情 報 学 基 礎 60－4
（2000．9．27）

話題の階層構成に基づく関連談話の対応づけ

仲尾 由雄
富士通研究所
〒211-8588 川崎市中原区上小田中 4-1-1
nakao@flab.fujitsu.co.jp

関連文書の組から、関連箇所を自動的に抽出する手法を提案する。語彙的結束性に基づき認定した話題階層を利用して、様々な粒度の話題を単位に、文書の部分間の関連度を計算し、話題の粒度に応じた関連度をもつ関連箇所の対を抽出する手法である。本手法を、国会における代表質問と答弁を対象に実験を行った結果、抽出された関連箇所の組の約8割は正しく同一の話題に対応し、また、新聞に要旨として掲載された内容の約6割は、この手法で自動的に抽出可能であることがわかった。これにより、完全に対応する文書であれば、話題階層に基づき関連話題を絞り込むことで、予め特別な閾値を設定することなく、効率的に様々な粒度の話題に対する関連箇所が検出できる見込みが得られた。

Automatic Discovering of Related Passages among Related Documents

Yoshio Nakao
Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211-8588 Japan

This paper presents an algorithm for discovering related passages among related documents. For the documents to be compared, the algorithm first detects their thematic hierarchies individually based on lexical cohesion measured by term repetitions. Then, it compares a pair of thematic hierarchies in terms of various grading topics, and selects closely-related pairs of thematic units across them. An experiment using proceedings of interpellations in the National Diet shows the precision rates of related topic selection are estimated to be about 80 percent and the recall rates for major related passages corresponding to manual summaries of these proceedings are estimated to be about 60 percent.

1 はじめに

本稿では、複数の関連文書から関連箇所を抽出する手法を提案する。本研究の最終的な目標は、複数の関連文書を比較しながら閲覧したい利用者に対し、関連箇所をわかりやすく提示して、比較作業の効率を高めることにある。例えば、ある調査項目について複数の地域の実情を調査レポートにまとめるために、各地域の調査担当者から寄せられた調査レポートを読むこと、あるいは、質問状と回答書を読み比べることの支援などが目標である。

複数の文献の比較支援に関し、Neuwirth[1] は、関連論文に見られる一致点・相違点を、著者と命題 (proposition) という 2 つの観点から整理して一覧表形式で提示する "Synthesis Grid" というインターフェースを提案している。本研究は、そのような情報を自動作成するための第一歩として、関連文書から共通する話題を取り扱った関連箇所を自動的に切り出す手法を提案するものである。

文書の関連箇所を抽出する従来の研究として、同一語彙の出現などを手がかりに、関連文書中の関連箇所にハイパーリンクを設定する手法が知られている。例えば、文献 [2] では、文書中の一節に相当する「セグメント」を単位に文書を分割し、語彙的類似度の高いセグメント間にハイパーリンクを設定する手法が示されている。

また、関連箇所の検出に関する技術として、文献 [3] では、語彙的類似性の高い段落群を検出することで、単一文書中の関連箇所を抽出する手法が示されている。また、文献 [4] では、文書中の語彙の連鎖などに基づく活性伝搬ネットワークを用いて、共通の関連語群を含む文などを検出する手法が示されている。

しかしながら、これらの手法には、関連箇所を認定する単位が固定的であるため、粒度の異なる話題に対して、適切な関連箇所の検出が難しいという問題がある。つまり、従来の研究では、節・段落・文 (または語の出現位置そのもの) のいずれか一つに比較の単位を固定しているため、基本的に検出できるのは、節対節、段落対段落など、比較の単位の大きさの箇所同士に限られることになる。そのため、例えば、第 1 の比較文書中で 2 段落からなる箇所が、ひとつのまとまりとして、第 2 の比較文書中の数段楽以上の大きさの箇所と関連している場合などには、対比すべき関連箇所を適切に切り出すことが難しい。それを実現するためには、関連箇所として検出された箇所を併合するなど、何らかの別の手段を講じることが必要になる。

そこで、本稿では、語彙的結束性に基づき文書中の話題の階層構成を認定する手法 [5] に基づき、様々な粒度の関連箇所の検出を試みる。以下、第 2 章で話題の階層構成に基づく関連箇所検出手法を説明し、第 3 章で国会会議録データを対象に行った実験について報告する。

2 話題階層に基づく関連箇所の検出

本稿で提案する関連箇所検出手法は、まず、比較する文書対のそれぞれについて、語彙的結束性に基づき、話題階層を認定する。次に、認定した 1 対の話題階層を比較して、それらを構成する話題同士の関連度を求め、関連度の高い話題を関連話題として抽出する。

