質問応答エンジンを利用した複数文書要約手法

森 辰則 † 野澤 正憲 ‡ 浅田 義昭 ‡
† 横浜国立大学 大学院 環境情報研究院 ‡ 横浜国立大学 大学院 環境情報学府
E-mail: {mori,nozawa,asada}@forest.eis.ynu.ac.jp

1 はじめに

大量の文書が溢れている昨今，その中から必要とされる情報を効率良く見つけたいという要求がある．情報検索や質問応答等の技術により情報要求に関連する文書群や答え自身を容易に得る事が出来るようになりつつあるが，最終的には原文書を調べる必要がある．

これらの技術と相補的な関係にあるのが，検索文書群を対象とした複数文書要約技術である．特に，情報検索過程においては利用者が情報要求を持っており，また，それらが質問文として記述できるという考え方から，近年，「質問の答に焦点を当てた要約」(Answer Focused Summarization) が注目されている．平尾ら [5] は質問型に一致する固有表現に高い重みを与え要約文書を生成する手法を提案している．Wu ら [3] は質問型と要約生成の単位となる文書断片の長さの関係を検討し，これに基づいた要約文書生成手法を提案している．いずれも一つの質問文に注目した個別文書の要約である．一方，Gaizauskas らは，複数文書要約と質問応答の融合を目指すとしているが，詳細は未公開である．

複数文書要約においては内容把握ができるように，ある程度の要約文書量が必要となるため，利用者の知りたい事柄の各々について要約文書を生成すると最終的に利用者が読むべき文書量が増えてしまう．そこで，本稿では，複数の要求の答とその背景知識を一度に概観できるような要約 (抜粋) を生成する一手法について述べる．特に，質問応答エンジンを用いた文重要度計算を汎用の文重要度計算に融合する手法を提案する．

2 提案手法の概要

本稿では，要約対象文書群は，情報検索等の結果として得られており，また，利用者の情報要求は，複数の質問文として与えられているとする．この状況下では，複数文書要約のために，1)「情報要求を考慮した重要箇所抽出」，2)「文書間の冗長箇所の削除」，3)「文書間の相違点の抽出」が必要であると考える． 提案手法では，これらについて以下の技術を用いる．

(a) 質問応答エンジンの出力スコアに基づく文の重要度計算 (上記 1 に対応．第 3 節．)

(b) 語の出現分布に関する情報利得比に基づく文の重要度計算 (上記 1，3 に対応．第 4 節．)

(c) MMR に基づく要約文書中の冗長性の制御 (上記 2 に対応．第 5 節．)

更に抽出文間の結束性の担保のために「(d) ハニング窓関数に基づく文重要度平滑化」(第 6 節) を採用する．

図 1 に提案手法に基づき構築したシステムの概略を示す．本システムへの入力は要約対象となる文書 (の ID) の集合，情報要求に対応する質問文の集合，ならびに，求める抜粋の長さ (文字数もしくは文数) である．出力は文書集合の抜粋 (文の列) である．例えば，今回の評価で用いたテストコレクションである NTCIR4 TSC3

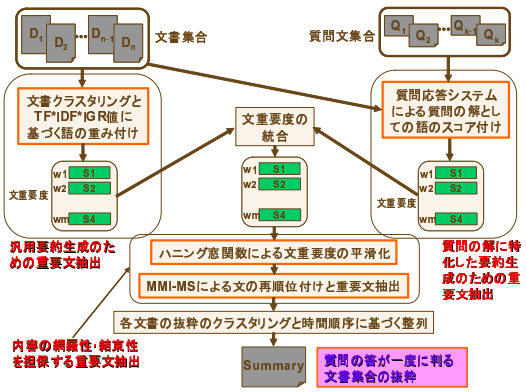


図 1: 質問応答システムを利用した複数文書要約

の Topic 0500 の場合，図 2 に示す 9 つの文書 ID，図 3 に示す 10 の質問文，ならびに，要約文書長 491 文字が入力である．この時，本システムは図 4 に示す抜粋を出力する．この中でゴシック体になっている部分が質問の答の一つである．

