

テキストマイニングのための情報抽出 —情報レベルの最適化—

4N-7

長野 徹

日本アイ・ビー・エム（株）東京基礎研究所

1 はじめに

テキストマイニングは、膨大なテキストデータからの知識発見を目的としているが、データマイニングと異なり、自由に記述されたテキストが対象となる。そのため、個々のテキストからいかに適切な情報を含むデータを抽出するかが重要な課題である。

ワードクラスタリングのようにテキストに含まれる情報を統計的に用いるような場合、まず形態素解析を行い、単語レベルのデータを情報抽出の対象として用いるのが一般的である。本研究の対象であるテキストマイニングのフレームワーク「TAKMI」[1]では、単語レベルの情報に加え、係り受けの情報やモーダルから得られる情報[2]を付加することで、より多くの情報を含んだデータを抽出することができる。

テキストデータの特徴の1つとして、その定義域（例えば、出現する単語の集合）が非常に広いということがある。そのため、データの分布が非常に広範囲になり、しかも抽出される情報が多ければ多い程、データの分布は「薄く」となるという問題がある。

また、テキストデータを統計的に扱う場合、抽出された各要素が同レベルの情報を持っていることが期待される。

TAKMIでは、主に辞書を用いて単語をカテゴリーに分類している。その際問題となるのが、複合語に関する扱いであり、複合語の要素のうち、度の組合せを抽出すべきかを考慮する必要がある。情報検索の世界では、一般的に中頻度語に重要語が多いとされているが、明確な指標にはなっていない。

本稿では、名詞句に対して、適切な単位の情報抽出する手法についての考察を行う。

2 抽出の単位

名詞句中に含まれる形態素の組合せについて考えてみる。例えば、「東京基礎研究所」という名詞句（複合名詞）は4つの形態素「東京」「基礎」「研究」「所」を含む。この形態素列から情報を取り出す単位として、最も簡単な3つを以下に挙げる。

1. 短単位

短単位（1形態素を1単語とする）で情報をとると、上の4つの名詞が情報として得られる。この場合、1つの名詞句から多くの名詞が得られるが、全体としての単語の異なり数は少なくなるため、高頻度語が

多くなり扱いやすいが、それぞれの単語では意味をなさないことが多い。

2. 長単位

長単位（1名詞句を1単語とする）で情報をとった場合、「東京基礎研究所」1語になるが、全体としての異なり数は多くなる。したがって、単語が広く分布し低頻度語が多く現れることになるが、単語あたりの情報量は大きくなると考えられる。

3. 任意の単位

全ての組合せを考えて、この形態素列から複合名詞を作る場合、形態素列の長さを自由にとれるとすると、 $4+3+2+1=10$ 通りの名詞を作ることができる（形態素列長さ1の名詞も含む）。したがって、上の短単位と長単位の両方を含んでいる。この方法では、任意の単位の情報をとることができるが、単語の異なり数が非常に多くなる上、「東京基礎」のように不適切な語を抽出することになる。

よって、理想的には、全組合せの中で意味のある単位で情報を取り、その後、一般的な語や低頻度語を取り除く必要がある。

3 実験

3.1 抽出方法の違いによる比較

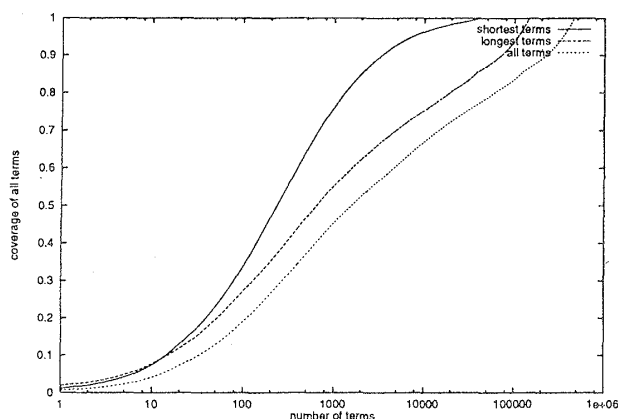
対象としたテキストは、230,360文で、平均17.2個の形態素を含む。このテキストを形態素解析した後、名詞句の自立語のみを取り出した。名詞句の数は、737,488個で、1文あたり3.2個。名詞句中の形態素の数は、1,310,408個で、1名詞句あたり平均1.77個の形態素を含む。

