話題の階層構成に基づく関連談話の抽出

仲 尾 由 雄†

関連文書の比較作業を支援するには，類似の話題を扱った関連箇所を対比して提示することが有効と考えられる．また，関連箇所が，関連文書中でどのように分布しているかを図示することは，関連文書の対応関係を総合的に理解する助けとなろう．本稿では，このような考えから，文書間をまたがる関連箇所を自動的に抽出する新しい手法を提案する．提案手法は，文書対に共通する話題を検出し，それぞれの話題に対して関連箇所の対を抽出する手法である．この際に，語彙的結束性に基づき認定した話題階層を利用して，様々な粒度の話題間の関連度を求め，比較していることに特徴がある．国会における代表質問と答弁を対象に行った実験では，抽出された関連箇所の組の約 8 割が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，抽出された関連箇所の対から読み取れることが分かった．

A Method for Related-passage Extraction Based on Thematic Hierarchy

Yoshio Nakao†

This paper presents a novel method for extracting related passages in multiple documents that is intended to be used to help a person who wants to compare the content of multiple documents. The aim of the algorithm is to extract the best matching pair of document portions for each topic commonly included in the documents to be compared. For the documents to be compared, the algorithm first detects individual thematic hierarchies based on lexical cohesion measured by term repetitions. It then compares a pair of thematic hierarchies in terms of various grading topics, and selects closely-related pairs of thematic units. In an experiment using proceedings of interpellations in the National Diet of Japan, the algorithm extracted correct pairs of related passages in a ratio of 80% and identified 60% of major topics that had been reported in newspaper articles.

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，関連箇所を分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．たとえば，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 ら[13]は，関連論文にみられる一致点・相違点を，著者と命題（proposition）という 2 つの観点から整理して提示する “Synthesis Grid” という一覧表形式のインタ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 1 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研究として，同一語彙の出現を手がかりに，関連文書中の関連箇所を検出する手法が知られている．たとえば，大森ら[15]は，文書を節単位に分割し，語彙的類似度の高い節の間にハイパーリンクを設定する手法を示している．また，Mani ら[6]は，語の共参照（coreference）や連接（adjacency）などの関係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章の箇所（典型的には文）を検出する手法を示している．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，適切な関連箇所を検出するのが難しいという問題がある．つまり，節・段落・文（または語）のいずれか 1 つに比較単位を固定しているため，基本的に，節対節，段落

† 株式会社富士通研究所
Fujitsu Laboratories Ltd.

対段落など，比較単位の大きさの等しい箇所どうしが検出されることになる．そのため，たとえば，第1の比較文書中で2段落からなる箇所が，1つのまとまりとして，第2の比較文書中の数段落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．それを行うには，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なくてはならない．

Allanは，文書間あるいは文書の部分の間に自動的にハイパーリンクを設定する一連の研究[1)]の中で，一対の文書の部分間に設定したハイパーリンクを併合して整理する手法を示している．具体的には，語彙的類似性の比較により設定したハイパーリンク群のうち，端点が文書中で近い位置にあるリンクについて，リンクおよびリンクの端点にはさまれた段落群を併合することで，段落以上のまとまりどうしの関係としてハイパーリンクを設定できることなどを示している．しかし，このハイパーリンク併合手法では，併合対象段落間の関係（語彙的類似性など）が考慮されていないため，いくつかの異なる話題に関する段落群が偶然同じ順に並んでいるような文書対を比較した場合に，話題の区別がつかなくなる可能性があるという問題点が残る．この点に関連して，Saltonら[16)]は，語彙的類似性の高い段落群の文書内の分布を手がかりに連続する関連段落群を抽出する手法と，ある閾値より高い語彙的類似性を示す段落群を再帰的に集めることで，話題（“text themes”）に関する関連箇所を抽出する手法を示している．しかし，これらの手法☆にも，関連箇所の比較において，比較対象とする箇所の大きさに対する考慮が不足しているという問題点がある．Singhalら[17)]は，情報検索の分野で広く使われている類似度計算手法（重み付き単語ベクトルの余弦値）が，極端に大きい/小さい文書に対して安定でなく，不適切に低い/高い値になってしまうことを指摘している．Callan[2)]は，パッセージ検索において，形式段落を手がかりにしたパッセージを用いるより，ある固定幅（たとえば150〜300語）の窓を重なりを持たせながらずらすことで作成した仮想的パッセージを用いた方が有効であることを示している．これらは，関連パッセージを抽出する際に，対象とするパッセージの大きさの違いに注意しないとならないことを示唆している．特に，比較対象とするパッセージの大きさに大きなばらつきがある場合には，何らかの対処が必要となると考えられる．

☆ 文献16)に明記された内容においては主として後者の手法．前者の手法も，認定した段落群を単位に文書間の関連箇所を抽出しようとすると，同様の問題に行きあたると考えられる．