3 質問応答エンジンの出力スコアに基づく文の重要度計算

我々の提案している質問応答システム [2] は，様々な文書や検索エンジンを利用できるようにするために，対象文書に対する前処理を必要としない仕組みとなっている．質問文が与えられた後に，形態素解析や構文解析，固有表現抽出などといった計算コストの大きい処理をするが，これを必要最小限に抑え実時間処理を行なうために，A* に基づく解の探索制御を導入している．

同システムのエンジンは質問文が 1 つ与えられると対象文書中の各語 (形態素) に対して解としての「良さ」を表すスコアを付与する．我々はこのスコアを質問の解に注目した時の語の重要度と考え，重要文抽出を行なうことを提案する．このスコアを利用することにより，質問型の情報のみを利用する従来手法よりも精度の高い要約生成を行なえると期待される．

本稿では，複数の質問文が与えられることを想定しているため，形態素毎にスコアの組が求められる．各スコアはある質問文に対応する．質問文の複雑さや質問の型によりスコアの値が変動するため，本来，異なる質問文に対して求められたスコアは比較可能ではない．しかし，ある形態素について複数の質問文に対する単一の重要度を付与したいので，元のスコアを比較可能な値に正規化する．さて，ある一つの質問文に注目しその答を見つける際には，各語のスコアの絶対値は重要でなく，他の語のスコアとの相対的関係が重要である．そこで，本稿では，スコアの平均値からの隔たりが重要であると考え，質問文毎に語のスコアを偏差値 (T-score) に変換し，これを正規化スコアとした．正規化スコアは複数の質問文に亙って平均値が同一に

JY-19990402J1TYEUG0400060, JY-19990527J1TYMAJ1400040, JY-19980424J1TYMAK1400070, JY-19980723J1TYMAJ1400050, JY-19980301J1TYMAP1400050, 980110135, 980723029, 980424152, 980215018

図 2: 入力例: 文書 ID の集合 (NTCIR4 TSC3 Topic 0500)

「ドリー」は何の名前か？ / クローン羊ドリーはどこで誕生したか？ / クローン羊ドリーは、胎児細胞ではなく何の複製であることが実験で確認されたか？ / クローン羊ドリーは何からつくり出されたか？ / クローン羊ドリーの元になった雌羊の細胞とドリーの細胞とで、何が同一であると確かめられたか？ / 英国のロスリン研究所を率いているのは誰か？ / クローン羊ドリーが妊娠中の羊から採った乳腺細胞をもとにつくりだされたことについて、どのような批判があったか？ / クローン羊ドリーの細胞の寿命は普通の羊と比べてどうであることが分かったか？ / クローン羊ドリーの出産はいつか？ / クローン羊ドリーの出産により何が確認されたか？

図 3: 入力例: 質問文の集合 (NTCIR4 TSC3 Topic 0500)

８日付のイブニング・スタンダード紙によると、英国のロスリン研究所は、世界で初めて体細胞を卵子に組み込んで誕生させた雌のクローン羊「ドリー」(生後１年半) と雄の羊を交尾させたことを明らかにした。一昨年、世界初のクローン羊ドリーをつくることに成功した英エディンバラのロスリン研究所は２３日、ドリーがこのほど出産、正常な生殖能力があることが確認されたと発表した。詳細は二十三日発行の英科学雑誌「ネイチャー」に掲載される。ドリーは、妊娠中の羊から採った乳腺 (にゅうせん) 細胞をもとにつくり出された。この雌羊はすでに死んでいるが、組織の一部は、英国内で凍結保存されている。英国ロスリン研究所のイアン・ウィルムット博士のグループと、英国レスター大のグループは、クローン羊ドリーが大人の雌羊の体細胞核移植によるクローンであることをそれぞれＤＮＡ鑑定で裏付け、２３日発売の英科学誌「ネイチャー」に発表した。生まれたのは雄二頭と雌一頭で、いずれも元気だという。ドリーを誕生させた英国のロスリン研究所などのチームが二十七日発売の英科学誌「ネイチャー」で発表する。すでに二回妊娠し、元気な子供も産んだ。

