自 然 言 語 処 理 147－12
（２００２．１．２２）

意味の確率的表現

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:

本論文では, 情報検索の分野で提案された PLSI (Probabilistic Latent Semantic Indexing) の方法を拡張した Semantic Aggregate Model を提案し, 単語の持つ意味の概略を最尤推定の立場から k- 次元の確率分布によって表現する.
この表現によって, 従来ベクトル空間モデルによって経験的に扱われてきた '意味' を数学的に見通しよく扱うことができる. 関連して, 単語間の意味的な距離, 意味的重みについての新しい指標を提案する.

キーワード: 語彙意味表現, PLSI, LSI, EM アルゴリズム, 統計力学

Probabilistic Representation of Meanings

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 Semantic Aggregate Model on word meanings by extending an Information Retrieval model PLSI (Probabilistic Latent Semantic Indexing). Through the maximum likelihood estimation, this model renders approximate meanings of a word with a discrete probability distribution on latent classes. By this representation, the semantic distance and semantic weights of words can be reformulated mathematically.

keywords: Lexical meanings, PLSI, LSI, EM algorithm, Statistical Mechanics

1 はじめに

計算機上の言語処理環境の高度化により，ますます言語の意味的な処理の必要性が高まっている．しかし，従来の意味記述のように論理的な素性を人手で列挙する方法では，常に新しく生成され，変化する「意味」に対応できない．また，個人毎に持つ意味の差異も表現できないため，言語データから自動的に言葉の意味的要素を学習して抽出し，計算的に表現できることが望まれる．言語空間は均質ではなく，各個人は接する言語的環境＝データによって同じ言葉にも異なった「意味」のイメージを持っているからである．

本論文では，このための方法として，PLSI (Hofmann99) の方法を拡張した Semantic Aggregate Model (SAM) を提案し，「意味」の概略を k- 次元の (離散的) 確率分布によって表現する．この表現によって，従来不可知な，あるいはベクトル空間上の座標として Heuristic に表現されてきた「意味」が，確率分布の全体のなす情報幾何的空間 [1] の問題に還元される．

以下ではまず，PLSI の基となる情報検索のモデルである LSI の概念とその方法，および問題点について述べ，次に PLSI による情報理論的再構成について述べる．次に情報検索モデルとしての PLSI を基に，語彙意味表現へ拡張する Semantic Aggregate Model を提案し，EM アルゴリズムを用いた推定式を示す．

4 章では，この表現を踏まえ，応用として，cosine 距離に代わる KL-Divergence を用いた情報理論的な単語間類似度，および inverse document frequency (idf) に代わる意味的な重みとしての Semantic Entropy について述べる．最後に 5 章で LSI との比較を行って今後の課題を明らかにし，6 章で今後の展望についてふれる．

2 Latent Space Model

Saussure[2] も述べているように，記号は記号との関係によってその意味が定まる[1]．記号間の意味的関係は共起関係として表れるが [4]，記号間の共起行列はきわめて疎である上，共起頻度は必ずしも意味的関係の強さに比例しないため，それ自身を直接意味的指標とすることはできない．この事態に対し，記号の共起関係をもたらす潜在的な意味空間を考えることによって解決するのが LSI，PLSI などの Latent Space Model である．

2.1 LSI

LSI (Latent Semantic Indexing) は，情報検索の分野で Dumais[5] らによって提案された意味的情報検索の手法である．LSI では，単語—文書の共起行列 A を特異値分解 (SVD) によって

$$A = TSD^T$$

と分解し，このうち大きい方から k 個の主成分を用いて

$$\hat{A} = T_k S_k D_k^T \tag{1}$$

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- 次元の空間に圧縮する．この圧縮は最小二乗誤差の意味でもとの行列 A にもっとも近い射影であることが示されている [6]．

式 (1) は，線形空間において図 1のような射影の変換をするものと考えることができるが，このとき，T_k は単語の k- 次元の Latent Space への射影であるから，T_k の持つ各列ベクトルを単語の持つ ‘意味表現’ であると考えることができる．また，その類似度をベクトルの $\cos\theta$ として計算することができ，多くの関連分野で用いられている．

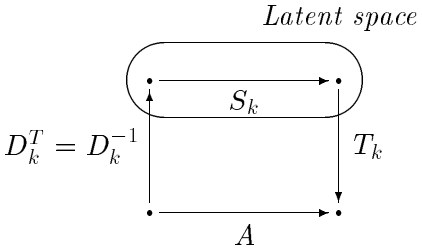

図 1: LSI の射影変換.