以下、例を交えながらこれらの処理について説明する。関連文書の例としては、「第 149 回衆議院本会議会議録第 2 号」(2000 年 7 月 31 日) から、水島広子議員による代表質問とそれに対する首相の答弁を、それぞれ 1 つの文書として切り出したもの用いた (以降、それぞれ「質問文書」「答弁文書」と称する)。国会の代表質問は、党を代表する議員がいくつかの項目を一括して質問した後、首相・関係大臣が答弁する形で進められるが、この代表質問では、子供の教育、民法改正、国会運営、有害情報、小児医療、歳費支給方式の 6 つの問題に関し、計 8 項目が質問されている。

2.1 話題階層の認定

提案手法では、まず、[5] で提案した話題構成認定手法に基づき、文書中の話題階層を認定する。ここで、話題階層とは、大きさの異なる複数の話題のまとまりが 2 段以上の階層構造を成していることを意味する。話題のまとまりとは、文書中である粒度の話題に関して記述している一続きの部分のことである。このような話題のまとまりの集合で、階層構造をなしているものを、本稿では話題階層と称する。

この手法では、まず、[6] にならい、文書中の各位置の前後に、求めたい話題の大きさ程度 (文書全体の 1/4〜段落程度) の窓を設定し、その 2 つの窓に出現する語彙の類似性を測定する。類似性は、次に示す余弦測度 (cosine measure) で測定する。

$$sim(b_l, b_r) = \frac{\Sigma_t w_{t,b_l} w_{t,b_r}}{\sqrt{\Sigma_t w_{t,b_l}^2 \Sigma_t w_{t,b_r}^2}}$$

ここで、b_l, b_r は、それぞれ、左窓 (文書の冒頭方向側の窓)、右窓 (文書の末尾方向側の窓) に含まれる文書の部分であり、w_{t,b_l}, w_{t,b_r} は、それぞれ、単語 t[1] の左窓、右窓中での出現頻度である。本稿では、この値を結束度と呼び、また、結束度に対応する窓の境界位置によって結束度を並べたものを結束度系列と称することにする。

次に、上記の結束度を、ある刻み幅 (窓幅の 1/8) で窓をずらしながら測定して、文書の冒頭から末尾に至る結束度系列を求める。そして、結束度系列の極小点を手がかりに話題境界を認定する。この際、結束度系列の移動

[1] 「語」は動詞・名詞・形容詞のいずれか。詳細については 3.1 節で説明する。

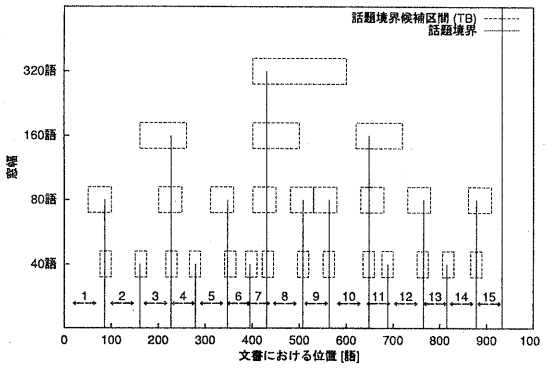

図 1: 話題境界の認定結果

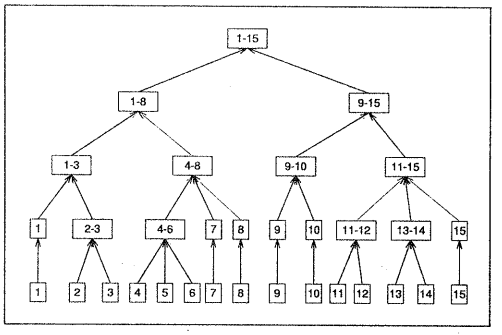

図 2: 話題階層の認定結果

平均をとることで、任意の大きさの話題のまとまりを選択的に検出できるようにし、また、極小となる移動平均の値に対して大きく寄与している文書中の範囲 (窓幅の 1/2〜1 程度) を求めて境界位置の候補区間を作成する。

以上の操作を、窓幅を変えて行うと、大きな窓幅では大きな話題の切れ目に、小さな窓幅では小さな話題の切れ目に対応する、境界候補区間が認定できる[2]。そして、境界候補区間の重なりを手がかりに、大きな窓幅よる話題境界と、より小さな窓幅による話題境界を統合することで、話題階層を形づくる (詳細は [5])。