図 4: 出力例: 要約文書 (抜粋)

なる．質問文 q に関する形態素 w の正規化スコアを $score^n(w,q)$ とする時，文 S_i の重要度 $Imp^n_{QA}(S_i)$ を式 (1) で求める．ただし，Q は与えられた質問文の集合，W_{S_i} は文 S_i に現れる形態素の集合である．いずれかの質問の答えが含まれているかと言う観点から文の重要度を決めるため，式 (1) ではある文の重要度をその文に含まれる形態素の最大スコアとしている．

$$Imp^n_{QA}(S_i) = \max_{w \in W_{S_i}, q \in Q} score^n(w,q) \quad (1)$$

4 語の出現分布に関する情報利得比に基づく文の重要度計算

我々は検索結果文書の各々を要約する手法として，情報利得比に基づく語の重み付けを用いた重要文抽出手法を提案している [4]．この手法では，検索結果文書間の類似性構造を階層的クラスタリングにより抽出し，その構造に則した出現分布を持つ語に高い重みをつけるため，情報利得比 (Information Gain Ratio,IGR) に基づく語の重要度計算を行なう．C_i を C の部分クラスタとするとクラスタ C における単語 w の情報利得比 $gain_r(w,C)$ は次のように求められる．

$$\begin{aligned} gain_r(w,C) &= \frac{info(w,C) - info_{div}(w,C)}{split_info(C)} \\ info(w,C) &= -p(w|C)\log_2 p(w|C) \\ &- (1-p(w|C))\log_2(1-p(w|C)) \\ info_{div}(w,C) &= \sum_i \frac{|C_i|}{|C|} info(w,C_i) \\ split_info(C) &= -\sum_i \frac{|C_i|}{|C|}\log\frac{|C_i|}{|C|} \end{aligned}$$

ここで次の二点に注意しなければならない．

1. 対象文書群が情報検索結果であれば，それらと検索されなかった文書群との対比が重要であるため，根クラスタの上に仮想的なクラスタを設ける．このクラスタには検索文書の属する部分クラスタとそれ以外の文書が属する部分クラスタが存在する．同仮想クラスタでは，対象文書群全体に関連する語に高い重みが与えられるので，検索要求に関する語が高く重みづけられる．
2. 階層的なクラスタリングを考える場合，クラスタ毎に語の重みが得られるので，これらを統合する必要がある．本稿では，各文書の所属するすべてのクラスタにおける語の重みの和を採用する．これを IGR_sum と呼ぶ．

そして，この重みと文書内単語頻度 (TF 値) や文書頻度の逆数 (IDF 値) など既存の重みづけ手法を組み合わせることにより，最終的な語の重みとする．この語の重みに基づく各文 S_i の重要度 $Imp_{IGR}(S_i)$ は，式 (3) に示すとおり，含まれる名詞の重みの総和を文の長さ (単語単位) により正規化したものである．また，文書間の文重要度を正規化するために，文書内の文重要度を偏差値 (T-score) に変換する．これを $Imp^n_{IGR}(S_i)$ とする．更に，前節の式 (1) と式 (3) を統合した文重要度として，式 (3) を考える．ここで，α は文重要度 Imp^n_{QA} の Imp^n_{IGR} に対する重みである．

$$Imp_{IGR}(S_i) = \frac{\sum_{w \in Noun(S_i)} TF(w,D)\cdot IDF(w)\cdot IGR_sum(w,D)}{|S_i|} \quad (2)$$

$$Imp^n(S_i) = \alpha \cdot Imp^n_{QA}(S_i) + (1-\alpha) \cdot Imp^n_{IGR}(S_i) \quad (3)$$