しかし，LSI にはいくつかの問題点が指摘されている．

1 つは，LSI で単語—文書の共起行列から適切な意味的関係を抽出するには，単語に対して事前の (アドホックな) 重みづけが不可欠だということである．「は」「する」のような機能語はどの文書とも共起頻度が非常に高いため，頻度をそのまま用いると単語間の意味的関係が正しく反映さ

[1] 「生理学者や心理学者の観点からすれば，われわれが，この [精神身体的な] 基礎に絶えず，かつ明示的に言及することもなく，ことばの問題を取り扱おうとするのは，不当な捨象を行っているように見えるかもしれない．しかし，このような捨象は正当化されるものだ．.. ことばの背後にある器官や心理のメカニズムは当然のこととして不問に付したまま，ことばの意図，形式，歴史を有益に論じることができるからである.」 —Sapir[3].

れない[2]. このための重み付け法には idf など様々な手法があるが [7], いずれもアドホックなもので数学的裏付けが薄く, しかも重みの選択によって結果が大きく影響される.

これは, LSI がベクトル空間上のモデルであり, 情報理論・統計的理論と離れていることに端を発している. 実際, 式 (1) の T_k によって得られる単語の意味ベクトルの要素は負になることもあり, 正規化ベクトルとしての意味表現は確率論とはなじみにくい.

2.2 PLSI

Hofmann[8] によって提案された PLSI (Probabilistic Latent Semantic Indexing) はこれに対して, Aspect Model [9] を基に LSI と同様の圧縮を確率的に行う手法であり, [8] においてその LSI に対する優位性が示されている.

PLSI ではベクトル空間上の空間軸ではなく, 意味的な隠れクラス変数 $c \in C$ を考え, 文書 d における単語 w の生起は, c に従って

$$P(d,w) = \sum_{c \in C} P(d|c)P(w|c)P(c) \qquad (2)$$

と確率的に起こると考える (図 2).

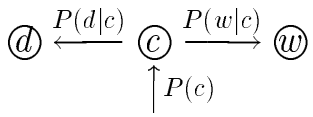

図 2: PLSI のグラフィカルモデル.

このとき, $N(d,w)$ を文書 d における単語 w の実際の観測値とすれば, データの尤度

$$L = \sum_d \sum_w N(d,w) \log P(d,w)$$

を最大にする[3] $P(c)$, $P(d|c)$, $P(w|c)$ は, EM アルゴリズムにより次式によって最尤推定することができる.

E step:

$$P(c|d,w) = \frac{P(c)P(d|c)P(w|c)}{\sum_{c'} P(c')P(d|c')P(w|c')}$$

M step:

$$P(d|c) = \frac{\sum_w N(d,w)P(c|d,w)}{\sum_{d',w} N(d',w)P(c|d',w)}$$

$$P(w|c) = \frac{\sum_d N(d,w)P(c|d,w)}{\sum_{d,w'} N(d,w')P(c|d,w')}$$

$$P(c) \propto \sum_{d,w} N(d,w)P(c|d,w)$$

過学習を避けるため, [8] では実際には Tempered EM[10] を用いて推定を行っている.

このモデルは, (2) 式を

$$P(d,w) = \sum_c P(d|c)P(c)P(w|c)$$

と書き直すとき, $\hat{U} \equiv (P(d_i|c_k))_{ik}$, $\hat{V} \equiv (P(w_j|c_k))_{jk}$, $\hat{\Sigma} \equiv \mathrm{diag}(P(c_k))_k$ とすれば, 共起確率行列 $P = (P(d_i,w_j))_{ij}$ を, (1) と同様に

$$P = \hat{U}\hat{\Sigma}\hat{V}^T \qquad (3)$$

と確率的に分解していると見ることができる.

