共起単語中の特異的頻出単語数を用いた用語の representativeness 計測尺度

3H−4

久光 徹　丹羽 芳樹
日立製作所 中央研究所

1．はじめに

大規模な文書集合の内容を要約することを目的として，文書集合中の「話題を代表する傾向の強い(=representative な)用語」を選ぶための尺度を定義する Baseline 法を開発してきた[1]。本報告では，「用語 T を含む文書集合 $D(T)$中で特異的に頻出する単語数」という特徴量を Baseline 法の枠組みで利用する新尺度を提案し，これまでに提案した尺度に対する，単語選別能力等の優位性を示す。

2．これまで得られた知見のまとめ

情報検索の分野では「用語の重要度」に関する尺度が索引語の重み付けのために導入されてきた[2]。しかし，従来の尺度には，語の頻度の寄与が大きすぎたり，閾値の設定が困難である等の問題があった。我々は「正規化」によりこのような問題の無い尺度を構成するための基本方式 Baseline 法を開発した[1]。

2.1 Baseline 法の概要

用語が何らかの話題を代表する力の強さを representativeness と呼び，これを定義する基本的な考え方として，「用語 T が特徴的ならば，T を含む文書の集合 $D(T)$は，“平均的な”文書集合に比べて何らかの特徴を持つ」という仮説を設定する。この仮説を具体的手続きに還元するため，単語集合から実数への写像を一つ固定し，「用語 T が特徴的ならば，$M(D(T))$は，“平均的な値”から外れる」と換言する。こうして，M と“平均的な値”を適切に定義すれば T の representativeness 尺度が定義できる。

ここで，$M(D(T))$は，一般に $D(T)$の大きさ(＝含まれる単語数)のみに依存して，その値が系統的に(多くは単調増大または減少)変動する。従って，“平均的な値”とは，この変動部分，すなわち，$D(T)$と同等の大きさのランダムサンプリングされた文書集合に対する $M(\bullet)$の値と考えることができる。そこで，文書単位をランダムサンプリングして生成した文書集合 D_{rand}に対して，D_{rand}が含む単語数$\#D_{rand}$を用いて $M(D_{rand})$を推定する関数を $B_M(\bullet)$(ベースライン関数と呼ぶ[1])とし，$B_M(\#D(T))$と $M(D(T))$の値の比をとるなどして $M(D(T))$を補正するのが，Baseline 法の原理である。この補正を，M に対する正規化と呼び，正規化された M を用いて定義した用語 T の representativeness 尺度を，M から Baseline 法により生成された尺度と呼び $B(\bullet, M)$と書き，M を $B(\bullet, M)$の原尺度と呼ぶ。Baseline 法により生成された representativeness 尺度は，異なる頻度を持つ用語 T の間で特徴量を比較することが意味を持ち，representativeness の有無を決める閾値が合理的に設定できる等の特長を持つ。

2.2 これまでに構成した尺度とその問題

原尺度 M としては，これまでに，

- 用語 T を含む全ての文書集合 $D(T)$中の単語分布 $P_{D(T)}$と，全文書集合中の単語分布 P_0の間の距離 $Dist(P_{D(\bullet)}, P_0)$（略称 *Dist*）
- $D(T)$における単語の異なり数を与える $DIFFNUM(D(\bullet))$（略称 *DIFFNUM*）
- $D(T)$における単語分布のエントロピー：$ENT(D(\bullet))$（略称 *ENT*）

等を考察した。中でも，*Dist* から Baseline 法により生成した尺度 $B(\bullet, Dist)$は，既存の諸尺度や $B(\bullet, DIFFNUM)$等と比較して，特徴単語の選別能力が大きく優れ，用語抽出実験においても有用であった。

しかし，$B(\bullet, Dist)$を新聞記事に適用した特徴単語抽出実験では，T を含む文書集合中で，不用語以外の比較的少数種類の語が大きな頻度を持つ場合，その偏りを大きく評価しすぎるという問題点が観測された。これを解決するためには，ある用語と共起する語の中で，特異的に高頻度で現れる語の「豊かさ」を併せて評価できるほうが好ましい。そうすれば，「内容豊かな話題を多く持つ」$D(T)$を与える用語 T の評価値が上がるからである。

3．共起単語中の特異的頻出単語数を用いた用語の representativeness 計測尺度

上記の考え方に基くもっとも単純な方法は，$D(T)$中の「特異的に高頻度で現れる語」の異なり数を数えることである。すなわち，原尺度として，「用語 T を含む文書集合 $D(T)$中で，特異的に高頻度で現れる語の数」を用いる。「語が特異的に高頻度で現れる」ことを計量するため，我々は[3]で報告した単語の重み付け尺度 HGS を用い，$D(T)$に対する原尺度 N_pの値 $N_p(D(T))$を，「$D(T)$内の HGS 尺度が p より大である†単語の異なり数」と定義し($p>0$ はパラメータ)，N_pを Baseline 法で正規化した $B(\bullet, N_p)$を新たな representative 尺度として，提案する。