今回の実験では、最小窓幅を 40 語とし、80 語、160 語、…のように等比級数的に文書サイズの 1/2 を超えない範囲で拡大した数種類の窓幅を用いて、それぞれに対する話題境界候補区間を認定した後、最大窓幅による境界候補区間から順に、一回り小さい窓幅で認定した境界候補区間との統合操作を行い、話題境界を認定した。

図 2.1 は、第 1 の実験対象文書中の話題境界の認定結果である。図中、横軸は語単位に測った文書における位置であり、縦軸は、結束度系列の計算に用いた窓幅であり、矩形が上述の話題境界候補区間である。そして、話題境界を貫く点線の棒グラフが仮境界位置 (結束力拮抗点:語彙的結束度の移動平均値に基づき検出した最も境界位置らしい点) を示している。

最後に、語単位で認定した仮境界位置を微調整して、文境界に合わせてから、各境界の間を一つの話題とする話題階層を作成する。例えば、最小窓幅に対応する境界からは、矢印で示された 15 の区間に対応して、15 の話題が最下層の認定される。また、80 語の窓幅に対応する境界からは、15 の話題のうち、区間 2 と 3、区間 4 から 6、区間 11 と 12、区間 13 と 14 の 4 グループに対応する話題が統合された、計 10 個の話題が第 2 層の話題として認定される。なお、仮境界位置の調整は、境界候補区間中の文から、話題の立ち上がり位置に相当する文を検出する手法 [7] によって行っている。

図 2 は、このようにして作成した話題階層である。図において、矩形であらわされたノードは、認定されたそれぞれの話題である。ノード内の数字は、図 19 で示した区間の番号に対応する。

表 1 は、今回の実験対象とした 11 対の質問・答弁文書 (詳細後述) に対して認定した話題階層について、それに含まれる区画数を、階層別に集計したものである。表中括弧内は、直下の階層中の区画と同一の区画の数である。この表に示されるように、上記の手順で認定した話題階層は、1 階層下る毎に区画数が約 2 倍に増える。すなわち、平均として、上位層中の 1 区画には直下の層の約 2 区画が含まれており、したがって、平均的な区画の大きさは、1 階層下るごとに約半分になることになる。これは、話題境界認定の際に、認定用窓幅を 2 の等比級数による拡張しことに対応する。また、表 1 の区画数と次章の表 2 に示した文書の平均語数に関し、質問文書対答弁文書の比を比較すると、両者ともほぼ 3:2 となっている。これは、話題境界認定手法は、同じ窓幅で認定した区画は、いずれの文書に対しても、ほぼ同じ大きさであることを示唆している。すなわち、今回の実験でも、話題境界認定手法は、窓幅にほぼ比例した (窓幅の 1/2〜2 倍程度の) 区画が認定されていることがわかる。

2.2 関連話題の抽出

関連話題の検出手順を図 3 に示す。

図 3 では、まず、第 1 の話題階層中の任意の話題と

[2] 調査報告書や新聞の特集記事などをテストデータとして用いた実験 [5] によれば、窓幅程度の大きさの話題の 7 割程度 (再現率) は、境界候補区間内に含まれることが期待される (適合率は 5 割程度)。

表 1: 話題階層の認定結果の概要

階層	区画数 質問		答弁	
6	5	(0)	1	(0)
5	17	(0)	7	(0)
4	35	(6)	20	(1)
3	63	(15)	42	(6)
2	124	(27)	88	(13)
1	258	(43)	164	(39)
延べ	505	(91)	322	(59)

入力 比較対象文書対に対応する 2 つの話題階層: $TH1, TH2$

出力 関連話題対の集合: T

1. $t1 \in TH1, t2 \in TH2$ なる話題対 $(t1, t2)$ のそれぞれに対し、関連度 $R(t1, t2)$ を求める。
2. $t1 \in TH1$ に対し、$t1$ 以下の部分木における最大関連度を求め、$t1.max$ に記録する。
3. $t2 \in TH2$ に対し、$t2$ 以下の部分木における最大関連度を求め、$t2.max$ に記録する。
4. 以下の話題対の集合 T を求め出力する。

$$T \equiv \{(t1, t2) \mid R(t1, t2) \geq \max(t1.max, t2.max)\}$$

図 3: 関連話題抽出アルゴリズム

第 2 の話題階層中の任意の話題からなる話題対の全てについて、関連度を計算する。話題 $t1$ と話題 $t2$ 間の関連度 $R(t1, t2)$ は、$t1, t2$ のそれぞれに対応する文書の区画 $s1, s2$ に含まれる語彙の類似性に基づき、以下の式で求める。

$$R(t1, t2) \equiv R(s1, s2) = \frac{\Sigma_t w_{t,s1} w_{t,s2}}{\sqrt{\Sigma_t w_{t,s1}^2 \Sigma_t w_{t,s2}^2}} \quad (1)$$