このようにして最尤推定の下に情報理論的に再構成された PLSI は, LSI に対して常に高い性能を見せることが示されている.

3 Semantic Aggregate Model

PLSI は情報検索のモデルであるが, 語彙の持つ意味を計算的に表現することは情報検索に限らず, 情報コミュニケーション一般において基礎的で重要な課題である. 計算機上における一般の言語環境では言語が文書集合に分かれているとは限らない[4]ため, 文書によらず単語間共起関係一般に基づいた定式化を考える必要がある.

ここで, 意味的に関連のある語 w と w' が共起するとき[5], そこに共通する意味クラス (あるいは, 意味の '核') $c \in C$ が (2) と同様に存在して,

$$P(w,w') = \sum_{c \in C} P(w|c)P(w'|c)P(c) \qquad (4)$$

と生成されると考える (図 3).

このとき, $N(w_i,w_j)$ を実際のデータにおける共起回数の観測値とすれば, データの尤度

$$L = \sum_{i,j} N(w_i,w_j) \log P(w_i,w_j)$$

[2] stop list を用いる方法は言語や観点に依存すること, 境界が曖昧なことから, 連続的な指標で置換した方がよいと考えられる. 4.2節参照.

[3] KL 情報量で測ったデータとのクロス・エントロピーを最小にする.

[4] 会話のような言語ストリームや, 長く境界が曖昧な文書 (小説) など.

[5] 一文内, 一文書内など. ??節参照. 以下では文内共起を用いている.

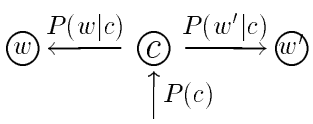

図 3: Semantic Aggregate Model.

を最大にする $P(c)$ および $P(w|c)$ は PLSI と同様に，EM アルゴリズムにより次のように最尤推定することができる．

E step:

$$P(c|w,w') = \frac{P(w|c)P(w'|c)P(c)}{\sum_c P(w|c)P(w'|c)P(c)}$$

M step:

$$P(c) = \sum_{w,w'} P(c|w,w')N(w,w')$$

$$P(w|c) = \frac{\sum_{w'} P(c|w,w')N(w,w')}{P(c)}$$

ここで

$$P(c|w) = \frac{P(w|c)P(c)}{P(w)} \propto P(w|c)P(c) \qquad (5)$$

であるから，$P(w|c)$ と $P(c)$ から単語の意味クラスへの帰属確率分布 $P(c|w)$ を求めることができ (ソフトクラスタリング)，単語の持つ意味の概要を表現することができる．

これは bigram に対する Saul ら [11] の Aggregate Markov Model $P(w_2|w_1) = \sum_c P(w_2|c)P(c|w_1)$ を意味的共起関係に拡張したものでもあるので，以下このモデルを Semantic Aggregate Model (SAM) と呼ぶ．

Pereira らは，(N,V) の共起ペアについて，クラスタ中心への KL 距離をもとに逆温度パラメータ β から階層的に同様の確率的帰属分布を求める Distributional Clustering[12] を提案している．これは語彙の過疎性を直接扱うものではないが，SAM は一般の意味的共起関係に対して，最尤推定という形で直接情報圧縮を行うものである．

このモデルは，$P \equiv (P(w_i,w_j))_{ij}$，$\hat{U} \equiv (P(w_j|c_k))_{jk}$，$\hat{\Sigma} \equiv diag(P(c_k))_k$ とおけば，(3) と同様に

$$P = \hat{U}\hat{\Sigma}\hat{U}^T$$

と PCA を確率化したものと考えることもできる．

このようにして求められた $P(c|w)$ は単語の持つ意味的性質を適切に再現する．図 (3) に，京大コーパス (38383 文，829351 語) における文内共起から計算される $P(c|w)$ の例を示す．例のために $|C| = 10$ としているが，実際にはクラス数はもっと大きい値を用いる．