4．比較実験

$B(\bullet, N_p)$に関しては，p を 10 から 200 まで 10 刻みに増やし，$B(\bullet, Dist)$，$B(\bullet, DIFFNUM)$，$B(\bullet, ENT)$，*tf-idf*，*tf*(単純頻度)の五尺度との比較実験を行った。

4.1 実験の詳細

(1)日経新聞 1996 年分に 3 回以上出現した単語から 20,000 語を無作為抽出し，そのなかから無作為抽出した 2,000 単語を P，N，U の三クラスに人手で分類する。ここで，分類の指針は以下のとおり：

・分類 P：「w が特定の話題に関する記事を検索する際の，良い手がかりになる」OR「複数の記事に共通する特定の話題そのものを表す言葉である」OR「情報検索の結果得られる記事集合の内容を要約する部分単語集合を抽出するとして，そのメンバとして適切である」

・分類 N：分類 P の各判定がすべて否定的である。

・分類 U：P とも N とも決定できない。

(2)前記 20,000 個の単語をランダムにソートし，先頭から k 位までの，クラス P の単語の累積出現個数を $Rand(\mathrm{P}, k)$，クラス N の単語の，先頭から k 位までの累積出現個数を $Rand(\mathrm{N}, k)$，尺度 M を用いて前記 20,000 個の単語をソートした時の同様の値を $M(\mathrm{P}, k)$，$M(\mathrm{N}, k)$とする。尺度 M に対して，$DP(M, k)=M(\mathrm{P}, k)-Rand(\mathrm{P}, k)$，$DN(M, k)=Rand(\mathrm{N}, k)-M(\mathrm{N}, k)$とする。$DP(M, k)$，$DN(M, k)$の値を，それぞれ *DP* スコア，*DN* スコアと呼ぶ。更に，

---

* A Measure of Term Representativeness Based on the Number of Co-occuring Salient Words
Toru Hisamitsu and Yoshiki Niwa
Central Research Laboratory, Hitachi, Ltd

† v が全文書中と $D(T)$中に出現する確率が同一で，単語の出現確率が互いに独立であると仮定した場合，v が D に実際に出現している数観測される確率が 2^{-p} より小さい。

$$ADP(M,k)=\sum_{l=1}^{k}DP(M,l),$$

$$ADN(M,k)=\sum_{l=1}^{k}DN(M,l).$$

により $ADP(M, k)$, $ADN(M, k)$を定義し，それぞれ *ADP* スコア，*ADN* スコアと呼ぶ。以上の各スコアは高いほど望ましい。

4.2 実験結果

図1は，M∈{$B(\bullet, Dist)$, *tf-idf*, $B(\bullet, N_p)$}(p は 4 種類)について $0 \leqq k \leqq 20{,}000$ の範囲で $DP(M, k)$をグラフ化したものである。図 2 は，M∈{$B(\bullet, Dist)$, $B(\bullet, DIFFNUM)$, $B(\bullet, ENT)$, *tf-idf*, *tf*, $B(\bullet, N_p)$}について*ADP*(M, *5,000*), *ADP*(M, *10,000*), *ADP*(M, *20,000*)を比較したものである(p は 6 種類)。

$B(\bullet, N_p)$が P と分類される語の優先順位を上げる力は，$B(\bullet, Dist)$を除く全尺度に対し，$20 \leqq p \leqq 200$, $0 \leqq k \leqq 20{,}000$ について，*DP* スコアで上回り，$k=5{,}000$, 10,000, 20,000 について，*ADP* スコアで上回った。

$B(\bullet, Dist)$に対しては，$20 \leqq p \leqq 200$, $0 \leqq k \leqq 15{,}000$ 程度の範囲で *DP* スコアで上回り，$k=5{,}000$, 10,000, 20,000 について，*ADP* スコアで上回った。このことは，$B(\bullet, N_p)$が P と分類される語の優先順位を上げる力は，$B(\bullet, Dist)$に対して全般的に優越するだけでなく，先頭順位付近に P と分類される語を集める力において特に優越することを示している。

$B(\bullet, N_p)$が N と分類される語の優先順位を下げる力も，*DN* スコア，*ADN* スコアを用いて同様のことが分った。

4.3 Baseline 関数による補正が不要となる可能性

様々な p に対して Baseline 曲線 $B_p(\bullet)$を求めると，p のある値 p_0 を境として，単調増大から単調減少に変わり，$B_{p0}(\bullet)$は x 軸に平行になる(4.1 で述べたコーパスでは p_0=140 程度)。このことは，N_{p0} は Baseline 法による正規化無しでも representativeness 原尺度として利用できることを示す。実際，N_{140} と $B(N_{140}, \bullet)$を比較したところ，双方の性質はきわめて良く類似していた。$B(N_{140}, \bullet)$は最良でないにせよかなり良い性能を示している。他の p についても，N_p と $B(N_p, \bullet)$の性能は類似しており，$B(Dist, \bullet)$の性能を概ね上回る。これは，原尺度 N_p 自体が，広範囲な p について良好な representativeness 尺度となることを示しており，N_p は扱い易く，有用な尺度である可能性を示唆している。

