類義語に対して関連性を判断しなくてはならず，コストや正当性・一貫性が問題となる．また，既存の正解セットについても，特定分野の類義語を自動獲得する場合などには，正解セットがそのまま利用できない場合がある．

これまで，言語モデルの性能評価には，perplexityと呼ばれる指標が一般的に用いられてきた．perplexityは，言語モデルの複雑さを表す指標であり，正解セットを用いることなく計算が可能であるという利点を持つ．情報検索の分野においては既に，perplexityと情報検索システムの性能との相関が示されている [1]．しかし，類義語獲得と情報検索では，性能評価の観点が本質的に異なるため，類義語の自動獲得に対するperplexityの有効性をあらためて検証する必要がある．

そこで本稿では，類義語の自動獲得に対して，perplexityによる評価が有効であるかどうかを明らかにする．具体的には，言語モデルの学習に用いたセットとは別のテストセットを用い，テストセットperplexityを計算することにより，その言語モデルの汎化能力を測定する．同時に，既存のシソーラスWordNet[5]を用いることにより，類義語獲得の性能評価を自動で行う．本稿では，このテストセットperplexityと類義語自動獲得の性能との相関関係を，3種類の言語モデル(VSM，LSI，PLSI)について実験に基づき考察した．

2　言語モデルと類義語獲得

本節では，言語モデルとしてVSM，LSI，PLSIを取り上げ，それぞれの言語モデルを用いた類義語の自動獲得手法について説明する．

2.1　共起関係の獲得

本稿では，分布仮定 (distributional hypothesis)[7]，すなわち，文脈の類似する語は類似した意味を持つという仮定に基づき，言語モデルの学習用セットとして，コーパス中の動詞と名詞の共起を用いる．ただし，共起には，主格・目的格としての共起および前置詞を介した共起を含む．なお，以下に示す言語モデルにおいては，文書と索引語など2項間の共起を対象としているため，ここでは動詞と格・前置詞を組として見なし，単一の動詞として扱う．例として，

John gave presents to his colleagues.

という文からは，(“John”, “give:SUBJ”), (“present”,“give:OBJ”), (“colleague”, “give:to”) という3つの共起が抽出される．

以下では，N を類義語獲得の対象となる名詞の異なり語数，M を名詞と共起する動詞の異なり語数，$\mathrm{tf}(n, v)$ を名詞 n と動詞 v が共起する回数とする．

2.2　VSM

本稿で扱う最も単純な言語モデルは，相対頻度をそのまま用いてベクトルを構成するベクトル空間モデル(VSM) である．ここでは，名詞 n_i に対する名詞ベクトル $\boldsymbol{n}_i$ を，共起する動詞とその共起頻度を用いて以下のように作成する．

$$\boldsymbol{n}_i = {}^t[\mathrm{tf}(n_i, v_1) \;\; \mathrm{tf}(n_i, v_2) \;\; \ldots \;\; \mathrm{tf}(n_i, v_M)] \qquad (1)$$

名詞 n_i と n_j との類似度は，対応するベクトル間の余弦の値を用いる．

2.3 LSI

Latent Semantic Indexing(LSI)[4] は，文書と索引語との共起を扱った情報検索のためのモデルとして提案されたが，文書を名詞に，索引語をその名詞と共起する動詞に置き換えることにより，そのまま類義語の自動獲得へ適用することが可能である．まず，2.2 で述べたベクトル $\{\boldsymbol{n}_i\}$ を用いて，索引語文書行列に対応する行列を

$$X = [\boldsymbol{n}_1\ \boldsymbol{n}_2\ ...\ \boldsymbol{n}_N] \tag{2}$$

として作成し，この行列 X を以下のように特異値分解する．

$$X = T_0 S_0 D_0^T \tag{3}$$

ここで，行列 D_0^T 中の列ベクトルは，各名詞ベクトルを主成分を用いて再構成したものとなっている．ここから，重要な順に K 個 ($K < \mathrm{rank}(X)$) の要素を取り出すことによって，ベクトル空間の次元を圧縮できる．名詞間の類似度は，VSM と同様に，余弦を用いて計算できる．

2.4 PLSI

Probabilistic LSI(PLSI)[8] も，LSI と同様，対象の潜在意味を扱うモデルであるが，PLSI は情報理論・確率論に基づいているという点で異なっている．PLSI は，情報検索ならびに類義語の自動獲得において，LSI を超える性能を発揮することが示されている [6], [8].