ここで、$w_{t,s1}, w_{t,s2}$ は、それぞれ、区画 $s1, s2$ における単語 t の重要度に相当する重みであり、以下の式により計算する。

$$w_{t,s} = tf_{t,s} \times \log(\frac{|D|}{df_t})) \quad (2)$$

ここで

$tf_{t,s}$ 単語 t の区画 s における出現頻度

$|D|$ 区画 s を含む文書を固定幅 (80 語) 刻みに区切ったブロックの数

df_t 単語 t が出現しているブロック数

である。

これらの式は、情報検索分野で検索対象文書と質問文との関連度計算などでよく使われる $tf \times idf$ と呼ばれる計算法の変種である。$tf \times idf$ では、上記の $\frac{|D|}{df_t}$ の部分を、文書内の区画ではなく、検索対象文書集合に含まれる文書を単位に計算する。すなわち、$|D|$ を検索対象文書集合中の文書数とし、df_t を単語 t が出現する文書数とすると、上記の式は通常の $tf \times idf$ の計算式となる。

式 2 の代わりに通常の $tf \times idf$ の式を用いることも可能だが、今回は以下の理由によりこの式を用いた。

- 式 2 は、比較対象文書だけから関連度が計算できる (idf を計算する文書集合を用意する必要がない)。
- 式 2 は、[8] で報告した見出し語抽出実験によれば、単語の重要語のよい尺度である[3]。
- 上述の話題境界を文境界に合わせる微調整の際にも式 2 を利用している[4]

次に、第 1 の比較文書中の話題 $t1$ と第 2 の比較文書中の話題 $t2$ の全てに対して、話題階層を利用しながら、話題対を選別するための閾値を求める。今回は、話題階層の部分木中の最大関連度を閾値として用いた。ここで、ある話題 t に対する話題階層の部分木中の最大関連度とは、t もしくは話題階層における t の子孫、すなわち、t を構成するいずれかのより小さい話題に対して計算された関連度の最大値のことである。

図 2 では、ステップ 2 で、第 1 の比較文書中の話題 $t1$ について上記の最大関連度を求め、$t1.max$ に記録し、ステップ 3 で、第 2 の閲覧対象文書中の話題 $t2$ についても同様に、最大関連度を $t2.max$ に記録する。そして、

$$T \equiv \{(t1, t2) \mid R(t1, t2) \geq \max(t1.max, t2.max)\}$$

なる話題対の集合 T を求め、関連話題として出力する。

関連話題の抽出例

関連話題抽出処理の意味を具体例に基づき補足する。

図 4 は、上記手順で抽出された話題対を実線のアークで、それ以外の話題対で関連度が 0.25 以上の値を持つものを点線のアークで示した図である。アークに添えられた数値は、関連度の値である。図中、2 つの木構造グラフは、左のグラフが第 1 の比較文書に、右のグラフが第 2 の比較文書に対応する。

[3] この実験の際に、形式段落を単位に idf を計算する手法も試したが、固定ブロックを用いた場合よりよい精度は得られなかった。

[4] 調査報告書、新聞の連載記事などを用いた実験 [9] によれば、見出しによる文境界を認定する確率が高い (話題境界候補区間に見出しが含まれる場合なら 8 割程度) という性質がある。

