自 然 言 語 処 理 139−11
（2000．9．22）

文脈共起ベクトルに基づく大域的言語モデル

黒橋 禎夫　　織 学
京都大学大学院 情報学研究科

要 旨

本稿では，文脈情報を考慮する大域的言語モデルを提案する．まず，単語・文書行列を圧縮することにより，単語共起ベクトルを作成する．単語共起ベクトルは，共起しやすい語の内積値は大きく，共起しにくい語の内積値は小さいという性質を持つ．文脈中の自立語の単語共起ベクトルの和によって文脈を表現し，これと次に出現しうる単語の単語共起ベクトルとの内積値をもとにして，文脈に対する語の出現のしやすさ，文脈共起確率を計算する．提案する言語モデルは，この文脈共起確率とトライグラムを線形補間したものである．毎日新聞コーパスを用いて実験したところ，トライグラムと比較して，テストセットパープレキシティが全体で5.0%，自立語に関して27.2%減少した．

Nonlocal Language Modeling based on Context Co-occurrence Vectors

Sadao Kurohashi　　Manabu Ori
Graduate School of Informatics, Kyoto University

Abstract

This paper presents a novel nonlocal language model which utilizes contextual information. A reduced vector space model calculated from co-occurrences of word pairs provides word co-occurrence vectors. The sum of word co-occurrence vectors represents the context of a text, and the cosine similarity between the context vector and the word co-occurrence vectors represents the long-distance lexical dependencies. Experiments on the Mainichi Newspaper corpus show significant improvement in perplexity (5.0% overall and 27.2% on target vocabulary).

1 はじめに

人間のパターンの認識は対象を孤立した単独のものとして認識することはほとんどない．風景全体の中での物体の認識，発話全体の中での単語の認識というように，時間的，空間的な様々な環境の中で，それらを条件として認識を行っている．それによって，曖昧性を排除し，ロバストな認識が実現されている．計算機によるパターン認識における最大の問題は，このような環境を，どこまで，どのような表現形式で取り込むことができるかという点にある．

言語の処理の場合には，文脈，すなわちその直前までの発話，文章などが重要な環境となる．文脈の表現として，音声認識，文字認識，形態素解析などで広く用いられているものとして確率的言語モデルがある．確率的言語モデルとは，任意の単語列の出現確率 $P(w_1^n)$ を与えるものである．ここで，w_i^j は単語列 $w_i \ldots w_j$ を表すものとする．これにより，ある時点までの単語列を条件とする次の単語の出現確率 $P(w_i|w_1^{i-1})$ が得られる．これは，ある文脈での各単語の現れやすさを意味する．

しかし，実際に用いられているのは，これを $(N-1)$ 重マルコフ過程で近似した N グラムモデルである．すなわち，ある時点での事象の確率は，その直前の $(N-1)$ 個の事象だけの影響を受けるとし，

$$P(w_i|w_1^{i-1}) \approx P(w_i|w_{i-N+1}^{i-1}) \qquad (1)$$

とする．一般には N が 2 のバイグラム，または N が 3 のトライグラムが用いられる．すなわち，直前の 1 語，または 2 語の局所的な文脈だけが考慮される．

このような言語モデルが有効な場合もある．例えば，「総理」に「大臣」「官邸」などは続きやすいが，「怪人」「鑑定」などはほとんど続かない．また，「だけ で」に「は」「も」などは続きやすいが，「が」「を」などはほとんど続かない．このように，複合名詞や付属語列での言語の性質はバイグラムやトライグラムによってうまく表現することができ，音声認識，文字認識，形態素解析などでその有効性が示されている．

しかし，例えば，「で は」の次の単語の確率分布にはほとんどかたよりはないであろうから，このような言語モデルは何も予測しないということになる．またかたよりがあったとしてもそれに意味があるとは考えられない．しかし，バイグラムやトライグラムにおいて確率分布に差がない部分であっても，たとえば政治に関する文章であれば「大臣」と「怪人」の出現確率は大きく異なるはずである．バイグラムやトライグラムのような言語モデルは，大域的な，長距離の文脈依存の関係については当然のことながら無力である．