PLSI モデルを用いた類義語獲得においては，名詞 n と動詞 v が，潜在意味クラス z を介して，以下のように共起すると仮定する．

$$P(n,v) = \sum_{z \in Z} P(z)P(n|z)P(v|z) \tag{4}$$

ここで，潜在意味クラス z は学習セットからは直接観察されないため，パラメータ $P(z), P(n|z), P(v|z)$ は EM アルゴリズムを用いて最尤推定する必要がある．EM アルゴリズムの適用後，ベイズの定理を用いて，各名詞に対する潜在意味分布を以下のように求める．

$$P(z|n) = \frac{P(n|z)P(z)}{\sum_{z'} P(n|z')P(z')} \tag{5}$$

名詞 n_i と n_j との間の類似度は，対応する潜在意味分布を比較することによって求められる．本稿では，Skew Divergence($\alpha = 0.99$)[10] を用いて，潜在意味分布 $p = P(z|n_i)$ と $q = P(z|n_j)$ との間の距離を計算し，それに基づき以下のように類似度を求めた．

$$sim(n_i, n_j) = \exp\{-\lambda s_\alpha(p \,||\, q)\} \tag{6}$$

$$s_\alpha(p \,||\, q) = KL(p \,||\, \alpha q + (1-\alpha)p) \tag{7}$$

ここで，$KL(p \,||\, q)$ は p と q との Kullback-Leibler 距離である．λ は適当なレンジ調節パラメータであり，各言語モデルの $sim(n_i, n_j)$ の平均が等しくなるように設定した．

3 perplexity

perplexity は，言語モデルの複雑さを表す指標であり，言語を情報源として見なしたときの 1 語あたりのエントロピーから計算される．言語モデルの評価の際には，学習に用いたものとは互いに素なテストセットに対して，テストセット perplexity PP を計算する．

$$PP = \exp\left\{-\frac{1}{L}\log\prod_{\boldsymbol{n}} P(\boldsymbol{n})\right\} \tag{8}$$

$$= \exp\left\{-\frac{\sum_{n,v}\mathrm{tf}_t(n,v)\log P(v|n)}{L}\right\} \tag{9}$$

ここで，$\mathrm{tf}_t(n,v)$ はテストセット中の名詞 n と動詞 v との共起回数であり，$L = \sum_{n,v}\mathrm{tf}_t(n,v)$ である．

式 (9) 中の $P(v|n)$ は，学習した言語モデルにより計算される．VSM の場合，相対出現頻度をそのまま用いて

$$P(v|n) = \frac{\mathrm{tf}(n,v)}{\sum_v \mathrm{tf}(n,v)} \tag{10}$$

と計算できる．

なお，LSI はベクトル空間モデルに基づいているため，確率を直接計算することができない．本稿では，[2] の手法に従い，LSI によって求められたベクトルを確率分布 $P(v|n)$ へ変換した．一方，PLSI の場合は，求められたパラメータから直接，

$$P(v|n) = \frac{P(v,n)}{P(n)} = \frac{\sum_{z\in Z} P(z)P(n|z)P(v|z)}{\sum_{z\in Z, v\in V} P(z)P(n|z)P(v|z)} \tag{11}$$

として計算した．

4　性能評価

本節では，2 節で述べた手法を用いて自動獲得された類義語に対して，性能評価を自動で行う手法について述べる．類義語獲得の自動性能評価のためには，まず，既存のシソーラス WordNet を用いて，正解となる類似度（以下，基準類似度と呼ぶ）を求める．続いて，この基準類似度を用い，識別率と相関係数の 2 つの性能評価指標を計算する．