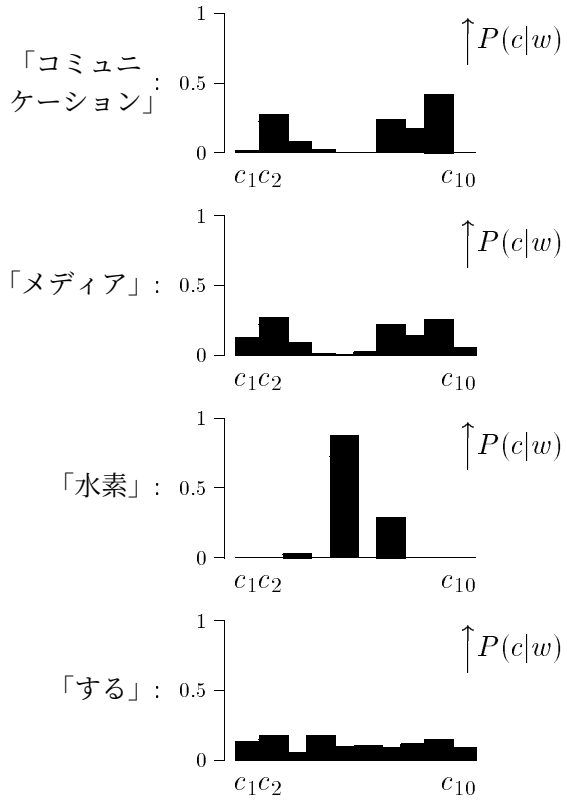

図 4: 意味確率分布.

なお，PLSI および Semantic Aggregate Model では，LSI と異なり単語に対する stop list や恣意的な重み付けは不要なことに注意されたい．

表中，意味的に共通点のある「コミュニケーション」と「メディア」は直接共起していないが，共通の意味クラス c_2, c_9 に対して高い帰属確率が与えられていることがわかる．また，機能語である「する」に対してはほとんど一様な確率分布が与えられ，これらの語が特定の '意味' を表現しないことが示されている．この分布の持つ情報量の利用については 4.2節で触れる．

4 Semantic Aggregate Model の応用

4.1 情報論的単語間類似度

(5) のように単語の「意味」が意味クラス数 $|C|$ 次元の確率分布で表されると，ベクトル空間における cosine 類似度に代わり，確率分布間の距

離尺度である KL 情報量を用いてその類似度を自然に定義することができる.

$$D(w_i||w_j) = \sum_{c \in C} P(c|w_i) \log \frac{P(c|w_i)}{P(c|w_j)}$$

ただし, KL 情報量 D は $w_i = w_j$ のとき 0, $w_i \neq w_j$ であるほど > 0 なので, 類似度としては直接扱いにくい.

ここで KL 情報量の意味から[6] $e^{-D(w_i||w_j)}$ を考えると, $p(i) \equiv P(c|w_i)$, $p(j) \equiv P(c|w_j)$ と書けば

$$\begin{aligned} e^{-D(w_i||w_j)} &= exp(-\sum_{c \in C} p(i) \log \frac{p(i)}{p(j)}) \\ &= exp \sum_{c \in C} p(i) \log \frac{p(j)}{p(i)} \\ &= exp\ E_{p(i)} \left[\log \frac{p(j)}{p(i)} \right] = E_{p(i)} \left[\frac{p(j)}{p(i)} \right] \\ &\quad (\text{ただし}, E_p \text{は} p \text{に関する期待値.}) \end{aligned}$$

ゆえ, $e^{-D(w_i||w_j)}$ は確率分布 $P(c|w_i)$ に対する $P(c|w_j)$ の一致度の期待値を表す. 従って,

$$s_I(w_i, w_j) = e^{-D(w_i||w_j)} \tag{6}$$

を意味確率分布の情報理論的な類似度と考えることができる.