本論文では，より長距離の，すなわち，それまでの文章全体が影響を与える言語モデルを提案する．長距離依存での重要な情報は単語間の共起のしやすさ，たとえば「政治」「政府」「行政」「省庁」などの政治に関する単語と「大臣」が共起しやすいという情報である．しかし，文章中の個々の語との共起のしやすさは簡単に計算できるが，それは文章全体，文脈との関係ではない．そこで，単語間の共起のしやすさを重ねあわせることで，文脈と単語の共起のしやすさを表現し，これを言語モデルに反映する．

2 単語共起ベクトル

2.1 単語・文書間の出現行列

語の共起関係を最も直接的に表現するものとして，図 1のように，一方 (縦) に単語，一方 (横) に文書 (例えば新聞の一記事など) をおき，ある単語がある文書に現れたときに対応する要素を 1 にする，単語・文書行列を考えることができる．

この行列の各行は，各単語の共起に関する情報をベクトル表現したものとみることができる．同じ文書に現れやすい，すなわち共起しやすい語は，このベクトルが類似している，すなわちベクトル間のなす角が小さいということになる．

このような方法によって，できるだけ正確な共起情報を得ようとするならば，できるだけ多数の文書を調べる方がよいということがいえる．そうすると単語・文書行列の横方向は非常に大きくなる可能性がある．実際，現在では膨大な数のオン

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
w_1	1	0	1	0	1	0	1	0
w_2	1	0	1	1	0	0	0	0
w_3	0	1	0	0	1	1	0	1
w_4	1	1	1	0	0	0	0	0
w_5	0	0	0	0	1	0	1	0
w_6	0	0	0	0	1	0	0	1

図 1: 単語・文書間の出現行列

	w_1	w_2	w_3	w_4	w_5	w_6
w_1	4	2	1	2	2	1
w_2		3	0	2	0	0
w_3			4	1	1	2
w_4				3	0	0
w_5					2	1
w_6						2

図 2: 単語間の共起の三角行列

ラインテキストが利用可能であるので，横方向は何百万，あるいはそれ以上になることもある．そこで，この行列をそのまま扱うということは現実的でなく，行列の横方向を圧縮し，より少ない次元でもとの行列の性質を近似するということを行う必要がある．

2.2 単語・文書行列の圧縮

単語・文書行列の使用目的は，単語間の共起のしやすさを行ベクトル間の角度として表現することである．そのため，行列の圧縮では，もとの行列における行ベクトル間の角度と，圧縮された行列での同じ行ベクトル間の角度ができるだけ近くなる，という必要がある．

このような行列の圧縮として，情報検索の分野においてHNCによって提案された以下の学習アルゴリズムを用いる[5]．

1. 実際には，文書集合から単語・文書行列を作成することはしない．そのかわりに，行，列を単語とし，対角要素には各単語の出現文書数，他の要素には各単語ペアの共起文書数を与える三角行列を作成する (図 2)．単語 w_i の出現文書数を $F(w_i)$，単語 w_i，w_j の出現文書数を $F(w_i, w_j)$ と表すことにする．

   単語・文書行列における重要な情報は，単語 w_i の行ベクトルと単語 w_j の行ベクトルの間の角度の cosine の値 (類似度) であるが，この値は三角行列から次のように計算することができる．

$$\alpha_{ij} = \frac{F(w_i, w_j)}{\sqrt{F(w_i)}\sqrt{F(w_j)}} \qquad (2)$$

   $\sqrt{F(w_i)}$ は単語・文書行列における単語 w_i の行ベクトルの大きさ，$F(w_i, w_j)$ は単語・文書行列における単語 w_i の行ベクトルと単語 w_j の行ベクトルの間の内積値に相当するからである．

