自 然 言 語 処 理 153－6
（２００３．１．２０）

PLSAによる確率的概念空間の評価

持橋大地, 松本裕治
奈良先端科学技術大学院大学 情報科学研究科
daiti-m@is.aist-nara.ac.jp
matsu@is.aist-nara.ac.jp

概要:

本報告では, 語彙の意味的概念の空間内での表現に関し, 空間の性質によらない評価基準を示し, 確率的表現が従来のベクトル空間での表現より優れていることを見る. また, 計算量上問題となる概念空間の次元数に対し, AIC による最適次元数の決定を試みた.

キーワード : 語彙, 意味マップ, LSI, PLSI, AIC, MDL.

Evaluation of Probabilistic Semantic Space with PLSA

Daichi Mochihashi and Yuji Matsumoto
Graduate School of Information Science, NAIST
daiti-m@is.aist-nara.ac.jp
matsu@is.aist-nara.ac.jp

Abstract

This paper proposes a neutral metric of semantic coherence independent of the inherent property of semantic spaces. Using this metric, we show that PLSA representation excels than LSA, and try to find an optimal setting of semantic dimensions based on minimum description length criterion.

Keywords: Lexical representation, Semantic Map, PLSI, LSI, AIC, MDL.

1 はじめに

言語モデルにおける予測や意味的多義性解消, 文書分類問題での語の分類などを考える際, 各単語の持つ意味的概念をある空間上への写像として表現し, その近接性や単語集合の空間内での分布を捉えることは基礎的な技術と考えられる.

語の意味には連辞的なものと範列的なものがあり [1], 前者が文法的な概念, 後者が (いわゆる) 意味的な概念に対応しているが, ここでは後者に注目したい. 実際の分類問題や文脈上の予測においては, 構文解析や係り受けによって捉えられる文法的関係以上に, それらのモデル単独では単に品詞や Named Entity, あるいは単語クラスへの一致 [2] として把握される各語の, 実際の意味的な局所性とその一致が問題となり, その意味的な局所性は世界の複雑さを反映して, 構文的関係以上に多様な広がりを持つからである.

このような概念的関係は意味ネットワーク, あるいは意味グラフとして捉えることもできる [3][4][5] が, 確率モデルとの親和性を考え, ここでは連続的な空間として考える.

2 Latent Space models

言語学的には, 語の範列的な[1] ‘意味’ とは他の全ての語との関係によって捉えられる.

情報検索の分野で行われているように, 語 w を

[1] Sapir が述べているように, 連辞的な概念が単語に凝縮して表現されることもある.

要素とする語彙集合を V とし, 語 $w, w' \in V$ の共起回数を $N(w, w')$ とすると, この意味的関係は (文書を軸とした) 語間共起関係行列 $S = \{s_{ij} = N(w, w') | w, w' \in V\}$ によって表すことができる. また, $N = \sum_{w,w' \in S} N(w, w')$ とする.

これは, $V \times V$ 上のある未知の確率分布 P を持つ定常無記憶情報源からの長さ N のサンプル $\{X_1 = \omega^1, \ldots, X_N = \omega^N | \omega^i = (w, w') \in V \times V\}$ と考えられるが, この行列には次のような問題がある.

I. 要素のスパースネス

共起の取り方にもよるが, たとえば読売新聞コーパス 1 月分 (1994/12 月) の意味的共起を文単位で考えた場合, S の要素充填率は 0.815% (総語彙数 $|V| = 36589$, 非零要素数 $= 10906588$) であり, bigram モデルでも問題となる過疎性がここでも現れる.

II. 機能 (的) 語の問題

ある名詞 n に対し, $(n,'$は$')$, $(n,'$こと$')$ のような共起は非常に多く起こると考えられるが, これは必ずしも意味的関係を反映しない. 'は', 'こと' のような機能語をストップワードとして明示的に除くことは主観に基づくため網羅的な把握は難しく, また「時」「内」「する」のような半-機能語の影響を考慮しないこととなるため, 概念表現の精度の悪化を招くと考えられる.

III. 共起の定義の問題

文書分類は情報検索などのタスクにおいては, 文書-単語間の関係が直接的に重要であるため, 文書を軸に共起を考えることは自然であるが, 意味的曖昧性解消や言語モデルなどのストリーム (文脈) を基とするタスクにおいては文書を軸とすることに必然性はない. 一般に文書の長さには制限がないため, 節や段落によって区切られ, トピックが変化する長い文書に対しても, そこに出現する語がすべて同じ意味的軸を持つと考えることは単語の概念表現としては精度の悪化を招くと考えられる.