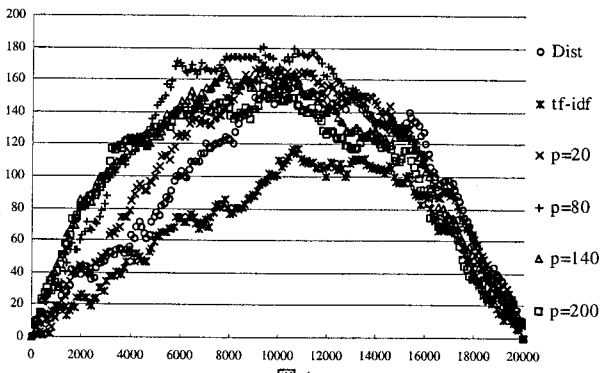

図 1

6 種類の尺度に関する $0 \leqq k \leqq 20{,}000$ での *DP* スコアの比較

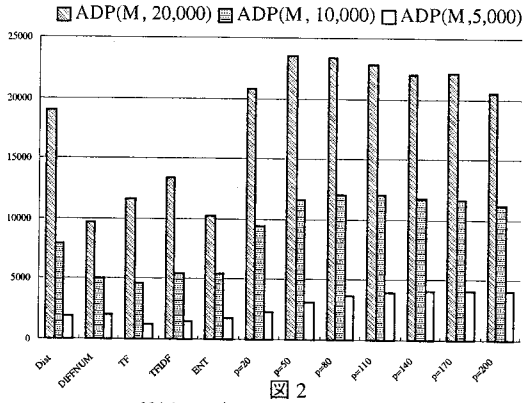

図 2

12 種類の尺度に関する *ADP* スコアの比較

5．まとめ

用語 T を含む文書集合 $D(T)$に対して「$D(T)$中で特異的に高頻度で現れる単語の異なり数 $N_p(D(T))$」を考え，N_p を Baseline 法で正規化した representativeness 尺度 $B(\bullet, N_p)$を提案した。「$D(T)$中で特異的に高頻度で現れる単語」は，パラメータ p を用いて，「$D(T)$内の HGS 尺度が p より大である(すなわち，v が全文書中と $D(T)$中に出現する確率が同一で，単語の出現が互いに独立であると仮定した場合，v が $D(T)$に実際に出現している回数だけ出現する確率が 2^{-p} より小さい)単語と定義する。

$B(\bullet, N_p)$について，「情報検索の立場から有用な単語を選択するタスク」における能力を調べた結果，広範囲の $p(20 \leqq p \leqq 200)$について，$B(\bullet, N_p)$は，[1]で提案した尺度 $B(\bullet, Dist)$より優れていることが分った。

更に，日経新聞 1996 年 1 年分においては，p=140 の付近で N_pの Baseline 関数は定数となり，Baseline 法による正規化無しで，N_p は $B(\bullet, N_p)$と極めて類似した性能を示すことも分った。N_p と $B(\bullet, N_p)$の性能は，広範囲の p において類似しており，両者とも $B(\bullet, Dist)$より優れている。異なるサイズ・種類のコーパスでこの安定性が確認できれば，$B(\bullet, N_p)$だけでなく原尺度 N_p も実用的かつ効果的といえる。今後，これらを実験的に検証する予定である。

謝辞

本研究の一部は IPA 独創的情報技術育成事業の支援を受けて行われました。本研究を進めるにあたり有益なコメントを頂いた，東京大学理学系研究科辻井潤一教授と，国立情報学研究所影浦峡助教授に感謝致します。

参考文献

[1] Hisamitsu, T., Niwa, Y., and Tsujii, J. 2000. A Method of Measuring Term Representativeness - Baseline Method Using Co-occurrence Distribution-, *Proceedings of COLING2000*, pp.320-326.

[2] Kageura, K. and Umino, B. (1998). Methods of automatic term recognition: A review. Terminology 3(2), pp.259-289.

[3] 久光徹, 丹羽芳樹. 2001. 組み合わせ確率を用いた特徴単語選択方法, 言語処理学会第 7 回年次大会論文集. pp.169-172.

[4] Niwa, Y., Iwayama, M., Hisamitsu, T., Nishioka, S., Takano, A., Sakurai, H., and Imaichi, O. (2000) *DualNAVI* -dual view interface bridges dual query types, *Proc. of RIAO 2000*, pp.19-20.