2. 圧縮する行列の (横方向の) 次元を決め，その行列の初期化を行う．まず，各要素の値を -1 から 1 までの正規分布に従う乱数で与え，次に各ベクトルの大きさを 1 に正規化する．正規分布に従う乱数は，例えば 10 個の独立な乱数を加えることによって高い精度で近似できる．このようにして初期化した単語 w_i の行ベクトルを $\boldsymbol{wc}_i^{Rand}$ と表わすことにする．

   このように乱数によって初期化した行ベクトル間は擬似的に直行しているとみなせる．すなわち，任意の行ベクトル，$\boldsymbol{wc}_i^{Rand}$，$\boldsymbol{wc}_j^{Rand}$ について，その内積値は 0 に近い値となる．

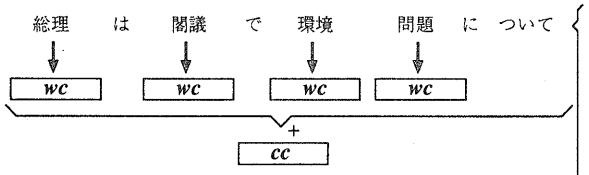
3. 学習後の行ベクトル $\boldsymbol{wc}_i$ は次の計算でえられる．

$$\boldsymbol{wc}_i = \boldsymbol{wc}_i^{Rand} + \eta \sum_j \alpha_{ij} \boldsymbol{wc}_j^{Rand} \qquad (3)$$

$$\boldsymbol{wc}_i = \frac{\boldsymbol{wc}_i}{||\boldsymbol{wc}_i||} \qquad (4)$$

   単語・文書出現行列において行ベクトルが似ていれば，それだけ α_{ij} の値が大きい．(3) 式では，α_{ij} に応じて他の行ベクトルを足し込むことにより，単語・文書出現行列で似ている行と行が圧縮後も似たものとなるような学習を行っている．η はパラメータである．(4) 式はベクトルの大きさを 1 に正規化するものである．

このようにして求めた単語 w_i に対する行ベクトル $\boldsymbol{wc}_i$ を単語共起ベクトルとよぶことにする．



w	$(cc \cdot wc)^2$	Pc
会議	0.237962	0.002702
選挙	0.150773	0.001712
予算	0.128907	0.001463
大臣	0.018549	0.000211
野球	0.004556	0.000052
怪人	0.000002	0.000000
数学	0.000001	0.000000
....	....	....
計	88.079230	1.000000

図 3: 文脈共起確率の計算例

3 文脈共起ベクトル

次に，単語共起ベクトルをもとにして文脈を表現することを考える．本研究では，これを文章中の自立語の単語共起ベクトルの和として表現することにする．ただし，ベクトルの大きさは1に正規化する．これを文脈共起ベクトルとよぶことにする．

もともと各単語の単語共起ベクトルは，共起しやすい単語同士の内積値が大きい，という性質を持つものであった．文章中の自立語の単語共起ベクトルの和である文脈共起ベクトルは，文脈中の各自立語と共起しやすい，すなわちその文脈と共起しやすい語の単語共起ベクトルに対して内積値が大きいという性質を持つはずである．

このような文脈共起ベクトルを考えると，ある文脈での語の出現を予測して数量化することができ，次節で述べるように確率的言語モデルとして利用することができる．

4 文脈共起ベクトルを利用した言語モデル

4.1 文脈共起確率

文脈共起ベクトルと単語共起ベクトルの内積値は，文脈における単語の出現のしやすさを表す．これを確率の形にしたものを文脈共起確率とよぶことにする．文脈共起確率は自立語だけを対象とし，次の式で与える．

$$Pc(w_i|w_1^{i-1}C_c) = \frac{f(\boldsymbol{cc}_1^{i-1} \cdot \boldsymbol{wc}_i)}{\sum_{w_j \in C_c} f(\boldsymbol{cc}_1^{i-1} \cdot \boldsymbol{wc}_j)} \tag{5}$$

ここで，$\boldsymbol{cc}_1^{i-1}$ はそれまでの文脈 (w_1^{i-1}) についての文脈共起ベクトル，C_c は自立語クラスを示すものとする．$Pc(w_i|w_1^{i-1}C_c)$ は，それまでの文脈と，次が自立語であるということを条件としたときに w_i が出現する条件付き確率であり，文脈共起ベクトルと単語共起ベクトルの内積値を(考慮している) すべての自立語について正規化したものである．

