LE-5

辞書に基づく単語の確率ベクトル
Probabilistic Word Vector based on Dictionaries

鈴木 敏[1]
Satoshi Suzuki

1 まえがき

一般に，テキスト処理ではTF·IDF(Term Frequency · Inverse Document Frequency)を基にして処理を行うことが多い．この手法は実に様々なアプリケーションに用いられ，その有効性が検証されている．また，IDFのエントロピーに基づく有効性の証明[1]も示されている．しかしながら，TF·IDFから得られたベクトルの示す意味は必ずしも明確ではなく，多くの場合，単なる数値とみなして計算を行っている．このとき，例えばベクトルの正則化の可否など，その応用に際しては十分な注意が必要であるが，検証が不十分な例も多い．さらに，このベクトル表現は極めてスパースなものであり，様々な学習手法への適用に際し制約にもなっている．

これらの問題を解決するために，TF·IDFに代わる手法として，確率に基づいた単語のベクトル化手法を提案する．本手法によれば，辞書の持つ単語（インデックス）と説明文（インデックスの集合）の関係から再帰的に単語を展開することにより，単語間の間接的な関係を含めたベクトル化が可能である．また，確率に基づき導出されているため，その意味するところは極めて明確であり，応用の見通しが良くなるという長所がある．

2 単語の確率ベクトル表記

2.1 拡張説明文

提案手法は辞書を基にしている．辞書は，見出し語と説明文の組で表現されている．説明文もまた見出し語の集合であると考えると，見出し語に対する説明文は無限に展開できる．以下，n回展開された説明文をn次説明文と呼ぶ．見出し語w_iに対して，元の説明文（1次説明文）もn次説明文も全てw_iを説明する文である．これらの中から説明文を1つ選び出すとして，n次説明文が選ばれる確率をP_nで表す．

一方，w_iのn次説明文中に単語w_jが現れる確率を$P(w_j^{(n)}|w_i)$とすると，元の辞書での単語間の関係は

$$A=\begin{bmatrix} P\left(w_1^{(1)}|w_1\right) & P\left(w_1^{(1)}|w_2\right) & \cdots & P\left(w_1^{(1)}|w_m\right) \\ P\left(w_2^{(1)}|w_1\right) & \ddots & & \\ \vdots & & \ddots & \\ P\left(w_m^{(1)}|w_1\right) & & & P\left(w_m^{(1)}|w_m\right) \end{bmatrix}$$

として表される．ただし，$N(w_j^{(1)}|w_i)$を説明文中の単語の出現頻度とすると，

$$P(w_j^{(1)}|w_i)=\frac{N(w_j^{(1)}|w_i)}{\sum_{all\ k} N(w_k^{(1)}|w_i)} \tag{1}$$

である．2次説明文に関しては，

$$P(w_j^{(2)}|w_i)=\sum_{all\ k} P(w_j^{(2)}|w_k^{(1)})P(w_k^{(1)}|w_i) \tag{2}$$

が成立ち，行列Aを用いれば2次説明文全体を表す式はA^2となる．同様に，n次説明文に関しては，

$$P(w_j^{(n)}|w_i)=\sum_{all\ k_{n-1}}\sum_{all\ k_{n-2}}\cdots\sum_{all\ k_1} P(w_j^{(n)}|w_{k_{n-1}}^{(n-1)})P(w_{k_{n-1}}^{(n-1)}|w_{k_{n-2}}^{(n-2)})\cdots P(w_{k_1}^{(1)}|w_i) \tag{3}$$