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[10)]を利用して，様々な粒度の関連箇所の検出を試みる．この話題構成認定手法は，TextTilingアルゴリズム[4)]をベースにした手順によって，文書を話題に関するいくつかの区画に分割する手法であるが，指定した大きさ程度の区画に分解できる点，したがって，文書全体を2〜4個に分割するような大まかな区画から，段落程度の大きさの細かな区画まで，様々な大きさの区画を体系的に求めることが可能な点に特徴がある．本稿では，この特徴を利用して，様々な粒度の話題に対して，適切な大きさの関連箇所を，できるだけ簡潔に切り出すことを主要な課題とする．

以下，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，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．4章で結論と今後の課題を述べる．

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，認定した一対の話題階層を比較して，それらを構成する話題どうしの関連度を求め，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を用いた（以降，それぞれ「質問文書」「答弁文書」と呼ぶ）．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，話題構成認定手法[10)]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区画が2段以上の階層構造をなしていることを意味する．話題区画とは，文書中である粒度の話題に関して記述しているひと続きの部分のことである．このような話題区画の集合で，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．また，特に区別が必要な場合以外は，話題区画のことを「話題」と略して呼ぶこととする．

この手法では，TextTilingアルゴリズム[4)]と似た手

順で文書を処理し，ほぼ同じ大きさの区画に分割する．文書中の各位置の前後に，求めたい話題区画の大きさ程度（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☆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呼ぶことにする．

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，結束度系列の移動平均をとることで，話題に関する記述箇所を，記述箇所の大きさ別に選択的に検出できるようにし，また，移動平均の極小値に対して大きく関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．そして，境界候補区間の重なりを手がかりに，大きな窓幅による話題境界と，より小さな窓幅による話題境界を統合することで，話題の階層的構成を認定することができる（詳細は文献 10) 参照）．

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た．そして，それぞれの窓幅により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，ひと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 1 は，第 1 の実験対象文書（水島議員の質問）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る．矩形が上述の話題境界候補区間を，話題境界候補区間を貫く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．

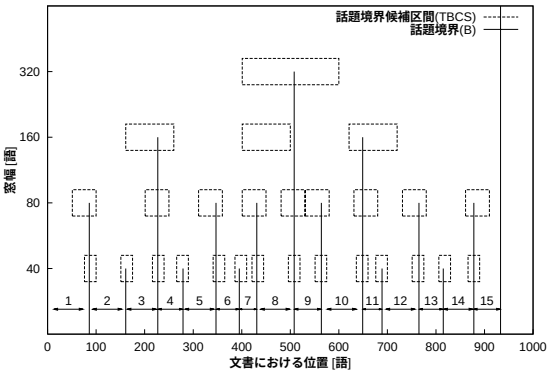

図 1 話題境界の認定結果
Fig. 1 Thematic boundaries of an interpellation.

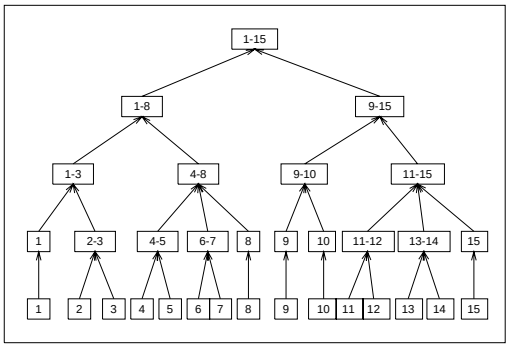

図 2 話題階層の認定結果
Fig. 2 Thematic hierarchy of an interpellation.

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 1 つの話題区画とする話題階層を作成する．たとえば，水島議員の質問文書に対しては，最小窓幅による境界で区切られた 15 の区画（図の矢印）に対応して，15 の話題が最下層の話題として認定された．また，80 語の窓幅では，最下層の 15 の話題のうち，区画 2 と 3，4 と 5，6 と 7，11 と 12，13 と 14 の 5 組の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た．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[11]によって行った．

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，矩形で表されたノードは，認定されたそれぞれの話題である．ノード内の数字は，図 1 中の区画番号に対応する．

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，

☆ 動詞・名詞・形容詞のいずれか．詳細については 3.1.2 項で説明する．

表 1 話題階層の認定結果の概要
Table 1 Summary of detected thematic hierarchies.