5 MMR に基づく要約文書中の冗長性の制御

重要文抽出において，Carbonell らが提案する MMR[1] と同種の冗長性制御機構を導入することを考える．MMR は，本来，文書もしくはパッセージを単位として，順位づけを行なうものであり，初期順位は検索質問に対する文書の類似度を用いる．これを式 (4) のように文を単位とし，初期順位を文の重要度により与えるように変更する．本稿ではこれを MMI-MS(Maximal Marginal Importance − Multi-Sentence) と呼ぶ．

$$MMI\text{-}MS(SS, A) \stackrel{\text{def}}{=} Arg \max_{S_i \in SS \setminus A} [\lambda Imp^n(S_i) - (1-\lambda) \max_{S_j \in A} Sim_s(S_i, S_j)] \quad (4)$$

ここで，SS は要約対象の文集合，A は既選択文の集合，$Imp^n(S_i)$ は式 (3) に定義される文 S_i の正規化重要度，Sim_s は文間の類似度を表す尺度である．A に空集合を，冗長度制御変数 λ に適切な値を設定してから式 (4) を繰返し適用すると，冗長性を考慮した文の再順位づけがなされる．なお，本稿では，Sim_s として文ベクトルの cosine 類似度を採用した．同ベクトルの各次元は，各文に含まれる名詞であり，その値は対応する名詞の重要度である．

6 ハニング窓関数による文重要度平滑化

ここまでの文重要度計算手法では各文を独立に扱うため，対象文書数が多い時には多くの文書から少しずつ重要文を抽出し，文間の結束性が低下する傾向が見られる．要約文書長が長い場合には，文の重要度を考慮しつつも，文間の結束性を高める事が必要である．そこで，ある文数の範囲内で重要度が滑らかに変化するように，ハニング窓関数を用いた重要度の平滑化を行なう．窓幅 W の同関数を用いて平滑化した文重要度は式 (5) により与えられる．なお，文書の先頭と末尾においては，その文が連続するものとして計算する．

$$Imp^n_c(S_i) = \sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} \frac{1}{2}(1 + cos 2\pi \frac{j-i}{W}) \cdot Imp^n(S_j) \quad (5)$$

7 実験と評価

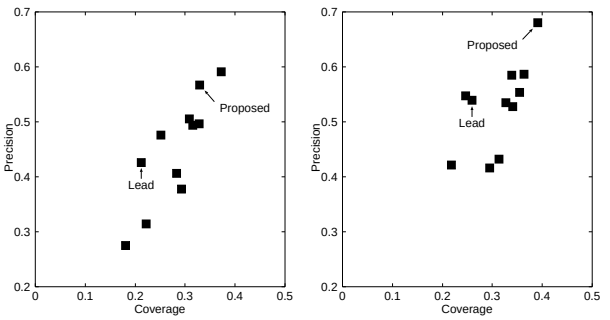
本節では，評価型ワークショップである NTCIR4 TSC3 における Formal Run の課題により提案手法に基づくシステムを評価する．同ワークショップは現在進行中であり，報告会は 2004 年 6 月に開催される．本稿執筆時点では，他参加システムの詳細は不明である．本稿では，タスクオーガナイザが用意したモデル抜粋との比較による抜粋の性能，ならびに，モデル要約との比較による質問に対する解の被覆率に基づき評価を行なう．

同 Formal Run の課題は，30 トピックからなる．各トピックは，要約対象文書 ID のリスト，トピックの表題 (検索要求を簡潔に表現したもの)，生成すべき要約文書の長さ (文字数，ならびに，文数．いずれも短いもの (Short) と長いもの (Long) の二種)，要約に含まれるべき事項を表した質問文のリスト (Short 用と Long 用の二種) から構成される．要約対象文書は 98,99 年の毎日及び読売新聞の記事から選ばれている．