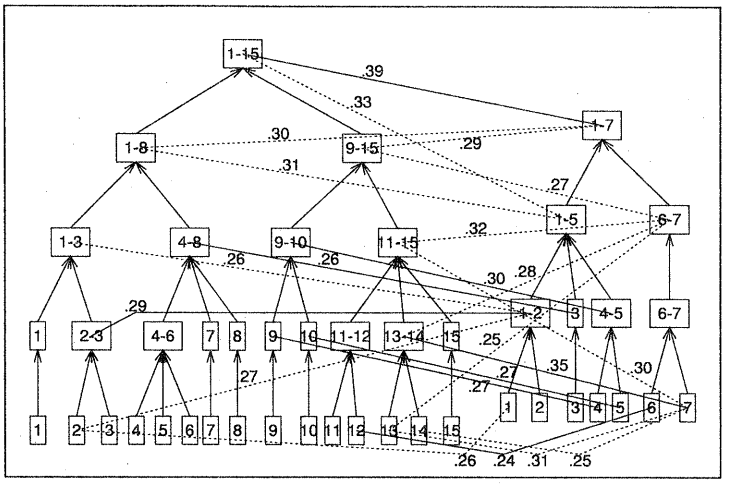

図 4: 関連話題抽出例

ここで、右のグラフの右下角のノード (以降「ノード右 7」のように参照) に着目する。このノードは、第 2 の文書の最後の最小区画に対応するノードであり、グラフ上では、末端ノードとなっている。よって、このノードにおける最大関連度は、このノードに直接結びつけられたアークの関連度の最大値である。ノード右 7 では、(ノード左 13-14, ノード右 7) の話題対の関連度 0.35 が最大関連度となる。そして、ノード左 13-14 からは、0.35 を超える関連度をもつアークはないので、(ノード左 13-14, ノード右 7) の話題対は関連話題として出力される。

一方、ノード右 6-7 に着目すると、このノード以下の部分木にノード右 7 が含まれているので、ノード右 6-7 に直結しているアークは、少なくともノード右 7 の最大関連度 (0.35) 以上でなければ、関連話題として出力されない。ノード右 6-7 にはこのようなアークはなく、ノード右 6-7 を含む話題対は関連話題としては出力されない。

以上のように部分木における最大関連度を基準に選別することで、関連話題の候補を大幅に削減し、実線で示した 8 対の話題対を関連話題として抽出できる。関連話題は、文書全体同士の話題対を除けば、7 対のみであるにも関わらず、関連箇所が検出できなかった話題は、ノード左 1、ノード左 11、ノード左 15 のみとなっている。関連箇所が検出できなかった話題のうち、質問項目を含むものは、ノード左 15 のみで、後のものは、後続の話題を導入するための役割を担った、答弁とは直接的には関連しない内容の部分であった。また、検出された 7 対の関連話題は、後で示す手順で評価した結果、いずれも適切に対応する内容を含む部分であった。

なお、ノード左 9, 10 の部分は、(ノード左 9-10, ノード右 4-5) だけでなく、(ノード左 9, ノード右 4) と (ノード左 10, ノード右 5) の話題対にも重複して関連する部分である。以下に示した内容[5] にみられるように、ノード左 9 に対する答弁はノード右 4 であり、ノード左 10 に対する答弁はノード右 4 であるが、ノード左 9 とノード左 10 との間、および、ノード右 4 とノード右 5 との間にも、強い関連性が読み取れる。このように同一の箇所に関係する、複数の意味のある関係も抽出できることは、様々な粒度で関連箇所を比較することの一つの利点と考えられる。

○水島広子君 [9-10]	○内閣総理大臣 (森喜朗君) [4-5]
¶9 総理御自身も触れられている**大人社会**のあり方ですが、これが子供たちに大きな影響を与えるのは事実だと思います。… モラルの低下の一つの例として、子供の目に触れる**テレビ**や**雑誌**、**ゲーム**などの影響も無視できません。だれでも**簡単**に目にするメディアに**暴力**や性**暴力**が**はんらん**し、町じゅうに売春情報があふれているというのが今の**大人**の**社会**です。…	¶4 **テレビ**や**雑誌**、**ゲーム**などの青少年を取り巻く環境について、**暴力**や性犯罪が**はんらん**しており、青少年にとって大きな問題であるとの御指摘でありますが、これらの問題は、申すまでもなく**大人社会**の**責任**であります。…
¶10 …諸外国でも進められているように、子供たちを**有害**な**情報**から守る**法律**を日本でも**早急**につくる必要があると思います。…	¶5 また、子供たちを**有害情報**から守るための**法律**の**早急**な制定を促す御意見をいただきました。…

[5] 切り出した関連箇所に含まれる重要語数語を与えて [10] の方法で要約した結果。

3 提案手法の評価

3.1 実験条件

実験に用いた文書

実験対象文書としては、国会会議録データ[6] から、衆議院の代表質問に関する会議録を用いた。具体的には、以下の 2 つの会議録から、計 11 対の代表質問・答弁を切り出して、実験対象文書とした。

- 「第 142 回衆議院本会議録第 6 号」(1998 年 2 月 18 日)
  橋本首相 (当時) の施政方針演説に対する代表質問を記録した会議録。羽田孜議員、加藤紘一議員、小沢辰男議員の 3 氏による代表質問を含む。
- 「第 149 回衆議院本会議録第 2 号」(2000 年 7 月 31 日)
  森首相の所信表明演説に対する代表質問を記録した会議録。鳩山由紀夫議員、小里貞利議員、水島広子議員、神崎武法議員、山岡賢次議員、不破哲三議員、土井たか子議員、野田毅議員の計 8 氏による代表質問を含む。

