複合語のパープレキシティに基づく重要語抽出法の研究

森山 聡†　　吉田 稔‡　　中川　裕志‡
†東京大学大学院学際情報学府
‡東京大学情報基盤センター
E-mail:†{moriyama, mino, nakagawa}@r.dl.itc.u-tokyo.ac.jp

従来の重要語抽出の研究分野では正解との対応付けのとれた大規模なテキストを対象として実験を行い、精度や再現率の比較がなされてきた。しかし、頻度に着目する種々の重要語抽出手法では、抽出対象文書のサイズが小さい場合に、サンプル数の不足から高い精度を期待できない。そこで、小さい文書サイズでも高い精度が得られるよう複合語の構造に着目する手法について報告する。提案するのは、複合語を構成する単位である語基の左右に連接した語基の種類数と頻度からそれぞれのパープレキシティ値を求め、複合語の重みを計算する方法である。新聞 Web 記事を対象とし、文書集合サイズを 1 記事から 50 記事まで変化させて実験を行ったところ、従来の手法と比較して最も高い精度を安定して得ることができた。

1.はじめに

重要語抽出の研究分野では従来、様々な手法が提案されており、正解集合を持つ大規模なテスト用コーパスにこれらを適用した結果得られる精度や再現率について議論が為されてきた。しかし、1 つの Web ページのような、抽出対象の文書サイズが小さい場合にこれらの方法が有効であるかは未知数である。

そこで本研究では、複合語の構造に着目した。具体的には、既に提案されている LR[2][3]の連接頻度及び連接種類数という 2 つの概念を、複合語の構造に対して情報理論的尺度であるパープレキシティを用いて精密化する方法を提案する。実験には Web 新聞記事を用い、その結果を評価するため携帯端末向け記事に出現した語句を正解と定義した。

2.既存重要語抽出手法の分類

これまで研究がなされてきた重要語抽出の手法は、大まかに以下の 3 種類に分類される。

1　頻度の統計量に着目する手法
2　複合語の構造に着目する手法
3　以上 2 点を組み合わせた手法

本節では、今回我々が新しく提案する尺度と比較するために用いた既存の重要語抽出の手法について簡単に説明をする。

2.1 頻度に着目する手法

TF(Term Frequency), F(Frequency)

文書中に多数回出現した語ほど重要であるという前提を置いている。出現回数の多い一般語を重要とみなしてしまう欠点がある。

DF(Document Frequency)

ある語がいくつの文書に出現したかをカウントする尺度である。これは、多数の文書に出現する語は特定の文書の特徴を記述するのにふさわしくないという前提を置いている。

またそれら二つを組み合わせる TF・IDF[1]などがある。

2.2　複合語の構造に着目する手法

LR

複合語構造に着目する手法は、*「他の複合語の一部になりやすい語基がつくる複合語ほど、重要な語句である」*という前提を置いている。

2.3 以上の定義を組み合わせた手法

頻度情報及び、複合語の構成情報を様々な形で組み合わせるものとして、C-Value[1]（本稿では若干の変更をしているので MC-Value と呼ぶ)、FLR[3]などがある。また、二つ以上の尺度を最適に組み合わせるものとして内山らの手法[4]などがある。

3.パープレキシティを利用する重要度

LR 法は、複合語からなる集合において、ある語基の前あるいは後ろに連接して複合語を構成する語基の種類数（連接種類数）と延べ頻度（連接頻度）を語基の重要度としていた。これに対し、本研究では、情報理論的により精密な尺度としてパープレキシティを利用した尺度を提案する。