この問題に対し, S を潜在的な空間に圧縮して考えるのが LSI (Latent Semantic Indexing) [6], PLSI (Probabilistic Latent Semantic Indexing) [7] などの Latent Space model である. 以下で, 本論文で比較対象とする LSI, PLSI の各モデルについて述べる.

2.1 LSI/LSA

LSI (Deerwester et al. 1990) はベクトル空間における主成分分析 (SVD/PCA) を基本とする情報検索の手法である. LSI では, 単語-文書の共起行列 A を考え, 特異値分解 (SVD) によって

$$A = T \Sigma D^T \tag{1}$$

と分解し, このうち大きい方から k 個の主成分を用いて

$$A \simeq T_k \Sigma_k D_k^T \tag{2}$$

と A を再構成することでノイズを除き, 各単語および文書の内容を k-次元ベクトル空間上の座標として表現する.

われわれの単語-単語正方行列 S の場合はこれは通常の主成分分析となり,

$$S = U_k \Sigma_k U_k^T \tag{3}$$

となる (LSA, Latent Semantic Analysis [8]). ここで Σ は固有値行列, U は固有ベクトルを要素とする正規直交基底行列である.

このとき, 各語 w_i の持つ '概念' は U_k の i 番目の行ベクトル $s(w_i) = [u_{i1}, u_{i2}, \ldots, u_{ik}]$ で表され, その間の距離は cosine 距離

$$\begin{aligned} \sigma(w_i, w_j) &= \cos(s(w_i), s(w_j)) \\ &= s(w_i) \cdot s(w_j) / |s(w_i)||s(w_j)| \end{aligned}$$

として通常測られる [9].

LSI では PCA を用いた固有ベクトル空間への縮退により I. の問題を解決しているが, II. の問題に関しては有効な方法を持っていない. ベクトル空間の線形変換である LSI では, 単語に対応する各次元は同じ重みを持つと仮定することになるため外部に単語の適切な重みづけが不可欠になる [10].

LSI では通常, a) idf [11] による高頻度語の重みづけ b) 頻度に基づく低頻度語の切り捨てを用いて, 各軸がほぼ同じ重みを持つように調整していると考えることができる. しかし, これは擬似的なものであり, 最適であることは保証されていないために性能の悪化を招くと考えられる.

2.2 PLSI/PLSA

PLSI(Hofmann, 1999) はこれに対し, 文書-単語の共起関係を確率的に捉え, 文書 d と単語 w

の共起確率 $P(d,w)$ を潜在変数 (意味クラス) $z \in Z = 1,2,\ldots,k$ に関して (4) 式のように分解して考える (Aspect Model).

$$P(d,w) = \sum_{z \in Z} P(d|z)P(w|z)P(z) \qquad (4)$$

このとき, 語 w の重みは共起確率 $P(d,w)$ の中に包摂されるため, 高頻度語・低頻度語に対する特別な重みづけは必要とされない. われわれの単語-単語行列 S の場合, S を共起確率行列とみて正規化した $P(w_i,w_j) = N(i,j)/N$ に対し, (4) を対称化して

$$P(w_i,w_j) = \sum_{z \in Z} P(w_i|z)P(w_j|z)P(z) \qquad (5)$$

として表現でき (PLSA, Probabilistic Latent Semantic Analysis), パラメータ $P(z), P(w|z)$ は次の EM アルゴリズムによって推定することができる.

E step:

$$P(z|w,w') = \frac{P(w|z)P(w'|z)P(z)}{\sum_z P(w|z)P(w'|z)P(z)}$$

M step:

$$P(z) = \sum_{w,w'} P(c|w,w')N(w,w')/N \quad (6)$$

$$P(w|z) = \frac{\sum_{w'} P(z|w,w')N(w,w')/N}{P(z)} \quad (7)$$

このとき, Bayes の定理から

$$P(z|w) = P(w|z)P(z)/P(w) \propto P(w|z)P(z) \qquad (8)$$

であるため, k-次元の多項分布 $P(z|w)$ を語 w の持つ '概念' として計算することができる. 図 1 に, 読売新聞コーパスから PLSA によって計算された語の概念分布の一部を示す.

