異なる類似尺度に基づいた要約文章の比較

大石 亨　　松本 守生　　日原 亜沙子

明星大学 情報学部 電子情報学科
oishi@ei.meisei-u.ac.jp
98j1111@edu.meisei-u.ac.jp
98j1096@edu.meisei-u.ac.jp

1 はじめに

電子化されたテキストの増大に伴い、テキストの自動要約の必要性が高まっている[奥村1999]。現在行われている研究の多くは、テキスト中から重要な情報を伝えている文を抜き出して、抄録を作成することを目的としている。抄録を作成する際には、テキスト中の語の頻度、各文の出現位置等、さまざまな情報が用いられるが、抽出された文の、話題に関する一貫性を保つという点では、語彙的結束性に基づく手法が優れていると考えられる。語彙的結束性とは、関連性のある語彙が用いられることで、複数の文間のつながりを明示することである[Halliday1976]。この情報を要約の知識源として用いた研究としては、[Hoey1991], [佐々木1993], [Barzilay1997], [望月1998] などがある。望月らは、関連する語彙の出現を語彙的連鎖(lexical chain)と呼び、これを含めた複数の抄録作成モデルを実装し、それぞれを索引付けに利用したときの情報検索の精度を比較している。[望月1998]

本稿では、「関連性のある語彙」を獲得するための情報として、シソーラス、ドキュメント空間上のベクトル類似度、係り受け解析に基づく確率分布の相違度、相互情報量の四つを用いた。それぞれの情報によって得られたクラスタに属する語彙を基準として、語彙的連鎖を計算し、その重要度に基づいて抄録を作成した。

10個の文章に対して作成した5種類の抄録を6人の被験者によって評価した結果を報告する。

2 語彙的連鎖に基づく抄録

ここでは、四つの処理に共通した抄録の作成手法と実験に用いたデータセットについて述べる。

要約対象は、情報検索システム評価用ベンチマークBMIR-J2[木谷1998]のテキストデータベースである94年版毎日新聞の5,080記事から30文以上の記事10個を選択した。なお、以下で述べる連鎖の重要度計算と、ベクトル類似度の計算に用いた文書集合も、この5,080記事である。要約率は25%とした。

抄録生成の手順は以下のとおりである。

1. 要約対象テキストを形態素解析する。
2. ストップワード（助詞・助動詞・「する」・「なる」などの頻出語、あらかじめ登録してある）を除く。
3. 先頭の形態素から順に、同一のクラスタに属する形態素と、その形態素を含む文の文番号を連鎖として登録する。
4. 連鎖の重要度を計算する。
5. 連鎖の重要度を文ごとに加算して、文の重要度を決定し、上位の文から要約率で規定された文数に達するまで出力する。

シソーラスの参照と、類似度の計算は3.の連鎖作成時に動的に行った。連鎖は、出現順にすべての異なり語に対して、それよりも後ろにある同一クラスタの形態素を登録するので、同一の形態素が複数の連鎖に登録される場合がありうる。

4. の連鎖の重要度計算は、望月らと同様、tf*idfの考え方に基づく次の式によった。

$$w_c = |c| \times \log \frac{N}{df_c} \qquad (1)$$

ここで、w_cは連鎖cの重要度、$|c|$は連鎖cを構成する形態素数、Nは全ドキュメント数（5,080）、df_cは連鎖cに含まれる形態素の出現文書数の最大値である。

5.では、連鎖を構成する形態素の出現する文に、連鎖の重要度を足しあわせて、文の重要度を計算した。単語数による正規化は行っていない。

3 単語の類似尺度

本節では、連鎖を構成する形態素のクラスタを決定する四つの類似尺度について述べる。

3.1 シソーラス

最も単純な手法は、既存のシソーラスを用いることである。本研究では、分類語彙表（増補版）[中野1996]を用いた。増補版分類語彙表では、延べ87,743語が、842の意味クラスに分類され、さらに段落と呼ばれる10,347のクラスに細分類されている。後者では、1クラスあたりの平均単語数が8.48であり、クラスタを構成するには小さすぎるので、実験では、前者をクラスタとして用いた。なお、分類語彙表に存在しないものは、単語そのものをクラスタとした。以下では、説明のため、この尺度を用いたときの結果を"Bunrui"と表記する。

3.2 ドキュメント空間上のベクトル類似度

二つ目の尺度は、ドキュメント空間上のベクトル類似度に基づくものである。類似尺度には、以下の式で計算される余弦を用いた。

$$\cos(x,y) = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} \qquad (2)$$

x_i, y_iは、i番目の文書中における、形態素xとyの出現頻度である。また、N=5,080である。$\cos(x,y)$の値が0.8以上の語を同一クラスタに属するものとした。この尺度を用いたときの結果を"SVM"と略記する。