3.1 連接頻度・種類数ＬＲとパープレキシティ

例えば、図 3-1 のような状態を仮定する。

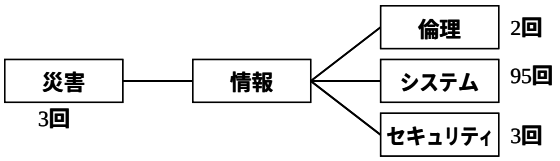

図 3-1：語基の接続例

このとき、語基「情報」の右側接続の重みはそれぞれ以下のように表せる。(LR は左右の連接頻度または連接種類数の相乗平均によって計算できる)
右側連接頻度（情報）＝101
左側連接種類数（情報）＝4

図 3-1 の場合に、連接頻度、種類数単体では頻度の偏り（「システム」が「倫理」・「セキュリティ」に比べて著しく多く現れている）を考慮していない。一方、下に述べる PerPlexity（情報理論的意味での分岐数）を用いた尺度では、95 回という極端に多い接続数が考慮され、右側 PerPlexity は 1.2163 と計算される。

3.2 パープレキシティ(PerPlexity)

言語を、長さnの単語列を生成する情報源と考えると、言語における単語列Wの生成確率を$P(w_1^n)$としたとき、言語全体のエントロピー値は

$$H_0(L) = -\sum_{w_1^n} P(w_1^n)\log P(w_1^n)$$

と表すことができる。ここから、１単語あたりの平均的なエントロピー値は、

$$H(L) = -\sum_{w_1^n}\frac{1}{n}P(w_1^n)\log P(w_1^n)$$

と書ける。H(L)は言語によって生成される単語を特定するために必要なビット数を意味している。また、*H(L)*に以下のような変形を与えたものはパープレキシティ（PerPlexity）と呼ばれており、これをPP と表して以下のように記述する。

$$PP = 2^{H(L)}$$

これは１単語あたりの平均連接種類数、すなわち単語の接続の複雑さを示す指標になっている。

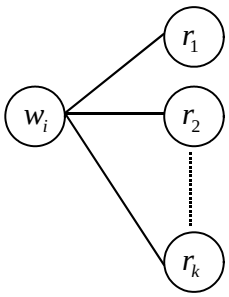

図 3-2：語基右側接続の一般例

図3-2のように複合語Ｗのｉ番目に当たる語基w_iの右側に k 種類の語基が接続している状態を考える。このとき、 w_iの右側に接続する語基を確率変数R^iと考え、R^iのとり得る事象の集合を$\{r_1,\cdots r_k\}$とするとき$H(R^i)$は、

$$H(R^i) = -\sum_k P(r_k)\log_2 P(r_k)$$

と表せる。複合語における語基の接続確率からエントロピーを求めた場合、その値は確率変数R^i自体の不確実性、すなわち、w_iの右側に接続する語基の情報理論的な分岐数を意味する。これは単純カウントを行う LR と比べて情報理論的により精密な分岐数と考えることができる。尚、ここで右側連接の場合について示したが左側についてもまったく同様である。

3.3 パープレキシティを用いた尺度

w_iの左右パープレキシティの組み合わせ値を$pp(w_i)$として、 $pp(w_i) = (pp_l(w_i)\cdot pp_r(w_i))^{\frac{1}{2}}$と定義する。さらに、n 個の語基からなる複合語 W 全体へこれを拡張し、PP(W)としたとき以下のように表せる。

$$PP(W) = \left[\prod_{i=1}^{n} pp(w_i) \right]^{\frac{1}{n}}$$

これによって複合語 W 全体の重みを計算することができる。実際は、計算の簡単のため、対数をとった値 *logPP(W)*をスコアリングに用いている。5 章で示す PerPlexity 尺度による実験結果はこの式を用いて計算された。

3.4 頻度情報との組み合わせ

前節までで定義した尺度と頻度情報との統合を試みる。後に述べるように組み合わせ対象として F と TF が考えられるが、TF がより高い精度を示すため、頻度として TF を組み合わせることを考える。