一般に，ある文脈に対して非常に共起しやすい語とそうでない語の内積値の差はそれほど大きな値にはならない．$f(x)$ はそれをより明確化するための関数である．後で示すように，x^2，x^3 のような関数を用いることでテストセットパープレキシティがより減少することがわかった．

図3に文脈共起確率の計算例を示す．

4.2 文脈共起確率を考慮した言語モデル

文脈共起確率は長距離の依存関係，トライグラムは局所的な関係を表しているものであり，これらは相補的関係にあると考えられる．そこでこれらを線形補間した言語モデルを考えることにする．なお，以下でトライグラムといっているのは，実際にはユニグラム，バイグラム，トライグラムの線形補間である．

文脈共起確率を考慮した言語モデルの計算方法を図4に示す．このモデルを以下では文脈言語モ

$$P(w_i|w_1^{i-1}) = \begin{cases} P(C_c|w_1^{i-1}) \times P(w_i|w_1^{i-1}C_c) & \text{if } w_i \in C_c \\ P(C_f|w_1^{i-1}) \times P(w_i|w_1^{i-1}C_f) & \text{if } w_i \in C_f \end{cases}$$

where

$$\begin{aligned} P(C_c|w_1^{i-1}) &= \lambda_1 P(C_c) + \lambda_2 P(C_c|w_{i-1}) + \lambda_3 P(C_c|w_{i-2}w_{i-1}) \\ P(w_i|w_1^{i-1}C_c) &= \lambda_{c1} P(w_i|C_c) + \lambda_{c2} P(w_i|w_{i-1}C_c) + \lambda_{c3} P(w_i|w_{i-2}w_{i-1}C_c) \\ &\quad + \lambda_{cc} Pc(w_i|w_1^{i-1}C_c) \\ P(C_f|w_1^{i-1}) &= 1 - P(C_c|w_1^{i-1}) \\ P(w_i|w_1^{i-1}C_f) &= \lambda_{f1} P(w_i|C_f) + \lambda_{f2} P(w_i|w_{i-1}C_f) + \lambda_{f3} P(w_i|w_{i-2}w_{i-1}C_f) \end{aligned}$$

with

$$\lambda_1 + \lambda_2 + \lambda_3 = 1,\ \lambda_{c1} + \lambda_{c2} + \lambda_{c3} + \lambda_{cc} = 1,\ \lambda_{f1} + \lambda_{f2} + \lambda_{f3} = 1.$$

図 4: 文脈言語モデル

デルとよぶことにする．文脈共起確率は自立語についてのみ考えるので，確率は自立語と付属語で別々に計算する (自立語クラスを C_c，付属語クラスを C_f と表わす)．

$P(C_c|w_1^{i-1})$ は w_1^{i-1} に自立語が続く確率で，これはトライグラムで近似する．自立語 w_i の出現確率は，この確率に，次が自立語であるという条件のもとでの w_i の確率，$P(w_i|w_1^{i-1}C_c)$ をかけたものとなる．そして，$P(w_i|w_1^{i-1}C_c)$ はトライグラムと文脈共起確率を線形補間したものとする．

一方，付属語の場合は，次が付属語である確率，$P(C_f|w_1^{i-1})$ に，次が付属語である条件のもとでの w_i の確率，$P(w_i|w_1^{i-1}C_f)$ をかけたものとなる．$P(w_i|w_1^{i-1}C_f)$ は文脈共起確率を考慮しないので，単にトライグラムとなる．なお，これらの式の λ の値は実験において EM アルゴリズムによって求める．

5 実験

5.1 テストセットパープレキシティ

毎日新聞の 1991 年から 1997 年の 7 年間の記事コーパス (約 44 万記事) を用いて，単語共起ベクトルを計算し，文脈言語モデルとトライグラムモデルとのテストセットパープレキシティの比較を行った．