が成立ち，全体を表現する式はA^nとなる．

ここで，1次説明文から∞次説明文までの全てをまとめた拡張説明文

$$C=P_1A+P_2A^2+\cdots+P_nA^n+\cdots \tag{4}$$

を考える．一般に，この式を計算することは不可能である．しかし，P_nがnに比例して一定の割合aで減少すると仮定すると，

$$C = b(aA+a^2A^2+\cdots+a^nA^n+\cdots), \tag{5}$$

$$b=1/\sum_{k=1}^{\infty}a^k, \tag{6}$$

$$0<a<1 \tag{7}$$

となり，これらの式から

$$(I-aA)C=abA \tag{8}$$

を得る．一般に，CはC,Aの第i列ベクトルC_i,A_iを用いた連立方程式

$$(I-aA)C_i=abA_i \tag{9}$$

の解の集合として求めることが出来る．特に，$det(I-aA)\neq 0$ならば，

$$C=abA(I-aA)^{-1} \tag{10}$$

により，行列Cを直接求めることも可能である．Cは単語と拡張説明文の関係を表した確率ベクトルの集合でもある．

2.2 単語の類似度

上記の拡張説明文Cを利用して，単語間の類似度を計算する手法を検討する．見出し語w_jが与えられたときの拡張説明文中の単語w_k^*の確率が$P(w_k^*|w_j)$であるとすると，w_k^*が与えられたときに，見出し語w_iが想起される確率は

$$P(w_i|w_k^*)=\frac{P(w_k^*|w_i)P(w_i)}{\sum_{all\ l}P(w_k^*|w_l)P(w_l)} \tag{11}$$

となる．従って，見出し語w_jから見出し語w_iを想起する確率は，

$$P(w_i|w_j) = \sum_{all\ k}P(w_i|w_k^*)P(w_k^*|w_j) = \sum_{all\ k}\frac{P(w_k^*|w_i)P(w_i)P(w_k^*|w_j)}{\sum_{all\ l}P(w_k^*|w_l)P(w_l)} \tag{12}$$

により与えられることになる．

この値を(i,j)成分とする行列Sを考えると，Sもまた単語の確率ベクトルの集合である．

[1]NTTコミュニケーション科学基礎研究所, NTT CS Labs.

3 計算機実験

3.1 国語辞典からのベクトル抽出

上記の手法を実際に国語辞典 [2] に適用した結果を以下に示す．前処理として，扱う単語を一般名詞とサ変名詞に限定 (形態素解析は茶筅 [3] を利用) し，その結果，44050 語の見出し語と，平均約 7 語の 1 次説明文を得た．

まず，式 (1) を用いて確率行列 A が計算される．これはスパースな 44050 次元の正方行列である．次に式 (9) から線形学習法により C を求めた．このときのパラメタは $a = 0.9$ とし，有効桁は 10^{-6} までとした．学習の結果，十分な収束を得られなかった語を除いて，43616 語の確率ベクトルを得た．非ゼロの値を持つ次元数は平均で約 24778 であった．

さらに，この結果を式 (12) に適用し，単語の類似度を求めた．見出し語の事前確率 $P(w_i)$ は，辞書の中では全ての見出し語の出現確率が等しいため，一定とした．この結果得られた行列 S において，単語確率ベクトルの非ゼロ値の次元数は平均で 43326 であった．結果，行列 S は 99%以上の要素が非ゼロの値を持ち，スパース性は解消されている．

3.2 単語類似度の検証

単語類似度の具体例を表 1に示す．見出し語に対し，その拡張説明文から想起されやすい上位 10 語が示されている．類似度は，式 (12) により求められた確率の値である．最も強く想起される単語が必ずしもその見出し語ではないという特徴を持っている．

これらの結果を検証するために，心理実験結果をもとに従来の計算手法との比較を行った．比較対象とするのは，辞書から得た単語頻度の正方行列 K を基に，$G = \alpha K + \beta K^2 + \gamma K^T$ を用いて単語ベクトルを計算し，それらの余弦を単語類似度とみなす手法 [4] である．また，心理実験は類義語，連想語を書き出すアンケート調査 [5] である．

図 1は類義語，連想語に関して従来手法と比較した結果である．計算により得られた語順のリストの中から心理実験の結果得られた類義語，連想語を見つけ出し，特定順位以内に現れた類義語，連想語の数をカウントし，グラフ化した．類義語に関してはほぼ同等の結果であり，連想語では僅かながら提案手法が優位であることが示されている．ただし，従来手法では単語数が約 88000 語と，提案手法の計算例の倍であるため単純には比較できない．しかしながら，差分の単語は品詞が異なるため，順位への影響はかなり小さいと考えられる．