なお, (8) 式から, w の '概念' $P(z|w)$ は意味クラス $z \in Z$ に対する事前分布 $P(z)$ と事後分布 $P(w|z)$ の積で表される. (6) 式によって求まる $P(z)$ はデータ全体の期待値として計算によればほぼ一様な分布を持つが, 文脈上, この事前分布が入力によって変化するものと考えると, 特定の意味クラスが強調される文脈の下では語 w の '概念' $P(z|w)$ も変化する可能性がある.

語 w の概念を $s(w) = \{P(z|w)|z \in Z\}$ としたとき, その間の距離 (意味的類似度) は KL-ダイバージェンスを用いて

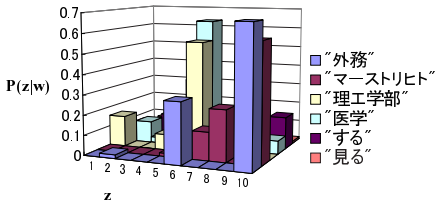


図 1: 語の概念確率分布.

$$\begin{aligned}\sigma(w,w') &= exp(-D(w||w')) \\ &= exp(\sum_{z \in Z} P(z|w) log \frac{P(z|w')}{P(z|w)})\end{aligned}$$

として表現することができ, 類似確率としての意味を持つ [12].

3 意味マップの比較

2. で述べたように, PLSI は LSI の持つ単語の重み付けの問題を自然にモデルに組み入れたという点で, 単語の概念表現に対してよい基準を与えると思われる.

単語の概念を LSI では k-次元固有ベクトル空間, PLSI では $(k-1)$-単体内の座標で表しているという点で, 両者は語に対するいわゆる '意味マップ' を与えていると解釈できるが, 両者は定義される空間が異なり, 特に LSI は確率モデルではないため, パープレキシティなどによる直接の比較は困難である.

ここで「よい意味マップ」の性質を考えてみると, それは自然な要請として

> 意味的に近い語は意味マップ上の距離も近い ... (i)

という基準を満たすものと考えられる.

そこで,「意味的に近い」ことの先験的な基準として

> 同じ文書に表れる語は意味的に近い ... (ii)

ことを仮定する. これは文書を軸とする情報検索における基本の仮定である[2].

[2]しかし, これが単語の概念表現としては必ずしも最適ではないことを 5 節で示す.

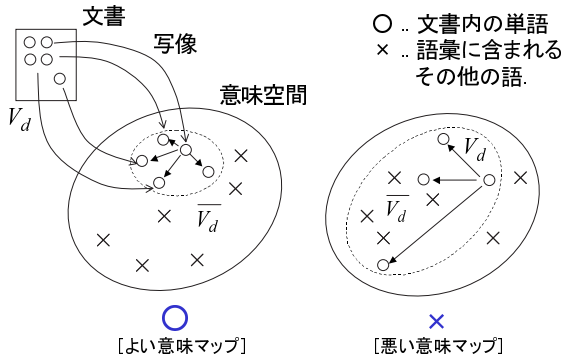


図 2: 意味空間と文書.

いま, 文書 d に含まれる語の集合を $V_d \subset V$, それ以外の語彙全体を $\overline{V_d} = V \backslash V_d$ とする.

意味マップによって V_d を意味空間上に写像した時, 'よい' 意味マップは図 2 左のように V_d が空間内で, $\overline{V_d}$ の分布に対して集中しており, 逆に '悪い' 意味マップは図 2 右のように V_d が集中せず, $\overline{V_d}$ と均一に散らばっているものと思われる.

われわれには空間への写像と, そこでの (相対的な) 距離尺度が与えられているので, この「まとまり具合」を次のように定式化する.

ある語 w からみた他の語 w' との距離を $d(w, w') = \sigma(w, w')$ とし[3], $w, w' \in V_d$ に対し, $d(w, v)(v \in V)$ を小さな順に並びかえた内での $d(w, w')$ の順位を

$$r(w, w') \in \{0, 1, \ldots, |V| - 2\}$$

とする.