提案システムの各種パラメタは，Formal Run に先だって配布された例題 5 トピックにより調整を行なった．Short 用にはハニング窓関数を適用せず，Long 用には窓幅 4[1] とした．二種類の文重要度 Imp^n_{QA} ならびに Imp^n_{IGR} の混合比を決めるパラメタ α の値は 0.8(Short 用) ならびに 0.7(Long 用) とした．MMI-MS 用のパラメタ λ は $0.4+0.5\cdot(1-Sim_{ave})$ とした．ここで Sim_{ave} はトピック毎の平均文間類似度である．

7.1 重要文抽出の性能に関する評価

モデル抜粋を正解として，提案システムの出力抜粋の平均被覆率 (Average Coverage) ならびに平均精度 (Average Precision) を調べた．図 5(a)，(b) に示す．図中のラベル 'Proposed' は提案手法，'Lead' は Lead 手法 (各文書の先頭部分を抽出する) によるベースライン，それ以外の点は他の参加システムである．



(a) Short (b) Long

図 5: 抜粋の平均被覆率ならびに平均精度

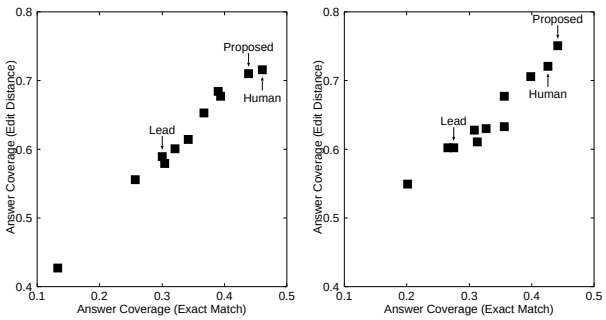
7.2 質問に対する解の被覆率に基づく評価

モデル要約に含まれる質問文の解が提案システムの出力抜粋に含有される度合 (解の平均被覆率) を調べた．図 6(a)，(b) に示す．尺度としては，正解文字列そのものが現れる割合の平均値 (Exact Match)，ならびに，式 (6) により定義される正解文字列 Ans_i と文 S の間の編集距離 $EditD()$ に基づく尺度の平均値 (Edit Distance) の二種類を考える．

$$Cov_{ED}(Ans_i) = \max_S \frac{Len(S) - EditD(S, Ans_i)}{Len(Ans_i)} \quad (6)$$

ここで，関数 $Len()$ は文字列の長さを返す．図中のラベル 'Human' は人間が作成した要約 (モデル要約とは異なる) である．

[1]前後 1 文ずつを考慮することになる．

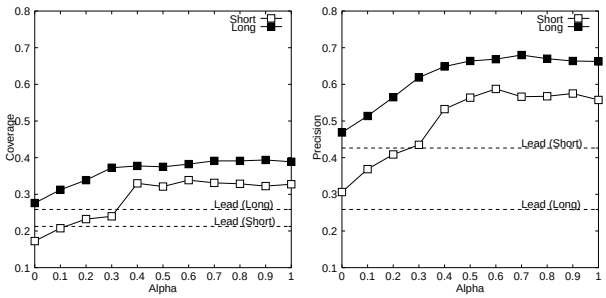


(a) Short (b) Long

図 6: 質問に対する解の平均被覆率

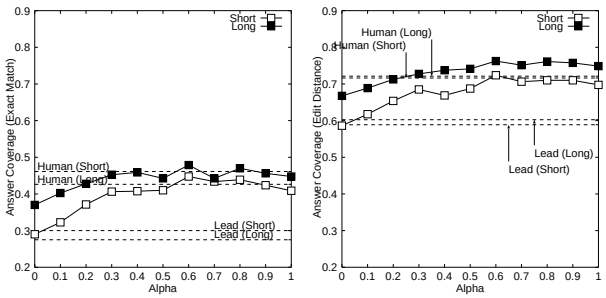
7.3 二つの文重要度の混合比に関する評価

二種類の文重要度の混合比が各種性能に与える影響について調べるため，他のパラメタは前述の通りに固定しつつ，パラメタ α の値を変化させて同様の評価を行なった．図 7 に抜粋の性能変化を，図 8 に質問に対する解の平均被覆率を示す．