提案手法の計算例では，単語の拡張説明文から単語が「想起」される確率を計算しており，計算結果が連想語にも対応した表現となっていることは，計算の定義から明らかである．このように，明確な意味付けが得られることが，提案手法のメリットの 1 つである．

レベル		ボディーガード	
類義語	類似度	類義語	類似度
水準儀	0.005534	ボディーガード	0.025666
レベル	0.005037	ガード	0.004786
レベルアップ	0.003480	身辺	0.00305
平準	0.001155	警衛	0.002760
水準	0.000914	親衛	0.002464
精度	0.000791	用心棒	0.002308
儀	0.000754	エスコート	0.002254
準	0.000684	座右の銘	0.001947
准	0.000684	警護	0.001810
別儀	0.000617	護衛	0.001569

表 1: 単語類似度

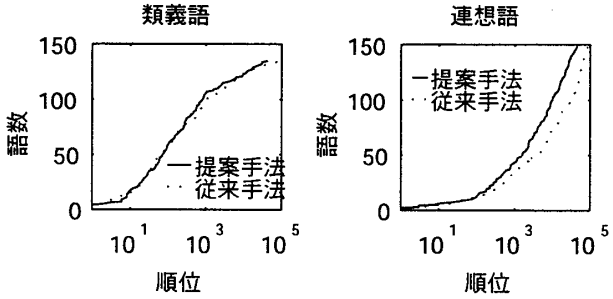

図 1: 従来手法との比較

4 考察

文頭でも述べたように，TF·IDF が有効な手法であることは，エントロピーの観点から証明されている．しかしながら，意味的な面で理解しやすいとは言い難い．

一方，提案手法は，拡張説明文に関する仮定の下で，確率的に導出されている．その有効性は，事前確率 $P(w_i)$ を一定とみなせば式 (12) が

$$P(w_i|w_j) = \sum_{all\ k} \frac{P(w_k^*|w_i)P(w_k^*|w_j)}{\sum_{all\ l} P(w_k^*|w_l)} \qquad (13)$$

となり，TF^2·IDF に類似した形式で表されることからも類推できる．すなわち，$P(w_k^*|w_i)$ は TF そのものであり，$\sum_{all\ l} P(w_k^*|w_l)$ は意味的に DF に近い．従って，上述の単語類似度の計算において，最も強く想起される単語が必ずしもその単語自身でないことの理由は，TF·IDF と同様に，より特徴的な単語を選ぶ傾向を持つためである．また，式 (13) から行列 S は対象性を持ち，$P(w_i|w_j) = P(w_j|w_i)$ であることもわかる．

5 あとがき

確率手法に基づき TF·IDF に代わり得る単語のベクトル化手法を提案し，単語類似度を用いてその有効性を示した．2 種類の確率ベクトルを提案したが，それぞれ異なる意味を持つため，その応用もそれぞれに考えられる．応用するに当たっては，単純な数値としての利用も出来るが，確率の計算手法に従うのが望ましい．

本手法は辞書を前提としており，拡張説明文の仮定の下に成り立っている．従って，言語処理全般にわたって TF·IDF を置き換えられるわけではない．コーパスからの計算手法の検討は今後の課題である．

参考文献

[1] Kishore Papineni, "Why Inverse Document Frequency?," NAACL, Pittsburg, 2001.

[2] 金田一, 池田, "学研 国語大辞典 第二版," 学習研究社, 1988.

[3] 松本, 北内, 山下, 平野, 松田, 高岡, 浅原, "日本語形態素解析システム『茶筌』," 2000.

[4] 笠原, 松澤, 石川, "国語辞書を利用した日常語の類似性判別," 情報処理学会論文誌 vol.38, No.7, pp.1272-1283, 1997.

[5] 笠原, 稲子, 金杉, 永森, 加藤, "単語の関連性判別の分析," 第 9 回ことば工学研究会, 大阪, 2001.