3.3 条件付確率の分布の相違度

三つ目は、確率分布の異なりを測る尺度であるinformation radius(IRad)[1],を用いるものである[Dagan1997]。IRadは、以下の式で定義され、0から2log2までの値をとる。

$$\mathrm{IRad}(x,y) = D(x \| \frac{x+y}{2}) + D(y \| \frac{x+y}{2}) \qquad (3)$$

ここで、$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$であり、相対エントロピー(KL divergence)を表す。

実験では、毎日新聞94年版全記事(101,058記事)を係り受け解析し、名詞が格助詞を介して動詞または「サ変名詞＋する」にかかっているもの1,903,130件を抽出し、係り語の分布による受け語のIRad、受け語の分布による係り語のIRadをそれぞれ計算し、この値が0.85以下のものを、同一クラスタとした。それぞれを、"IRad-1","IRad-2"とする。

3.4 相互情報量

四つ目の尺度は、係り語と受け語の相互情報量を用いるものである。前節のIRadが係り語同士、受け語同士の相違度を範列的(paradigmatic)に表すのに対し、この尺度は係り語と受け語の統辞的(syntagmatic)な関連を表す。相互情報量は、以下の式で定義される[Church1990,Hindle1993]。

$$I(x,y) = \log_2 \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N}\frac{f(y)}{N}} \qquad (4)$$

ここで、$\frac{f(x)}{N}, \frac{f(y)}{N}$は、形態素$x$と$y$の独立の生起確率、$\frac{f(x,y)}{N}$は同時確率を表す

[1] Daganらは、"total divergence to the average"と呼んでいる。

が、実験では、前節と同様の係り受けデータを用い、語 y が語 x に係っている回数を $f(x,y)$ とした。$N=14{,}692{,}941$ である。$I(x,y)$ または $I(y,x)$ の値が5.0以上のものを同一クラスタとした。この結果は、"MI"と略記する。

4 抄録の評価結果と分析

10個の文章に対して、前節で述べた4種類の類似尺度に基づく5つの抄録を作成し、6人の被験者に評価してもらった。評価は、原文をまず読んでもらい、次に5種類の抄録をランダムな順にすべて提示する。それぞれの抄録について、原文の内容情報を提供する度合い(informativeness)を、［1:非常に悪い－2:悪い－3:普通－4:良い－5:非常に良い］の5段階でランク付けしてもらった。結果を表1に示す。値は、6人の評価の平均値である。

文章によってばらつきはあるが、平均すると、SVMとIRad-2の評価が低く、IRad-1とMIの評価が比較的高い。ここでは、統計的に有意であるかどうかという定量的分析よりもむしろ、定性的な分析を行う。抄録結果の評価が良かったとしても、その理由を説明することが、さらにシステムを改善するためには必要であるからである。

ベクトル類似度は、同じ話題領域に属する単語を捉える方法として、従来から広く用いられてきた。しかし、文書集合の設定によっては、全く関連のない語が、特定文書に出現することによってのみ、同一クラスに属するものとみなされてしまう可能性がある。実際、特に評価値の低かった（平均1.7）文章1では、「蒸し暑い-陳列-昔話-合法-長蛇-里帰り-…」という連鎖の重要度が最も高く、これが出力された抄録に悪影響を与えている。一方、この文章で最も評価の高かったBunruiでは、1.3330（「生活」）、1.2301（「国民」）、等のクラス及び「カストロ」「ハバナ」等の語からなる連鎖の重要度が高い。この文章はまさに「キューバの国民生活」について述べているものである。

また、IRad-2は受け語の分布による係り語の相違度を用いた分類であるが、この場合には、一般的な語ほど連鎖を構成しやすいという傾向が見られた。最も評価値の低かった文章6では、「もの-感-今-全体-意味-ひとつ-…」という連鎖の重要度が高かったが、この連鎖には65個もの形態素が含まれている。重要度計算の式（1）におけるtfに対応する部分の値があまりに大きいために、idfに対応する部分が効いていないのである。一方、IRad-1では、同じ文章でもっとも重要度が高いのは「宇宙」と「実験」という語のみからなる連鎖である。この文章は「宇宙実験」についてのものである。IRad-1でも「予定-計画」や「実施-行う」という一般的な動詞が連鎖を構成するのであるが、連鎖自体がそれほど大きくないので、影響しなかったのである。

IRad-1で評価の高かった文章3、文章8をみると、いずれも一つの単語からなる連鎖が上位を占めている。特に固有名詞が多い。これらは、係り語の分布による受け語の相違度では、重要度計算に影響を与えるような連鎖が構成されなかったことが、かえって良い抄録を生成する場合があることを示している。SVMでは、固有名詞に対しても、他の語と