それぞれの文書の大きさを表 2 に示す。節・記事の大きさは、字数および次節で説明する「語」単位で示した。

表 2: 実験文書の構成

種類	字	語	文	段落
質問文書 (平均)	6443	1643	115	51
答弁文書 (平均)	4067	1114	51	37

これらの文書の特徴は、話題の開始位置と質問・答弁間の対応関係を客観的に認定しやすい点にある。

答弁文書は、それぞれの質問項目について、ほとんどの場合、質問項目を復唱する文 (「…についてお尋ねがございました。」など) を述べてから、答弁に移る形式がとられている。そのため、復唱する文のパタンを念頭に目視すれば、大部分の話題 (個々の質問項目に対する答弁) については、正確に認定できる。また、復唱内容を手がかりに、質問文書を目視すれば、ほとんどの場合質問の順序と答弁の順序が一致していることもあり、対応する質問箇所を簡単に見いだすことができる。

質問文書についても、背景を述べる文 (複数)、質問点・提案点を述べる文 (1〜数文)、答弁を要求する文 (「総理の見解を伺います。」など) という形式をとることが多いので、話題 (質問項目) の切れ目を、ある程度客観的に判定できる。ただし、このような様式は、質問者によって大きく異なる面があり、必ずしも客観的に話題が認定できるとは限らない。例えば、長い演説調の質問中に時折疑問文の形で質問点をあげているような質問文書の場合、対応する答弁と比較しても、どこからどこまでに対して答弁者が答えているのか判然としない場合もある。

単語認定

今回の実験における内容語抽出には、日本語形態素解析ツール jmor[11] を用い、名詞・動詞・形容詞) を切り出した。jmor によって切り出される名詞には、形容動詞語幹が含まれ、機能語や数字・時詞・相対名詞 (左右/上下/以上/以下など) は含まれない。また、jmor には名詞などの連続を複合語としてまとめて抽出する機能もあるが、この機能は用いず、個々の名詞を別々の語として扱った。例えば、「水」の質問文書の先頭の 1 文から以下の【】で囲まれたものが切りだされた。【】内の "/" の後ろは、活用語の終止形語尾である。結束度の計算においては、終止形語尾つきで表記が一致するものを同一の語とみなした[7] 。

> 私は、【民主党/】・【無所属/】【クラブ/】を【代表/する】して、【森/】【総理/】の【所信/】【表明/する】【演説/する】に【対/する】し【質問/する】いたします。

最小区画の認定精度

話題階層の最下層に位置する最小区画が、正確に認定できているかを評価する。

まず、表 3 に、最小区画と、それを隔てる話題境界と段落境界との一致状況を示す。段落境界は、話題境界とは限らないが、少なくとも今回用いた答弁文書の範囲では、個々の質問項目 (または質問点) に対する答弁の開始位置は段落開始位置と一致していた。

表中、「境界一致」とは、区画の先頭の境界が、段落境界と一致した場合に対応する。これによれば、答弁文書中の区画は、9 割以上という高率で、段落境界と一致していることがわかる。これは、答弁文書の方が語数当たりの段落数がやや多いことの影響も考えられるが、基本的には、前述のように答弁文書が整然とした形で、個々の質問事項に答えていることに由来すると見られる。質

[6] 国会図書館のホームページよりリンクのある会議録検索ページ http://kokkai.ndl.go.jp/ 、または衆議院のホームページ http://www.shugiin.go.jp/から入手可能。

[7] 「い/る」は "要る"、"居る" のいずれの意味でも同一の語とみなすことになる。また、「い/る」と「要/る」のように表記が違う語は例え意味が同じでも別の語とみなした。

問文書についての一致率は、答弁文書に比べると低いものの、語彙的結束性 (のみ) に基づく関連研究で報告されている認定精度とほぼ同レベルである。

表 3: 話題境界と段落境界の比較

種類	区画	段落	境界一致	単数段落
質問	258	565	167 (65%)	50 (19%)
答弁	164	411	151 (92%)	22 (13%)

※括弧内はランダム抽出に相当する基準値

表 4: 答弁文書の話題境界認定精度

正解の種類	境界数		再現率	適合率率
	認定	正解		
大話題	153	65	78% (37%)	33% (16%)
小話題	153	174	61% (37%)	69% (42%)