このとき, d に含まれる $|V_d|(|V_d| - 1)$ 個の語のペア $(w, w') \in V_d$ に対し, $r(w'|w)$ の平均値

$$\bar{r}(d) = \frac{1}{|V_d|(|V_d| - 1)} \sum_{w, w' \in V_d} r(w, w')$$

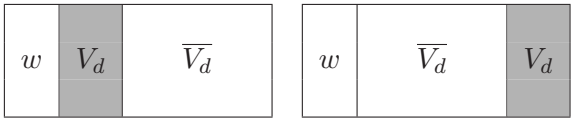
は, 意味マップ上のまとまりが最良の場合 (図 3(a)) に $|V_d|/2$, 最悪の場合 (図 3(b)) に $|V| - 1 - |V_d|/2$ に近くなるはずである (平均 $(|V| - 1)/2$).

この平均値は相対的なものであるため, 概念空間の数学的な性質と距離の絶対値に依存せず, 空間と距離尺度のペアに対して, その中での文書の意味的まとまりを表現する.

3.1 実験

LSA および cosine 距離, PLSA および KL-距離に対し, (i),(ii) の仮定の下で実験を行った.

[3]cosine 距離は対称性を持つ真の距離だが, KL-距離には '視点' が存在するため対称ではない.



(a) 最良の場合　(b) 最悪の場合

図 3: 意味マップ上の散らばり.

実験には Penn Treebank Wall Street Journal コーパス [13] から 2000 記事をトレーニングデータ, 500 記事をテストデータとし, 5-fold クロスバリデーションで計算を行った. PLSA, LSA ともに意味クラス数 $k = 32, 64, 128$ を用い, LSA に対しては単語の重みづけとして tf·idf を用いた.[4]

$\bar{r}(d)$ の計算に対しては, 品詞タグをもとに名詞・動詞・形容詞・副詞の内容語を対象とした.

3.2 結果

図 4 に空間と意味クラス数 k を変化させた時の $r(w, w')$ のヒストグラム, 表 1 に平均順位 $\bar{r}(d)$ を示す. $\bar{r}(d)$ は文書に含まれる任意の 2 単語間の意味距離の, 全語彙中での順位の期待値を表す.

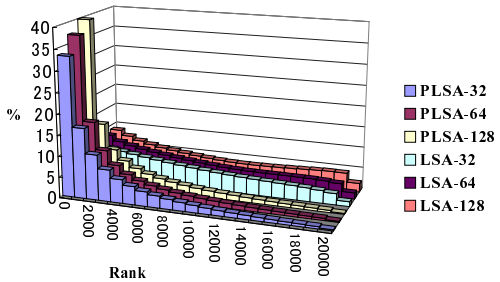


図 4: PLSA と LSA によるランクヒストグラム.

意味マップがランダムな場合, 全語彙数に対する割合は $(|V| - 1)/2 = 50\%$ であり, どちらも意味的なまとまりを反映しているが, LSA では平均ランクがなだらかに分布しており, 文書中の意味的距離が正しく評価されない場合が多いことを示している. PLSA の場合は文書中の意味的距離が全語彙中の順位で 5%以内に入る割合が 40.54%, 10%以内に入る割合が 55.64% ($k = 128$) と非常に高く, クラス数に比例して精度が上昇していることがわかる.

[4]文書集合に対する語の生起エントロピーを用いる方法 [14] では, tf·idf より良い値を示さなかった.

	k	$\bar{r}(d)$	(%)
PLSA	32	3714	(17.95%)
	64	3459	(16.71%)
	128	3507	(16.94%)
LSA	32	8303	(40.11%)
	64	8570	(41.40%)
	128	8934	(43.16%)

表 1: $\bar{r}(d)$ と全語彙数に対する%.

4 情報量基準による次元決定

PLSA + KL-距離による意味的距離が, 文書の意味的まとまりを表す写像として LSA + cosine より優れていることを見たが, PLSA の欠点としてその計算量の大きさが挙げられる.

PLSA は式 (6) から, 語彙 V に対して $O(k|V|^2)$ の計算量が必要となり, $|V| = 10,000 \sim 100,000$ 程度の場合, 共起行列 S の充填率を 1% としても 1 ステップあたり $k \times 1,000,000 \sim 100,000,000$ 回程度の演算を必要とする. k が大きいほど KL-距離の意味で実際の確率分布 P をより正確に近似することができるが, 過学習を抑え, 現実的な速度で解析を終えるためには, クラス数 k を適切な大きさに抑える必要がある.