なお, KL- 情報量の性質からこの類似度には「観点」が存在し, 一般に $s_I(w_i, w_j) \neq s_I(w_j, w_i)$ である. 「鯛」→「魚」の類似度は高いが, 「魚」→「鯛」の類似度はそれほど高くないと考えられることからも, この性質は妥当であろうと考えられる.

表 5に, 京大コーパスにおけるいくつかの語の類似語と KL 距離, 類似度 s_I を示す.

また, この $e^{-D(w_i||w_j)}$ は統計力学における Boltzmann 因子 [13] とも見ることができる. この関係については 5.2節で述べる.

4.2 Semantic Entropy

3章でみたように, 単語 w の意味確率分布 $P(c|w)$ は, 意味的に特徴のある語に対しては分布

[6] 意味的に考えると, $D(P||Q) = E_p[\log p - \log q]$ は確率分布 P を確率分布 Q で近似した時に失われる情報量 (ビット落ち) を表す [7]. したがって, $e^{D(P||Q)}$ を考えれば, これは失われる情報量に対する平均分岐数であり, その逆数 $e^{-D(P||Q)}$ は P を Q で近似できる確率を表すとも解釈できる.

‘イスラエル’:	s_I	KL 距離
パレスチナ	0.7327	0.3110
和平	0.7324	0.3114
パキスタン	0.7062	0.3479
カンボジア	0.6896	0.3716
パレスチナ解放機構	0.6762	0.3912
調印	0.6341	0.4555
共同コミュニケ	0.6247	0.4705
中国	0.6189	0.4797
チリ	0.5863	0.5339
カ国	0.5721	0.5585
会議	0.5695	0.5629
‘演奏’:	s_I	KL 距離
編集	0.6705	0.3997
収録	0.5879	0.5312
芸術	0.5195	0.6548
書き初め	0.5101	0.6731
コンクール	0.4822	0.7293
連載	0.4653	0.7650
紹介	0.4562	0.7849
作品	0.4476	0.8038
創刊	0.4447	0.8104
集	0.4407	0.8194
訪ねる	0.4116	0.8878
‘は’:	s_I	KL 距離
と	0.9782	0.0220
に	0.9733	0.0271
が	0.9698	0.0307
の	0.956	0.045
なる	0.9523	0.0489
を	0.9387	0.0633
する	0.9355	0.0666
だ	0.9286	0.0741
いる	0.9219	0.0813
も	0.9162	0.0875
この	0.9151	0.0887

図 5: KL 距離と類似度.

が偏るが, 機能語についてはほとんど一様な分布を持つという特徴をもつ.

そこで, その情報量

$$H(w) = -\sum_{c \in C} P(c|w) \log P(c|w)$$

を考えると, $H(w)$ は内容語に関しては 0 に近く, 機能語に関しては最大値 $\log |C|$ に近づく.

そこで, $H(w)$ の最大値との比をとって

$$e(w) = 1 - \frac{H(w)}{\log |C|} = 1 + \frac{\sum_c P(c|w) \log P(c|w)}{\log |C|} \tag{7}$$

とすれば, $0 \leq e(w) \leq 1$ であり, Semantic Entropy $e(w)$ はこの範囲で, 機能語に対して 0 に近い値を, 内容語ほど高い値を持つ.

Coccaro ら [14] は文書集合における単語の重みについて LSA Confidence (LSAC) と呼ばれる同様の指標を提案しており, Semantic Entropy はそれを単語の持つ意味カテゴリへの予測力という形で SAM から構成するものである.

図 6に, Semantic Entropy の計算例を示す.

w	$e(w)$
から	0.0118
の	0.0123
で	0.0134
に	0.0146
が	0.0191
する	0.0253
⋮	
持つ	0.1411
大きい	0.1420
場	0.1439
個人	0.2542
クラス	0.2543
基づく	0.2548
起こす	0.2555
存在	0.2551
⋮	
需要	0.501
提訴	0.501
核燃料	0.501
司法研修所	0.7643
都市ガス	0.7644
レトロ	0.7645
インスブルック	1.00
猛禽類	1.00
返り咲き	1.00
穂高岳	1.00

図 6: Semantic Entropy.