PP と TF の組み合わせ尺度を FPP とすれば

$$FPP(W) = TF(W)PP(W)$$
$$\Rightarrow \log_2 FPP(W) = \log_2 TF(W) + \log_2 PP(W)$$
$$\Rightarrow \log_2 FPP(W) = \log_2 TF(W) + \frac{\sum_{i=1}^{n} \left(H(R^i) + H(L^i) \right)}{2n}$$

$TF(W) = 0$ の場合 $\log_2 TF(W) = -\infty$ になってしまい、数量的な比較ができなくなるため、以下のようなスムージングを行う。

$$\log_2 FPP(W) = \log_2 (TF(W) + 1) + \frac{\sum_{i=1}^{n} \left(H(R^i) + H(L^i) \right)}{2n}$$

5 章に示す TF･PerPlexity の結果は上述の式によって計算された。

4.実験方法

4.1 実験対象コーパス

今回、今回実験に使用したテストコレクションは 2001 年 4 月から 2002 年 3 月までの毎日新聞 Web 記事である。毎日新聞 Web 記事は、経済・社会・国際・政治の 4 ジャンルにわかれており、それぞれのジャンルの各記事には、新聞社でそれを人手で要約してつくったと思われる携帯端末向け記事が対応している。それぞれのジャンルの最大記事数は表 4-1 のようになっている。

	経済	国際	社会	政治
記事数	4177記事	5952記事	6153記事	4428記事

表 4-1

以下は国際ジャンルにおける記事の一例である。

Web記事

イラン：治安部隊が「アルカイダ」戦闘員数人逮捕　日刊紙報道　【カイロ小倉孝保】イランの日刊紙「ホラサン」は１３日、イラン治安部隊がアフガニスタンから逃亡したテロ組織「アルカイダ」の戦闘員数人を逮捕した、と報じた。　イラン国会国家安全委員会のタルカシバン氏の話として伝えた。同氏は「治安部隊はイランに不法入国したアルカイダの戦闘員を逮捕し、さらに国境付近を捜索している」と語った。　米国はイスラム原理主義組織タリバンやアルカイダの戦闘員がイランに逃亡するのを防がなかった、としてイラン政府への批判を強めている。これに対しイラン政府は、公式には「イランはアフガン国境を封鎖しており、不法入国者は存在しない」と米国の指摘を否定している。

携帯端末向け記事

イラン治安部隊が「アルカイダ」の戦闘員数人を逮捕…イラン日刊紙。部隊は国境付近での不法入国者捜索を続行。

本研究では、人手によって要約された記事に含まれる語の大半は重要語であろうという前提のもと実験を行った。

4.2 前処理

Web 記事を ChaSen で形態素解析し、接頭語接尾語などを隣接する語句に連結する処理を行っている。また、助詞など明らかに候補語とならないものを ChaSen の出力結果によって除外している。

4.3 コーパスのランダム生成

今回の実験では小さいサイズでの精度を比較するため、1 記事のみからなるコーパスに対し、ランダムに記事を 1 つずつ追加してより大きなコーパスを生成していくことにより、最大 50 記事まで、50 種類のサイズのコーパス集合を構成した。各記事サイズで今回提案する手法も含め重要語抽出実験を行い、抽出結果を携帯端末記事に出現するかどうかによって評価し、平均精度を計算した。尚、コーパスをランダムで生成したことで結果がそれに依存しないよう、5 回実験した平均を示す。

5.実験結果及び考察

5.1 実験結果

まず、毎日新聞経済ジャンルにおける 50 記事までの実験結果を示す。尚、他 3 つのジャンル（政治・社会・国際）も同様の傾向を呈した。横軸は記事数、縦軸は平均精度で図 5-1 は 5 回試行の結果である。