表1 評価結果

尺度	文章1	文章2	文章3	文章4	文章5	文章6	文章7	文章8	文章9	文章10	平均
Bunrui	3.8	3.0	2.8	2.3	3.7	3.3	3.7	3.5	3.2	2.8	3.2
SVM	1.7	2.5	3.3	2.8	3.3	3.7	3.0	2.0	2.5	3.0	2.8
IRad-1	3.7	3.7	4.0	3.8	3.0	3.2	3.5	4.2	3.2	3.3	3.6
IRad-2	2.8	3.2	2.8	3.2	3.3	2.0	2.8	2.3	2.7	3.0	2.8
MI	3.5	4.2	4.0	4.3	3.8	3.7	2.5	3.8	3.0	4.0	3.7

の連鎖を構成してしまうことがあるが、これが逆効果となっているのが文章8の場合である。上述した文章1の場合と同様、無関係の語の中に、重要な固有名詞が紛れ込んでしまっている。

さて、本実験で最も評価値が高かったのは相互情報量を用いたMIである。この場合には、IRad-1に見られたような消極的な理由ではなく、MIによって構成された連鎖が積極的に抄録の評価に貢献している。実験で最高の評価を得た文章4は、「スイスの外国人雇用情勢」に関するものであるが、「スイス」「ジュネーブ」「ログェ」という固有名詞とともに、「銀行-就職-解雇-…」や「外国-入国-許可-流入-…」、「失業-避ける-減らす-追い込む-失業」といった興味深い連鎖が構成されている。これらの連鎖により、計算が難しいとされている代名詞や省略による結束性にも対処できる。すなわち、相互情報量を連鎖の構成に用いることで、文中の主語が省略されているような場合にも、述語の連鎖によってその文の重要度が極端に低下することが避けられるのである。

5 おわりに

本稿では、語彙的結束性に基づく抄録作成において、シソーラス、ドキュメント空間上のベクトル類似度、係り受け解析に基づく確率分布の相違度、相互情報量という異なる類似尺度を用いた。10個の文章に対して、それぞれの情報によって得られたクラスタに属する語彙を基準として、語彙的連鎖を計算し、その重要度に基づいて5種類の抄録を作成した。

6人の被験者によって評価した結果、

① ベクトル空間モデル（SVM）は、話題領域に特有の語彙を明らかにするが、重要でない語彙をも含んでしまうことがある。

② IRadは、高頻度で一般性の高い語ほど相違度が少なく（類似度が高く）なる傾向がある。

③ 相互情報量は、最も評価が高かったが、これは、文中の主語が省略されているような場合にも、述語の連鎖によってその文の重要度が極端に低下することが避けられるためである。

という知見が得られた。

謝辞

実験データとして、毎日新聞記事、BMIR-J2、分類語彙表を、また、形態素解析に「茶筅」、係り受け解析には「cabocha」を使わせていただきました。関係各位に深く感謝いたします。

参考文献

奥村学,難波英嗣.(1999) テキスト自動要約に関する研究動向.自然言語処理,Vol.6,No.6,pp.1-26.

Halliday H. A. K. and Hassan R.(1976) *Cohesion in English.* Longman.

Hoey M.(1991) *Patterns of Lexis in Text.* Oxford University Press.

佐々木一朗,増山繁,内藤昭三.(1993) 語彙的結束性に着目した文章抄録法の提案.情報処理学会研究会報告NL-98-9,pp.65-72.

Barzilay R. and Elhadad M.(1997) *Using Lexical Chains for Text Summarization.* In Proceedings of the ACL Workshop on Intelligent Scallable Text Summarization, pp.10-17.

望月源,岩山真,奥村学.(1998) 抄録を利用したテキスト検索.言語処理学会第4回年次大会併設ワークショップ「テキスト要約の現状と将来」論文集,pp.22-29.

木谷強,他.(1998) 日本語情報検索システム評価用テストコレクションBMIR-J2,情報処理学会研究会報告DBS-114-3,pp.15-22.

Dagan I., Lee L. and Pereira F.(1997) *Similarity-Based Methods for Word Sense Disambiguation.* In Proceedings of 35th Annual Meeting of the ACL and 8th Conference of the EACL, pp.56-63.

Church K. W. and Hanks P.(1990) *Word Association Norms, Mutual Information, and Lexicography.* Computational Linguistics, Vol.16, No.1, pp.22-29.

Hindle D. and Rooth M.(1993) *Structural Ambiguity and Lexical Relations.* Computational Linguistics, Vol.19, No.1, pp.103-120.

中野洋.(1996) 「分類語彙表」形式による語彙分類表（増補版）.