5 SAM と言語モデル

5.1 Semantic Perplexity

Semantic Aggregate Model の評価法として, 文脈に基づく意味的予測力を考えることができる. 構文的関係は n–gram で捉えることができるから [15], 学習データはすべての語[7]を用いるが, 予測の対象は自立語のみに限定し, 後方 t 個の文脈に対して

$$\begin{aligned} w_1^n &= \{w_1 w_2 \cdots w_n | w_i : \text{自立語}\} \\ P(w_1^n) &= \prod_n P(w_n | w_{n-t+1}^{n-1}) \\ &= \prod_n \frac{e(w_n) d(w_{n-t+1}^{n-1}, w_n)}{\sum_{w' \in L} e(w_n) d(w_{n-t+1}^{n-1}, w_n)} \end{aligned} \tag{8}$$

によって意味的な予測確率を計算する. ここで $e(w_n)$ は式 (7) による単語 w_n の意味的な重み (Semantic Entropy), $d(c, w)$ は文脈 c における単語 w の距離であり, われわれのモデルでは確率分布間の KL- 距離から, LSI ではベクトルの cosine 類似度から計算される. 以下, d および c の定義について述べる.

5.2 単語間類似度と統計力学的メタファ

d としては, われわれの SAM のモデルにおいては (6) から

$$d(c, w) = e^{-D(c||w)} \tag{9}$$

を, LSI においては cosine 類似度

$$d(c, w) = cos(c, w) = \frac{\vec{c} \cdot \vec{w}}{|\vec{c}||\vec{w}|} \tag{10}$$

を用いることができる.

ここで (10) において, 黒橋 [15],Coccaro[14] らは d として $\cos(c, w)$ を直接用いるのではなく,

$$d(c, w) = \cos(c, w)^\gamma \quad (\gamma = 2 \sim 7\text{程度の整数})$$

とすることで単語間距離の差が広がり, パープレキシティが大きく減少することを報告している.

[7] 意味に関する多くの研究では, 重み付けと過疎性の問題から高頻度語のみを対象としているものが多いが, 実際には低頻度語の方が大きな意味を担っており [16], 低頻度語を切り捨てることは適当ではない. SAM は単語に重み付けを用いず, 粗い文書情報ではなく前後の文脈を利用して安定した性能をみせるため, 本論文ではノイズの可能性をもつ頻度 1 以外のすべての単語を対象としている.

われわれのモデルでは，これは (9) を用いて

$$d(c,w) = (e^{-D(c||w)})^{\beta} = e^{-\beta D(c||w)} \quad (11)$$

と書けることを意味するが，(9) は確率としての解釈ももつから，式 (11) は統計力学における Boltzmann 因子 [10] であり，KL 距離 $D(c||w)$ を，w に関して文脈 c からのポテンシャルとみたとき，類似度の総和

$$N = \sum_{w} d(c,w)$$

および全類似度の，c からのポテンシャルエネルギの総和

$$E = \sum_{w} D(c||w)d(c,w)$$

を一定としたときの温度 β における最尤確率である．このとき，(8) の $\prod$ の中は Gibbs 分布

$$\frac{e(w_n)e^{-\beta D(c||w_n)}}{\sum_{w\in L} e(w)e^{-\beta D(c||w)}}$$

となり，β は系を支配する逆温度となる．直観的には，β を大きくする＝温度を下げると文脈 c の周辺の類似語分布が集中し，差が広がるといってよい．

したがって，[14],[15] において恣意的に導入された係数 γ は，われわれの確率的なモデルでは逆温度 β として再解釈することができる．

われわれのモデルにおいても，クラス数 $|C|$ が少ない場合 (5.3) 節参照)，$\beta = 5 \sim 7$ 程度におくことにより Perplexity が減少した．

5.3 確率分布としての文脈

LSI においては，文脈として後方 t 個の語 w_{n-t+1}^{n-1} の意味ベクトルの平均を取って文脈 c とする方法が一般に行われている [14, 15, 17]．しかし，確率分布として意味を考え，式 (9) によって文脈 c との距離を計算する確率的なモデルでは，c として同様に確率分布の平均を取る方法では特にクラス数 $|C|$ が大きい場合，予測確率が著しく下がることがわかった．