実験では，単語共起ベクトル，ユニグラム，バイグラム，トライグラムの学習に 6 年分，ヘルドアウト用 (EM 法による λ の計算) に半年分，テスト用 (パープレキシティの計算) に半年分を用いた．

単語共起ベクトルはコーパスの自立語のうち高頻度の上位 5 万語を対象とした．ただし，時相名詞，形式名詞，副詞的名詞，数詞，「する」「なる」「いう」「ある」「ない」は対象外とした (これらは図 4の計算式では付属語クラス C_f に属する語と考える)．また，ユニグラム，バイグラム，トライグラムは全単語のうち高頻度の上位 6 万語を対象とした．

テストセットパープレキシティの比較結果を表 1に示す．表の左は全体のテストセットパープレキシティ，右は単語共起ベクトルを考慮した 5 万語 (ターゲット語) のみのテストセットパープレキシティを示したものである．表に示すとおり，最適なパラメータの設定 (* の行) において，全体で 5.0%，ターゲット語に対して 27.2% パープレキシティが減少しており，文脈共起確率の効果が十分あることがわかる．

なお，最適なパラメータ設定時の λ の値は以下のとおりであった．

$$\lambda_1 = 0.08, \lambda_2 = 0.50, \lambda_3 = 0.42$$

$$\lambda_{c1} = 0.03, \lambda_{c2} = 0.50, \lambda_{c3} = 0.30, \lambda_{cc} = 0.17$$

$$\lambda_{f1} = 0.06, \lambda_{f2} = 0.57, \lambda_{f3} = 0.37$$

図 5にテスト記事における文脈言語モデルとト

表 1: 文脈言語モデルとトライグラムモデルのパープレキシティの比較

言語モデル				全体の パープレキシティ	ターゲット語の パープレキシティ
トライグラムモデル				107.7	1930.2
	次元	η	$f(x)$		
文	500	0.5	x^2	106.3 (-1.3%)	1663.8 (-13.8%)
脈	1000	0.3	x^2	102.7 (-4.7%)	1495.9 (-22.5%)
言	1000	0.5	x	103.6 (-3.9%)	1496.1 (-22.5%)
語	* 1000	0.5	x^2	102.4 (-5.0%)	1406.2 (-27.2%)
モ	1000	0.5	x^3	102.4 (-5.0%)	1416.8 (-26.9%)
デ	1000	1.0	x^2	102.5 (-4.8%)	1430.3 (-25.9%)
ル	2000	0.5	x^2	102.4 (-5.0%)	1408.1 (-27.1%)

ライグラムモデルの比較例を示す．単語共起ベクトルを考慮した各語について，文脈言語モデルの方が出現確率が大きい場合 (うまく予測している場合) は四角で，トライグラムモデルの方が出現確率が大きい場合は下線でマークしている．

この図から，付属語の直後では，トライグラムモデルによる予測は難しく，文脈言語モデルの方が出現確率が大きいという場合が多いことがわかる．逆に，複合名詞内の2語目のようにトライグラムモデルの予測が強い場合には，文脈共起確率を考慮する分 (文脈共起確率では大きな差はつかないので) 出現確率が小さくなっている．

本稿で述べた言語モデルでは，トライグラムモデルと言語モデルを一定の割合で線形補間しているが，トライグラムによる予測の分布に応じて補間の割合を変えることでさらにテストセットパープレキシティが減少する可能性がある．

6 関連研究

長距離依存の問題を扱う古典的な手法としては，キャッシュモデルがある[6]．キャッシュモデルは文脈中にすでに出現している単語の確率を大きく見積もるというものである．

トリガモデル[7] は，より一般的に単語の共起度を利用するが，文脈中の個々の単語との個別の共起度を考えるもので，我々のモデルのように文脈の表現というものはない．個別の共起度を考慮

米 株式 市場 の 急騰 を 背景 に ウォール街 が 活況 を 呈し，わが世の春を おう歌 している。 証券 会社，投資 銀行 は１９９６年に 入って から 過去 最高 益 を 更新 。９６年の 株式 相場 は９５年に 続き 急伸。また 株価 急伸 を 背景 に 企業 の 新株 発行 が 過去 最高 となった。