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)

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 階層下るごとに区画数が約 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 の等比級数により拡張したことに対応する．また，表 1 の区画数，および，次章の表 2 に示した文書中の平均語数において，質問文書と答弁文書の値の比を比較すると，両者ともほぼ 3 : 2（505 語 対 322 語，1,643 語 対 1,114 語）となっている．これは，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，話題境界認定手法により，窓幅にほぼ比例した（窓幅の 1/2〜2 倍程度の）区画が認定されたことが分かる☆．

2.2 関連話題の抽出

関連話題の検出手順を**図 3** に示す．

まず，ステップ 1 で，第 1 の話題階層中の任意の話題と第 2 の話題階層中の任意の話題からなる話題対のすべてについて，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．

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

☆ 入力文書を窓幅にほぼ比例した大きさの区画に分割するという性質は，話題構成認定手法の特性であり，調査報告書・新聞記事などを対象とした場合でも同様の結果となる[10]．

入力 2 つの話題階層（比較対象文書対）：$TH1, TH2$.
（話題階層に含まれるノードの集合を $node(TH1)$ のように表記）
出力 関連話題対の集合：T.
(1) $t1 \in node(TH1), t2 \in node(TH2)$ 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) $t1 \in node(TH1)$ 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．ここで最大関連度とは，$t1$ 以下の部分木に含まれるいずれかのノード $\exists u1$ と $\exists t2 \in node(TH2)$ について，(1) で計算された関連度 $R(u1,t2)$ の最大値のことである．
(3) 同様に，$t2 \in node(TH2)$ 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
Fig. 3 Algorithm for related passage extraction.

ここで

$tf_{t,s}$ 単語 t の区画 s における出現頻度
$|D|$ 原文を固定幅（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．通常の $tf \times idf$ 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 計算用の文書集合を用意する必要がない）．
- 式 (2) は，先に実施した見出し語抽出実験[9]によれば，単語の重要度の良い尺度である☆☆☆．
- 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．

☆☆ 最小窓幅による結束度の計算において考慮される語の繰返し間隔の最大値（最小窓幅の 2 倍）という意味合い．予備実験[9]で，比較的成績の良かった代表的ブロック幅でもあり，今回の実験ではこの値を用いた．
☆☆☆ この実験の際，形式段落単位に idf を計算する手法も試したが，固定ブロックによる場合より良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[12]によれば，見出しを話題境界として認定する確率が高い（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．

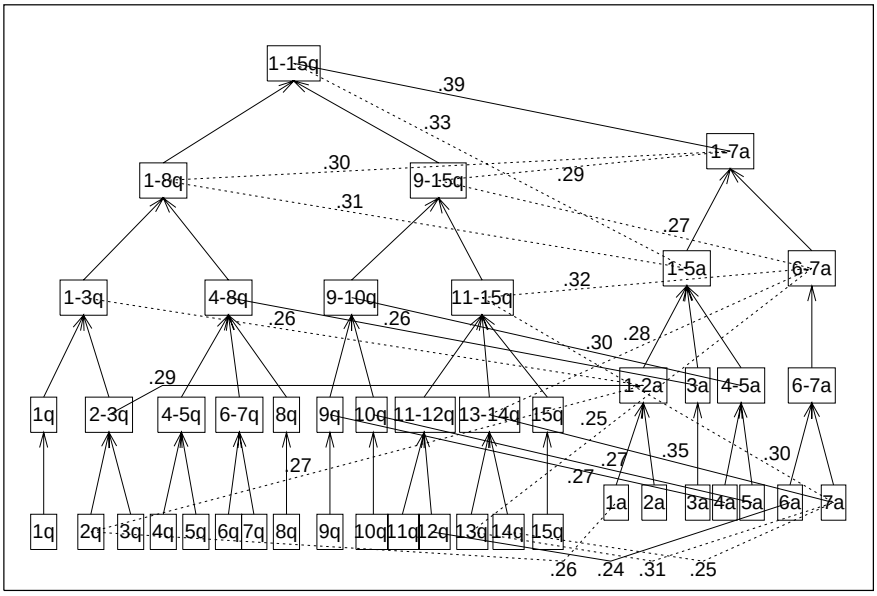

図 4 関連話題抽出例
Fig. 4 Example of related passage extraction.

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 $t2$ のすべてに対して，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．

具体的には，ステップ 2 で，第 1 の比較文書中の話題 $t1$ について最大関連度を求めて $t1.max$ 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，

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

なる話題対の集合 T を求め，関連話題として出力する．

2.2.1 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ている．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（質問文書）に，右のグラフが第 2 の比較文書（答弁文書）に対応する．

この図が示すように，上記手法で求めた関連度には，上層の話題間ほど高めになる傾向がみられる．このことは，画一的な閾値は話題対の選別に有効でないことを示唆している．上記のアルゴリズムの狙いは，話題階層の位置に応じて話題対選別の閾値を調整することで，雑多な「話題」が混在した上層の話題ノードからなる話題対ばかりが選択されるのを避けることにある．

ここで，右のグラフの右下角のノード 7a に着目する．このノードは，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付けられたアークの関連度の最大値である．ノード 7a では，（ノード 13-14q，ノード 7a）の話題対の関連度（0.35）が関連度の最大値となる．そして，ノード 13-14q 以下の部分木には，この値を超える関連度を持つアークがないので，（ノード 13-14q，ノード右 7a）の話題対が関連話題として出力される．