(a) 被覆率 (b) 精度

図 7: 文重要度混合比 α の変化に対する抜粋の性能変化



(a) Exact Match (b) Edit Distance

図 8: 文重要度混合比 α の変化に対する質問応答の性能変化

8 考察

図 5 によると，提案手法では Lead 手法と比較して，Short，Long の抜粋長について，抜粋の平均被覆率ならびに平均精度共に改善されていることがわかる．また，TSC3 に参加した他システムと比較しても，その優位性が示されている．ところで，Long については抜粋精度が 0.680 と高いのに対して，抜粋被覆率は 0.391 と低い．これは，別の文書に由来する同一もしくは非常に似通った文を抽出する例が見受けられるためである．出力文書の冗長制御を行なっている MMI-MS では，文名詞の重要度を成分とする文ベクトルの類似度を用いているが，各語の重要度は文書によって異なるために，全く同一の文であっても類似度が 1 にならない．文間類似度計算の精緻化が必要である．

図 6 によると，質問の解の被覆率について，提案手法は Lead 手法と比較して，Short，Long の要約長のいずれにおいても，改善されていることがわかる．特に Long においては，人間が作成した要約と比較しても，解の被覆率が高いことがわかり，質問応答エンジンを利用した効果が現れている．更に，図 7 ならびに図 8 によると，二種類の文重要度のうち，質問応答のスコアに基づく文重要度が支配的であることがわかる．ただし，いずれも $\alpha = 0.6 \sim 0.8$ の箇所に性能の頂点が存在するので，両重要度を考慮したほうが良い．特に質問応答評価においても $\alpha = 1.0$ ではない箇所に頂点があることが興味深い．採用している質問応答エンジンは，NTCIR QAC1 の質問セットにおいて MMR が 0.5 程度[2]であり精度が十分ではなないことから，IGR に基づく文重要度がこれを補っていると考えられる．

9 まとめと今後の課題

本稿では，複数の情報要求に対して一度に答えることができる複数文書要約を目標として，質問応答エンジンを用いた文重要度計算を汎用の文重要度計算に融合する手法を提案した．NTCIR4 TSC3 Formal Run に基づく評価により，その有効性を示した．

今後の課題としては，先に述べた MMI-MS における文類似度の精緻化の他に，質問の解解析の高速化が挙げられる．現在は，与えられた文書中の全ての形態素についてスコアを求めているため，平均的な PC を利用した時に 1 質問当り数十秒の処理時間がかかっている．我々の質問応答エンジンでは，解の探索制御を行なっているために，上位の解のスコアが求まった段階で解析を打ち切ることができるので，その際の近似スコアの利用を今後検討したい．

参考文献

[1] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proc. of SIGIR 98*, pp.335–336, 1998.

[2] T. Mori, T. Ohta, K. Fujihata, and R. Kumon. An A* search in sentential matching for question answering. *IEICE Trans. Info. & Sys.*, Vol. E86-D, No. 9, pp. 1658–1668, 2003.

[3] H. Wu, D. R. Radev, and W. Fan. Towards answer focused summarization. In *Proc. of the 1st Int'l Conf. on Information Technology and Applications*, 2002.

[4] 森辰則. 検索結果表示向け文書要約における情報利得比に基づく語の重要度計算. 自然言語処理, Vol. 9, No. 4, pp. 3–32, 2002.

[5] 平尾努, 佐々木裕, 磯崎秀樹. 質問に適応した文書要約手法とその評価. 情報処理学会論文誌, Vol. 42, No. 9, pp. 2259–2269, 2001.

[2]順位つきの解出力を行なった時に，正解の順位が平均すると 2 位程度であることに相当する．