図1は、名詞句に含まれる情報を、短単位・長単位・任意の単位で抽出した場合の総単語数とその異なり数を示す。

	単語数	異なり数
1. 短単位	1,310,408	41,556
2. 長単位	737,488	144,874
3. 任意の単位	2,360,588	461,551

図1: 抽出単位による単語数と異なり数

図2は名詞句の抽出方法の違いによる単語数の分布の違いを示している。横軸（対数スケール）は抽出した単語の異なり数であり、出現頻度の高いものから順に並べられている。縦軸は、この出現頻度順に並べられた単語が、それぞれの手法で抽出された単語の集合全体に占め

図 2: 頻度上位 n 語の単語数に占める割合

る割合を示しており、頻度上位 n 語までが全体の単語数に占める割合を表している。

また、抽出手法の違いによる情報の分布を調べるために、異なり語の出現頻度の分布について調べた。

	平均頻度 μ	分散 σ
1. 短単位	31.53	269.2
2. 長単位	5.09	71.9
3. 任意の単位	5.14	83.0

図 3: 平均頻度と分散

また、抽出手法の違いによる、情報量の分布を調べるために、この出現頻度順に並べた単語の記述長（形態素を単位としている）を調べた。例えば「研究所」という単語は2つの形態素「研究」「所」を含むので、長さは2になる。ここでは簡単のため、各形態素が同じ情報を持っていると仮定すると、構成する形態素列の長さを、記述長としている。

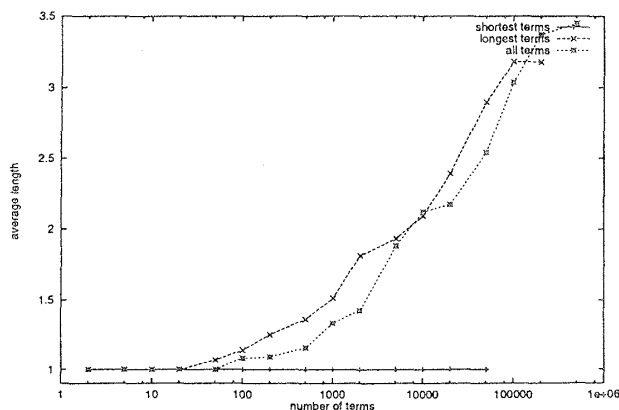


図 4: 抽出単位による単語長

3.2 考察

図 2 から、抽出する情報が多い程、データが薄く分布することが分かる。最も情報量の少ない短単位では、上位 1000 語で総単語数の約 76% をカバーできるが、情報量の多い長単位では約 55% しかカバーしない。

また、図 3 から、同様のことがいえる。短単位では平均頻度が高いので、分散が大きくても異なり語は少ない。長単位では、平均頻度が低いので、分散は小さいが、異なり語の数は大きくなっている。

図 4 は、出現頻度の低い単語ほど記述長が大きくなっている。

4 おわりに

当面の目的は、統計量を用いた名詞複合語の抽出である。

テキストデータを統計的に扱おうとする場合、データの情報レベルに留意する必要がある、またテキストから抽出するデータが多ければ多い程、データの分布が薄くなるという問題がある。

今回は、テキストデータの性質を示してのみであったが、今後はテキストマイニングに情報抽出のモデルを構築する予定である。

参考文献

- [1] 那須川哲哉, 諸橋正幸, 長野徹. テキストマイニングー膨大な文書データの自動分析による知識発見ー. 情報処理, Vol.40 No.4, pp.358-364 (1999)
- [2] 長野徹, 諸橋正幸, 那須川哲哉. テキストマイニングのための情報抽出手法. 人工知能学会 第 13 回全国大会 pp.411-412 (1999)