4.1 AIC

このための基準として, われわれは AIC [15] を用いた. PLSA では, 推定すべきパラメータは $|P(z)|+|P(w|z)| = (k-1)+k(|V|-1) = k|V|-1$ 個存在し,

$$\begin{aligned} l_{AIC}(S) &= -\sum_{(w,w')\in S} N(w,w')\log P_\theta(w,w') \\ &\quad +k|V|-1 \qquad (9) \\ P_\theta(w,w') &= \sum_{z\in Z} P(w|z)P(w'|z)P(z) \end{aligned}$$

である.

情報量基準としては BIC[16](=MDL[17]) が一致性を持つ, より良い推定量であることが知られているが [18], BIC を用いることができなかった理由は以下である.

PLSA の場合, BIC による符号長は

$$\begin{aligned} l_{BIC}(S) &= -\sum_{(w,w')\in S} N(w,w')\log P_\theta(w,w') \\ &\quad +\frac{k|V|-1}{2}\log N \qquad (10) \end{aligned}$$

となり, パラメータ長 $k|V|-1$ に $\log N/2$ の重みが加わるが, 自然言語のようなオープンクラスな情報源の場合, 一般に語彙の数 $|V|$ はサンプル長 N のある関数として増加するため,(10) 式ではパラメータの記述長を加重評価することになってしまう. 実験でも, BIC を用いた場合, 符号長がクラス数に伴って線形に増加し, 最適点を見出だすことができなかった.

4.2 実験と結果

必要となる意味クラス数 k は, 情報源からのサンプル数 N に従って大きくなると考えられる. 図 5 に, 読売新聞コーパスのデータ (1994 年 1 月, $N = 25,016,634$) について, いくつかの N に対して AIC を計算した結果を示す. どの N に関しても, 共起はパラグラフ単位で取り, 頻度 5 以上の語を対象として充分に収束する同一の EM ステップ (75 回) まで計算している.

N が 10 倍になっても AIC による最適クラス数の増加は 2 倍程度であり, 線形には増加しないことがわかる.

情報検索の分野では, ヒューリスティックな値としてベクトル空間の次元を 200~300 次元程度に取ればよいことが知られているが, われわれの単語間共起行列 S の場合, S の次元である語彙数がデータ長の増加に対して限られていることから, より少ない次元で情報を表現できることが期待できる.

5 共起スキームと意味マップ

最後に, III. の問題について実験を行った.

共起の取り方としては, 1) 文書単位 2) 文単位 3) 窓関数 (方型窓, 前後の窓幅= 50, 100, 200) を用い, Penn Treebank コーパスから 1000 記事をトレーニングデータ, 500 記事をテストデータとして 3.1 節の方法で評価を行った.

概念表現の精度の問題以外に, これらの取り方の違いにより共起行列 S の非零要素であるデータ量に大きな差が生じ, 時間および空間計算量とのトレードオフが問題となる.

表 2 に順位の平均値と共起行列の充填率を示す.

わずかな差はあるが, データ量に比べて共起単

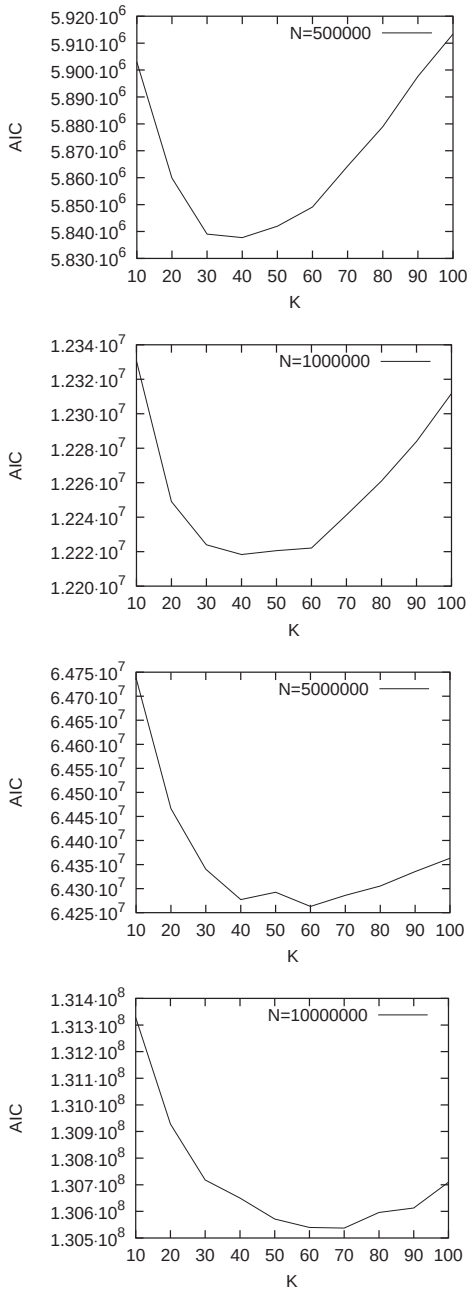


図 5: サンプル長 N と AIC.