図 5: 文脈言語モデルとトライグラムモデルの比較例

することは意味的曖昧性解消などでも一般に行なわれていることである[4, 8]．

共起情報をもとにして単語ベクトル，文脈ベクトルを作り，その内積をもとに文脈共起確率を計算するという点で，我々の研究は Coccaro らのモデルに近い[1]．Coccaro らのモデルでは，バイグラムとの比較でパープレキシティが 12% 減少すると報告されている．

主な相違点は，1) 単語・文書行列の圧縮に Latent Semantic Analysis[2] の SVD(Singular Value Decomposition) を用いている点，2) すべての語について文脈共起確率を考え，各語について，文書集合の中での出現のかたよりを尺度として文脈共起確率と N グラムモデルを混合する点である．

1) について，我々は比較的軽い繰り返し計算

によって圧縮を行っている．一つの理由は，膨大な文書から共起情報を得ようとする時にSVDの計算量が非常に大きいことである．また，共起と意味の関係を認知的興味として考えたとき，SVDのような数学的モデル化よりも，繰り返し計算によるモデル化の方が妥当性が高いように思えるという理由もある．2)については，文書集合中での出現のかたよりを直接尺度とする必然性はないと考えた．必要なことは意味的な共起を考えるべき語とそうでない語を区別することであり，これは自立語と付属語を区別することで十分であると判断した．

文脈の表現をトピックの予測という形で行い，トピックごとに単語の出現分布が異なるということを利用する研究も行われている．例えば，Florianらの次のような言語モデルがある(3)．

$$P(w_i|w_1^{i-1}) = \sum_t P(t|w_1^i) \cdot P_t(w_i|w_{i-N+1}^{i-1}) \quad (6)$$

tはトピックを表わし，トピックは対象コーパスをクラスタリングして得る(実験では新聞記事コーパスに対して30トピック程度)．まず，$P(t|w_1^i)$で文脈からトピックを予測し，これにトピックごとのトライグラム$P_t(w_i|w_{i-N+1}^{i-1})$をかけたものが単語の出現確率となる．パープレキシティとしては，バイグラムとの比較で全体で10.5%，ターゲット語で33.5%減少している．

7 おわりに

単語の共起情報に基づいて文脈を表現することにより，長距離依存の関係を考慮する言語モデルを提案した．トライグラムと比較して，テストセットパープレキシティが全体で5.0%，自立語で27.2%減少した．また，非常にシンプルな方法であり，計算コストも低いという特徴がある．

人間の文脈理解がより構造的であり，複雑であることはまちがいない．そのような方向へ本モデルを発展させることは今後の課題である．

参考文献

(1) Noah Coccaro and Daniel Jurafsky, Towards Better Integration of Semantic Predictors in Statistical Language Modeling. *Proceedings of ICSLP-98*, Vol. 6, pp. 2403–2406, 1998.

(2) Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), pp 391–407, 1990.

(3) Radu Florian and David Yarowsky, Dynamic nonlocal language modeling via hierarchical topic-based adaptation, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 167–174, 1999.

(4) Pascale FUNG, LIU Xiaohu, and CHEUNG Chi Shun, Mixed language query disambiguation, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 333–340, 1999.

(5) Mard R. Ilgen and David A. Rushall, Recent advances in HNC's context vector information retrieval technology, *TIPSTER PROGRAM PHASE II*, pp. 149–158, 1996.

(6) R. Kuhn and R.de Mori, A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12-6, pp. 570–583, 1990.

(7) R. Lau, Ronald Rosenfeld and Salim Roukos, Trigger based language models: a maximum entropy approach. In Proceedings ICASSP, pp. 45–48, 1993.

(8) D. Yarowsky. Decision lists for lexical ambiguity resolution : Application to accent restoration in Spanish and French. In *Proc. of the 32nd Annual Meeting of ACL*, pp. 88–95, 1994.