※括弧内はランダム抽出に相当する基準値

次に、3.1 節で述べた性質に基づき、答弁文書の話題境界認定精度を評価した結果を表 4 に示す。

ここでは、答弁文書中で質問項目を復唱している文を正解境界[8] として、再現率 (一致境界数/正解境界数)、適合率 (一致境界数/認定境界数) を求めた。なお、冒頭の境界 (今回の実験では必ず文書頭になる) は、比較に意味がないので、はずして集計した。

表中、「大話題」とは、関連答弁 (同じ質問項目に含まれる複数の質問点などに対する答弁) がある場合に、最初の復唱文のみを正解境界として集計した精度に対応する。例えば、「また、政治改革について種々の御提言がございました。」という 1 文のみからなる段落からはじまり、いくつか点について答弁が行われたあと、「以上、議員から政治改革についてのご提言をいただきましたが、…。」という締めくくりの 1 文で終わるような部分[9] が典型である。また、「小話題」とは、復唱文を正解境界とした場合の精度である。

なお、「大話題」には、上例のようなはっきりとした目印が添えられていないことも多く、質問文と答弁を見比べながら、1 つの質問項目から派生したとみられる部分をまとめたため、必ずしも客観的とはいい難い面がある。また、「大話題」による境界の直後には「小話題」の境界がくることが多いが、「小話題」による精度の集計においては、この点は特に考慮せず、いずれの正解境界と一致しても、「一致」と判定した。

表によれば、「大話題」に 8 割弱という語彙的結束性による話題境界認定としてはかなり高い再現率が得られており、小話題に関しても再現率約 6 割、適合率 7 割弱という標準的な精度が得られている。

[8] 質問項目に複数の復唱文がある場合には、先頭の文のみが正解。

[9] この例は、羽田議員の質問に対する橋本首相 (当時) の答弁の一部。

3.2 関連話題抽出精度の評価

抽出された関連話題が、同一の話題 (質問項目) に由来する質問-答弁の対であるかについて評価した。計 91 対 (文書全体同士の話題対は除外) 抽出された関連話題全てについて、質問・答弁文書を参照したところ、少なくとも 1 文ずつは同一話題の質問-答弁の箇所と対応していることが確認できた。ただし、最小区画の切り出し誤り、または、話題階層上位での、区画の統合不良によると見られる、関係のない話題に関する箇所を含むものも見られた。

そこで、関連話題に対応する区画に、不要な内容が含まれていないか、話題の粒度 (切り出された関連箇所の大きさ) が適切に対応しているかを評価した。具体的には、各関連話題対に対応する、答弁・質問文書の対応箇所 (以下、「答弁箇所」「質問箇所」と称す) に含まれている文を以下のように分析し、評価した[10] 。

- 答弁箇所に、質問文書の復唱文と一群の答弁文が全て含まれていれば、「完全」とする。ただし、全く関連しない質問項目が答弁箇所に混在している場合には、「完全」とはしない。
- 答弁箇所に、一群の答弁文が一部でも含まれていれば「許容」とする。逆に、答弁内容を含まない復唱文のみしか含まれていない場合には「不良」とする。
- 答弁箇所に対応する全ての質問項目に関し、質問箇所に、背景、質問点 (提案点)、答弁要求の一群が (存在するかぎり) 全て含まれていれば、「完全」とする。
- 答弁箇所に対応するいずれかの質問項目に関し、質問箇所に、対応する質問点が 1 つでも含まれていれば「許容」とする。逆に、質問点を含まない、背景、答弁要求のみの場合には、「不良」とする。
- 答弁箇所と質問箇所の両者が「完全」の場合、「完全」と判定する。
- 答弁箇所と質問箇所のいずれかが「不良」の場合、「不良」と判定する。
- それ以外は「許容」と判定する。ただし、全く関連しない質問項目が同じ程度の比率で混在している場合には、「不良」と判定する。

[10] 「質問点を含むか」「関連しない質問項目か」の判定が難しい場合があり、判定は、必ずしも一定していない部分も残っている。

表5に、この判定結果を示す。「完全」「許容」を合わせれば、8割強が正しく対応づけられていた。

表 5: 質問-答弁対応箇所切り出し精度

判定	箇所数	(構成比)
完全	22	(24%)
許容	52	(57%)
不良	17	(18%)
合計	91	(100%)

表 6: 質問文書関連箇所数の対象文書別比較

比較対象	抽出箇所	正解 (準正解)	
演説	75	51	(13)
別演説	57	31	(7)