4.1　WordNet を用いた基準類似度の計算

基準類似度は，シソーラスの木構造を基に 2 語の類似度を計算する一般的な手法 [11] を用いて計算した．具体的には，WordNet 中の語義 w_i に対応する節点の深さ d_i, 語義 v_j に対応する節点の深さ d_j，および，これら 2 節点に対する共通祖先の深さの最大値 d_{dca} から，

$$sim(w_i, v_j) = \frac{2 \cdot d_{\mathrm{dca}}}{d_i + d_j} \tag{12}$$

として語義間の類似度を求めた．さらに，語 w の語義 $w_1, ..., w_n$ と語 v の語義 $v_1, ..., v_m$ 間の類似度を用いて，語 w と語 v との基準類似度を以下により計算した．

$$sim(w, v) = \max_{i,j} sim(w_i, v_j) \tag{13}$$

4.2　識別率

識別率とは，語のペア (w_1, w_2) の関連の度合いを，言語モデルから計算された類似度 (以下，評価対象類似度と呼ぶ) によって識別できる割合である [9]．この識別率の計算のために，高関連の語のペアと無関連の語のペアから成る 2 種類のテストセットを用いる．これらのテストセットの作成には，まず，100,000 個のペアを無作為に生成し，そこから基準類似度の高いペア 2,000 個と低いペア 2,000 個を抽出し，それぞれ高関連テストセット，無関連テストセットとした．

計算手順は以下の通りである．まず，セット内の各ペアについて評価対象類似度を計算し，類似度がある閾値 t 以上であればそのペアを高関連，それより小さければであれば無関連と判断する．続いて，この判断の正解率をテストセット間で平均し，識別率を得る．なお，識別率は，閾値 t の値に依存するため，t を変化させて識別率が最大となる値を採用する．

4.3　相関係数

ここで用いる相関係数は，基準類似度と評価対象類似度との相関係数である．相関係数計算用テストセットとして，まず 4,000 個の語のペアを無作為に生成し，そこから基準類似度の高いペア 2,000 個を抽出した．続いて，テストセット中の各ペアについて基準類似度と評価対象類似度を計算する．この相関係数が高いほど，得られた結果は WordNet に類似しており，したがって類義語獲得の性能は高いと言える．

5　実験

5.1　手順

コーパスには，WordBank[3](約 19 万文，500 万語収録) を用いた．使用したツール，名詞・動詞の共起抽出手法については，文献 [6] と同様である．結果として得られた共起の総数は 702,879 個である．ここでは，5 分割の交差検定を用いて perplexity と性能評価指標との相関を調べた．具体的には，コーパスを 5 等分し，そのうち 4 つを言語モデルの学習セットに，残る 1 つを perplexity 計算用のテストセットとする．学習セットから言語モデルを学習し，そのモデルを用いて類義語自動獲得を行い，4 節で述べた 2 つの性能評価指標を計算する．同時に，学習された言語モデルとテストセットを用いて perplexity を求める．

このとき，LSI については，圧縮後の次元数 K を 50 から 500 まで 50 刻みで変化させ，それぞれの場合につき perplexity と性能評価指標を計算した．

また，PLSI については，潜在意味クラス数 K と TEM アルゴリズム [8] の逆温度 β の 2 つのパラメータを無作為に変化させ，様々な性能の言語モデルを得た．K については 50 から 500 の間を 50 刻みで，β については 0.7 から 1.0 の間を 0.01 刻みで設定した．

5.2　結果

図 1 は，性能評価指標である識別率・相関係数と perplexity をそれぞれ横軸，縦軸に設定し，LSI，PLSI に対応する点をそれぞれプロットしたものである．VSM については，perplexity158.4, 識別率 59.2%, 相関係数 0.062 となった．

この結果から，LSI については，圧縮後の次元数 K が perplexity および性能に与える影響は小さいことが分かる．相関係数については，これまでの研究 [1], [8] と同様，言語モデルの性能と perplexity との間には明確な負の相関 (相関係数 $R = -0.95$) が確認できる．しかし，LSI モデルの識別率については，他の結果と振る舞いが異なっており，perplexity と強い正の相関