	平均順位	データ長	(充填率)
文	25.64%	2231312	(5.69%)
窓 (25)	25.38%	3588905	(9.15%)
窓 (50)	25.32%	5228944	(13.33%)
窓 (100)	25.15%	7249940	(18.48%)
窓 (200)	25.24%	9545871	(24.34%)
文書	25.24%	13862300	(35.34%)

表 2: 平均順位とデータ量.

位を大きく取ったときの精度の上昇は小さく, 意味的なまとまりとして文単位を考えることにはある程度妥当性があるといえる.

Beeferman らは, 単語間の意味的関連性の表示として文書中に表れる同一の語の文脈上での繰り返しに注目し, その距離別の頻度を計算した結果, 距離に関して指数的な減衰を持つ分布となることを示した [19]. 小嶋 [20] は意味グラフ上の活性拡散 [21] を用いた意味予測モデルにおいて, 場面分割のための窓関数として方型窓, 三角窓, ハニング窓の中でハニング窓が最もよい性能を示したことを報告している.

共起数の取り方を確率的に考え, 最適な窓関数を求めることは今後の課題としたい.

6 まとめと課題

概念空間への '意味マップ' の良さを測る尺度として, 空間の性質に依存しない基準を提案し, 確率モデルによる概念表現がその基準において優れていることを示した. また, 最適な次元の決定法として MDL を単純に適用することができないこと, 情報量基準として AIC を用いた場合, 最適な次元数はデータ長に対して緩やかに増加することを示した.

以上の報告は意味評価尺度における発見的なものであり, 厳密なモデルに基づいた導出であるとはいえない. 今後の課題として, 今回得られた知見を基に, 確率モデルとしてより厳密に意味的基準を評価することを目指したい.

参考文献

[1] Ferdinand de Saussure. **一般言語学講義**. 岩波書店, 1972.

[2] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L.Mercer. Class-based n-gram models of natural language. In *Computational Linguistics*, volume 18, pages 467–479, 1992.

[3] M. R. Quillian. *Semantic Information Processing*, pages 216–270. MIT Press: Cambridge, MA., 1968.

[4] Beate Dorow Dominic Widdows. A graph model for unsupervised lexical acquisition. In *COLING 2002: The 19th International Conference on Computational Linguistics Vol.2*, pages 1093–1099, 2002.

[5] Jaz Kandola, John Shawe-Taylor, and Nello Cristianini. Learning semantic similarity. In *NIPS 2002*, 2002.

[6] S. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[7] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proc. of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR '99)*, pages 50–57, Aug 1999.

[8] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[9] M.W. Berry and M. Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM Book Series: Software, Environments, and Tools. June 1999.

[10] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. pages 159–168, 1998.

[11] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372, 1973.

[12] 持橋大地, 松本裕治. 意味の確率的表現. **情報処理学会研究報告** *2002-NL-147*, pages 77–84, 2002.

[13] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English:The Penn Treebank. In *Computational Linguistics, Vol.19, No.2 *Special Issue on Using Large Corpora: II*, pages 313–330, 1993.

[14] Noah Coccaro and Daniel Jurafsky. Towards Better Integration of Semantic Predictors in Statistical Language Modeling. In *Proc. of ICSLP '98*, volume 6, pages 2403–2406, 1998.

[15] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AP-19:716–723, 1974.

[16] Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

[17] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[18] 韓太舜, 小林欣吾. **情報と符合化の数理** [対象 11]. 岩波講座 応用数学 13. 岩波書店, 1994.

[19] Doug Beeferman, Adam Berger and John Lafferty. A Model of Lexical Attraction and Repulsion. In *Proc. of ACL-EACL '97, Madrid, Spain*, pages 373–380, July 1997.

[20] 小嶋秀樹, 古郡延治. 単語の結束性にもとづいてテキストを場面に分割する試み. **情報処理学会自然言語処理研究報告** *95-7*, pages 49–56, 1993.

[21] A.M. Collins and E. F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, (82):407–428, 1975.