3.3 その他の評価

主要な話題が網羅的に抽出できているかに関し、実験対象の会議録に対応する日本経済新聞に掲載された要旨との比較を行った。具体的には、2つの会議録に対応して、1998年2月19日朝刊4面、2000年8月1日朝刊6面の「衆院代表質問と答弁の内容」という記事中で、一問一答形式に要約された内容と、関連話題対(上記で「許容」「全体」とした対)に対応する質問箇所・答弁箇所を比較した。全54組の一問一答形式の要約のうち、その内容の一部ずつでも質問箇所・答弁箇所から読み取れたものが32組(59%)あり、うち、要約内容全部が含まれていたものは、22組(41%)あった。

また、完全には対応しない文書の比較における提案手法の効果に関し、簡単な補足実験も行った。この実験は、答弁文書のかわりに、代表質問の元になっている、橋本首相(当時)の施政方針演説と森首相の所信表明演説を使って関連箇所を抽出したものである。表6の結果中、「演説」は質問文書と対応のある演説を用いた場合を、「別演説」は質問文書と対応のない演説を用いた場合を示している。例えば、鳩山議員の質問文書の場合、「演説」では森首相の演説を、「別演説」では橋本元首相の演説を比較対象としている。この場合、特に「別演説」に関しては、何を「関連」と判定するかが問題になるが、今回は、完全に対応がないものを除去し、さらに、関連に疑問が残るを「準正解」として分離し、残りを「正解」とした。関連に疑問が残ったものの例としては、冒頭の挨拶同士が共通の話題(「有珠山」など)を含むために関係づけられたもの、「別演説」における「与党三党」「補正予算」など別の実体を指すと思われる語で関係付けられたものなどである。

4 まとめ

本稿では、関連文書中の様々な粒度の話題に対して、話題の関連性を判定する手法を提案し、実験により評価した。国会における代表質問と答弁を対象とする実験では、抽出した関連箇所の組の約8割は正しく同一の話題に対応し、また、新聞に要旨として掲載された内容の約6割は、この手法で自動抽出可能なことがわかった。このように、少なくとも完全に対応する文書であれば、話題階層に基づき関連話題を絞り込むことで、予め特別な閾値を設定することなく、効率的に様々な粒度の話題に対する関連箇所が検出できる見込みが得られた。

謝辞

本研究を進めるにあたり、東京大学理学系研究科の辻井潤一先生にご指導いただきました。ここに記して感謝いたします。

参考文献

[1] Neuwirth, C. M. and Kaufer, D. S.: The Role of External Representations in the Writing Process: Implications for the Design of Hypertext-based Writing Tools, in *Proc. of Hypertext '89*, pp. 319–341the Association for Computing Machinery (1989).

[2] 大森信行, 岡村潤, 森辰則, 中川裕志 : *tf·idf* 法を用いた関連マニュアル群のハイパーテキスト化, 情処研報 FI-47-8/NL-121-16 (1997).

[3] Salton, G., Singhal, A., Buckley, C. and Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes, in *Proc. of Hypertext '96*, pp. 53–65the Association for Computing Machinery (1996).

[4] Mani, I. and Bloedorn, E.: Summarizing Similarities and Differences Among Related Document, chapter 23, pp. 357–379, The MIT Press, London (1999), reprint of *Information Processing and Management*, Vol. 1, No. 1, pp. 1–23, 1999.

[5] 仲尾由雄 : 語彙的結束性に基づく話題の階層構成の認定, 自然言語処理, Vol. 6, No. 6, pp. 83–112 (1999).

[6] Hearst, M. A.: Multi-paragraph segmentation of expository text, in *Proceedings of the 32nd Annual Meeting Annual Meeting of Association for Computational Linguistics*, pp. 9–16 (1994).

[7] 仲尾由雄 : 話題の階層構成に基づく文書自動要約: 本一冊を一頁に要約する試み, 情処研報 NL-132-7 (1999).

[8] 仲尾由雄:文書の話題構成に基づく重要語の抽出, 情処研報 FI-50-1 (1998).

[9] Nakao, Y.: An Algorithm for One-page Summarization of a Long Text Based on Thematic Hierarchy Detection, in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (2000), (10月予定).

[10] 仲尾由雄 : 見出しを利用した新聞・レポートからのダイジェスト情報の抽出, 情処研報 NL-117-17 (1997).

[11] 西野文人 : 日本語テキスト分類における特徴素抽出, 情処研報 NL-112-14 (1996).