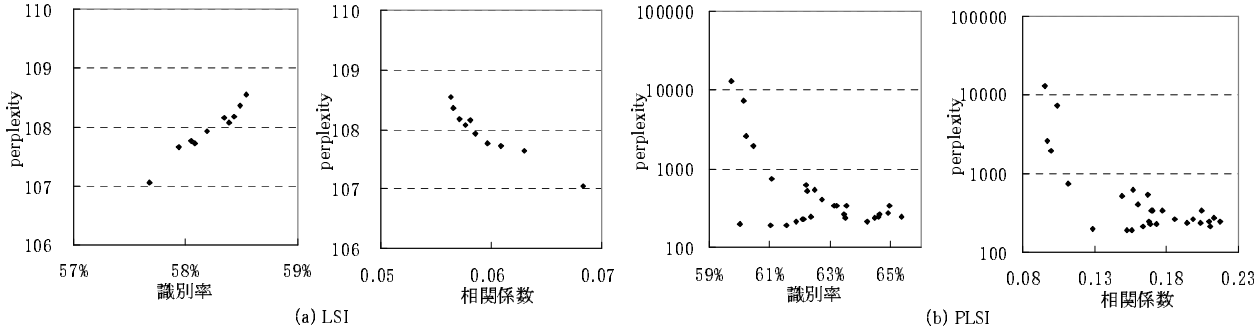

図 1: 性能評価指標と perplexity との相関

($R = 0.98$) を示している．この原因としては，識別率と相関係数では性能評価の観点が異なるという点が挙げられる．具体的には，VSM や LSI において類似度計算に用いられる余弦は，変化幅が小さく [2]，語の関連性の識別には適していない可能性があると考えられる．

LSI は VSM に比べて perplexity は低いものの，類義語獲得の性能には大きな差はない．また，PLSI と比較した場合は，perplexity は低いにも関わらず性能は低い．したがって，異なる言語モデル間においては，perplexity による言語モデルの評価と，性能評価指標は必ずしも対応しないことが分かる．

一方，PLSI については，perplexity と性能評価指標との間には負の相関 (perplexity-識別率間 $R = -0.52$, perplexity-相関係数間 $R = -0.58$) が確認できた．具体的には，perplexity が高い場合には常に性能が低いことが分かる．このことから，PLSI におけるパラメータ調節などの際には，perplexity を最小化することにより，極端に低い性能を回避することができると言える．

6 おわりに

本稿では，言語モデルの性能と perplexity との間に負の相関が存在することを実験的に確認することにより，類義語自動獲得に対する性能評価指標としての perplexity の可能性を示した．特に PLSI を適用する際には，調節すべきパラメータの数が多いため，本手法による簡易評価が特に有効であると考えられる．

一方，LSI と PLSI など，異なる言語モデルの間の比較については，perplexity は類義語自動獲得の性能を必ずしも正確に評価できるとは限らないことを明らかにした．これは，モデル変換手法や類似度指標，性能評価の観点などに依存する部分が大きいためであると考えられ，今後引き続き検討する．

参考文献

[1] Azzopardi, L., Girolami, M., Rijsbergen, K.: Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures. *Proceedings of 26th ACM Conference on Research and Development in Ingormation Retrieval (SIGIR)*, pp. 369-370, 2003.

[2] Coccaro, N., Jurafsky, D.: Towards better integration of semantic predictors in statistical language modeling, *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, pp. 2403-2406, 1998.

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

[4] Deerwester, S., Dumais, S., Furnas, G., Landauer, K., Harshman, R.: Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp.391-407, 1990.

[5] Fellbaum, C.: *WordNet: an electronic lexical database.* MIT Press, 1998.

[6] Hagiwara, M., Ogawa, Y., Toyama, K.: PLSI Utilization for Automatic Thesaurus Construction, *Proceedings of The Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, pp.334 - 345, 2005.

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

[8] Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol. 42, pp.177-196, 2001.

[9] 小島一秀, 渡部広一, 河岡司: 関連度における共通閾値の存在と応用, 第 3 回情報科学技術フォーラム (FIT2004) 講演論文集, F-003, 2004.

[10] Lee, L.: On the Effectiveness of the Skew Divergence for Statistical Language Analysis, *Artificial Intelligence and Statistics 2001*, pp. 65-72, 2001.

[11] 長尾真編: 自然言語処理, 岩波講座ソフトウェア科学 15, 岩波書店, 1996.