これは，確率分布の持つ性質に起因している．図 6に示されているように，意味的に予測されやすい高い意味を持った語は情報量が大きく，特定の数個の意味クラスに対してきわめて高い帰属確率を持つ．一方，平均を取った確率分布はきわめて 'なだらかな' 分布になり，その特定の意味クラスとの一致度の期待値は必然的に 0，あるいは小さくならざるを得ない．

このことは，LSI の各次元が持つ意味とも関係している．LSI における行列の SVD 分解 (式 (1)) では，分解後の各軸は互いに無相関の固有ベクトルとなる [6] ため，ベクトルの各成分ごとに平均を取ることには意味があるが，SAM においては各意味クラスの間には相関があり，意味クラス間の共分散行列を求めることもできる．特定のクラスとの帰属確率の不一致が影響を及ぼすのもこの理由によっている．

確率分布としての定式化から考えても，ベクトル空間のように単純に平均をとることは適切ではなく，確率論に基づいた取り扱いが求められているといえる．

このほか，文脈における時間的な相関 (「惑星」→「帰還」のような関連する語が続けて表れたとき，その意味の強化は後から上書きされにくい) も考えられ，筆者らは確率システムとしての取り扱いを検討中である．

6 考察と展望

本研究では，文書集合によらず一般の共起情報から最尤推定という形で情報圧縮を行い，確率分布として '意味' を表現する Semantic Aggregate Model を提案した．関連して，単語の重みとして意味確率分布から Semantic Entropy が計算され，KL 情報量によって定義される単語間類似度が統計力学的温度 β をパラメータとして持つことを示した．

'意味' の概要を確率分布として数学的に捉えることによって，意味相互の関係が現在研究が進んでいる情報幾何的空間 [1] の問題として捉え直される．本研究がそのための一歩となることを期待したい．

意味の問題は単語に限ってみても複雑であり，本論文のような静態的な方法ではそのすべてを尽くすことはできない．より動態的・構造的な方向へモデルを発展させること，共起によって回収されない，単語の持つ独自の意味について考えることは今後の課題である．

謝辞

EM アルゴリズムの実装について，メモリの省力化のアドバイスをいただいた松本研究室の工藤拓氏に感謝します．

参考文献

[1] 甘利俊一, 長岡浩司. 情報幾何の方法. 岩波講座 応用数学 (6). 岩波書店, 1993.

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

[3] Edward Sapir. 言語. 岩波書店, 1998.

[4] J. R. Firth. A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Oxford: Basil Blackwell, 1957.

[5] 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.

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

[7] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[8] 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.

[9] Thomas Hofmann, Jan Puzicha and Michael I. Jordan. *Learning from Dyadic Data*, pages 466–472. Number 11 in Advances in Neural Information Processing Systems. Morgan Kaufman Publishers, 1998.

[10] Kenneth Rose, Eitan Gurewitz and Geoffrey C. Fox. Statistical Mechanics and Phase Transitions in Clustering. *Physical Review Letters*, 11(9):589–594, September 1990.

[11] Lawrence Saul and Fernando Pereira. Aggregate and mixed-order Markov models for statistical natural language processing. In *Proc. of the Second Conference on Empirical Methods in Natural Language Processing (SIGDAT)*, pages 81–89, 1997.

[12] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional Clustering of English Words. In *Proc. of 31th ACL*, pages 183–190, 1993.

[13] 戸田盛和. 熱・統計力学. 物理入門コース 7. 岩波書店, 1983.

[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] 黒橋禎夫, 織学. 文脈共起ベクトルに基づく大域的言語モデル. 情報処理学会研究報告 *2000-NL-139*, pages 77–83, 2000.

[16] Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[17] 小嶋秀樹, 伊藤昭. 文脈依存的に単語間の意味距離を計算する一手法. 情報処理学会論文誌, 38(3):481–489, 1997.