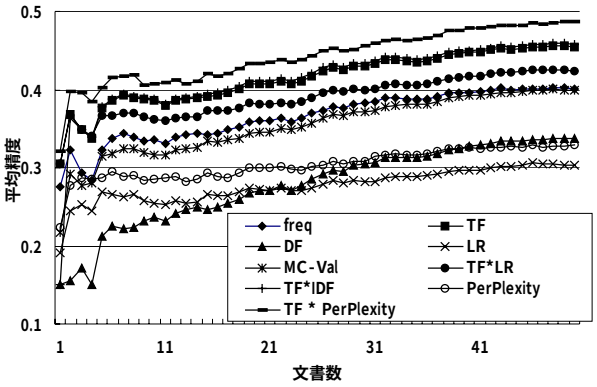

図 5-1：経済ジャンル 50 記事までの実験結果

頻度に着目する手法である Freq、TF、TF･IDF、DF などは文書サイズの増大に伴って平均精度が上昇している。MC-Value については、TF などと同様に文書サイズ増加に伴って平均精度は上昇している。この実験結果からも、頻度に着目する手法は、文書サイズが大きくなるほど有利になることが確認された。 逆に、複合語の構造に着目する手法である LR や PerPlexity は文書サイズの変化にともなう平均精度の変化は緩やかである。しかし、頻度情報を加味した FLR や TF･PerPlexity では、文書サイズの増加に伴って平均精度が上昇していることがわかる。

5.2 PerPlexity 尺度の有効性

今回提案した PerPlexity を用いた尺度は同じく複合語の連接に着目する LR を全体的に超えることはできたが、頻度を用いた手法である TF、TF･IDF、MC-Value などには平均精度という視点では及ばなかった。しかし、TF と組み合わせる事で、全ての既存手法を安定して上回ることができた。

5.3 抽出対象サイズが大きい場合

今回提案した PerPlexity や TF･PerPlexity は、文書サイズが大きい場合にどのような平均精度を示すのか。1000 記事までに範囲を広げて 50 記事ごとの実験を行った。

図 5-3 に示すように、TF･PerPlexity は 1000 記事までの全実験サイズで既存手法を上回ることができた。よって本手法は、大きな記事の集合に対しても既存手法より精度が高いと言える。また、特に小さい 100 記事程度までの抽出対象では、大きく既存手法を越えることができた。

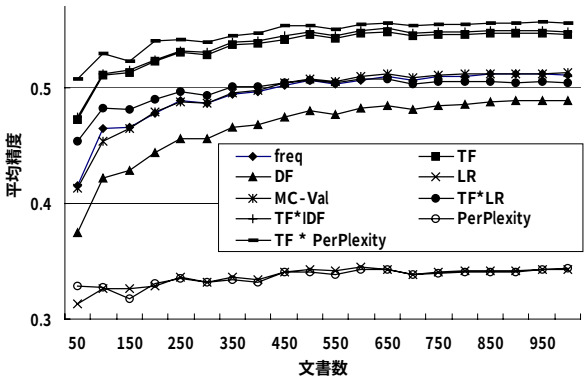

図 5-3：政治ジャンル 1000 記事までの実験結果

6.おわりに

本研究では、連接頻度・種類数を使用する LR に対して PerPlexity の手法を導入することで、情報理論的により精密な尺度を提案し、実験を行った。その結果、今回我々が提案した TF･PerPlexity 尺度は、小さい文書サイズにおいてのみならず、大きな文書サイズにおいても既存手法を上回ることが確認された。

参考文献

[1] S. Ananiadou.“a methodology for automatic term recognition”. *In Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pp. 1034–1038, 1994.

[2] Hiroshi Nakagawa.“Automatic Term Recognition based on Statistics of Compound Nouns”. *Terminology*, Vol. 6, No. 2, pp. 195–210, 2000.

[3] 中川裕志，湯本紘彰，森辰則.“出現頻度と連接頻度に基づく専門用語抽出”. 自然言語処理，10(1), pp.27–45, 2003.

[4] 内山将夫，井佐原均.“複数尺度の統計的統合法とその専門用語抽出への応用”. 情報処理学会研究報告，pp. 69–76, 2003