一方，ノード 6-7a に着目すると，このノード以下の部分木にノード 7a が含まれているので，ノード 6-7a に直結しているアーク（話題対）は，少なくともノード 7a に対する最大関連度（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 6-7a には，このようなアークはなく，よって，ノード 6-7a 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，文書全体どうしの話題対を除けば，7 対のみであるにもかかわらず，関連箇所が検出できなかった話題（いずれの関連話題話題対の構成要素ともなっていない話題）は，ノード 1q，ノード 11q，ノード 15q のみとなっている．これらの話題のうち，質問項目を含んでいたのは，ノード 15q 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，後に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．

3. 提案手法の評価

3.1 実験条件

3.1.1 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☆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．

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，質問点を復唱する文（「· · · 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，復唱文のパターンを念頭に目視すれば，大部分の話題（個々の質問点に対する答弁）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．

表 2 実験文書の構成
Table 2 Test documents for evaluation.

種類	字	語	文	段落
質問文書 (平均)	6,443	1,643	115	51
答弁文書 (平均)	4,067	1,114	51	37

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．たとえば，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った．

3.1.2 単語認定

今回の実験における単語認定には，日本語形態素解析ツール jmor [14]を用い，内容語（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切り出す機能もあるが，この機能は用いず，個々の名詞を別々の語として扱った．たとえば，水島議員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出された.【】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見なした☆☆.

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.

3.2 最小区画の認定精度の評価

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に関し，3.1.1 項で述べた答弁文書の性質に基づき，答弁文書の話題境界認定精度を評価した結果を**表 3** 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

☆ 国会図書館のホームページよりリンクのある国会会議録の検索ページ 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．

☆☆ 「い/る」は “要る”，“居る” のいずれの意味でも同一の語と見なすことになる．また,「い/る」と「要/る」のように表記が違う語は，たとえ意味が同じでも別の語と見なした．

表 3 答弁文書の話題境界認定精度
Table 3 Accuracy of thematic boundaries detected for answer documents.

正解の種類	境界数 認定	境界数 正解	再現率	適合率
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

正解境界として，再現率（$\frac{\text{一致境界数}}{\text{正解境界数}}$），適合率（$\frac{\text{一致境界数}}{\text{認定境界数}}$）を求めた．

表中,「小話題」とは，復唱文を正解境界とした場合の精度である☆．また,「大話題」とは，連続して出現する関連答弁（同じ質問項目に含まれる複数の質問点に対する答弁）を 1 つにまとめ，最初の復唱文のみを正解境界として集計した精度である．たとえば,「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始まり，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☆☆が,「大話題」で 1 つの話題としてまとめた部分の典型である．再現率・適合率に添えた括弧内の値は，すべての文境界から，認定境界と同数の境界をランダムに選択した場合に相当する基準値である．すなわち，再現率の基準値は，$\frac{\text{認定境界数}}{\text{文境界数}}$ によって，適合率の基準値は，$\frac{\text{正解境界数}}{\text{文境界数}}$ によって求めた値である．

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．

小話題に関する再現率約 6 割，適合率 7 割弱という値は，Hearst が英文の説明文書に対する実験で報告している値[4]と同程度の値であり，語彙的結束性に基づく話題境界認定としては標準的な精度が得られていると考えられる．また,「大話題」に 8 割弱というより高い再現率が得られていることは，大きな話題の切れ目ほど優先的に検出されていることを示唆していると考えられる．

3.3 関連話題抽出精度の評価

本節では，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する．計 91 対（文書全体どうしの話題対は除外）抽出された関連話題すべてについて，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みられる，関係のない話題に関する箇所を含むものもみられた．

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」と呼ぶ）を以下の基準により評価した☆☆☆．

答弁箇所に対する判定基準

- 抽出された答弁箇所に，一連の答弁内容のすべてが含まれ，かつ，関連のない別の質問項目に対する答弁内容が含まれていない場合,「完全」と判定する．
- 抽出された答弁箇所に，復唱文以外の答弁内容が含まれていれば,「許容」と判定する．ただし，まったく関連のない質問項目に対する答弁内容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．
- そうでない場合（典型的には，復唱文しか答弁内容が含まれていないか，雑多な質問項目に対する答弁が混在している場合）には,「不良」と判定する．

質問箇所に対する判定基準

- 答弁箇所に含まれるすべ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）すべて含まれていれば,「完全」とする．
- 答弁箇所に含まれ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．ただし，答弁箇所に関連しない質問項目が，質問箇所の大半を占めている場合には,「不良」と判定する．
- その他の場合（典型的には，質問点を含まない背景説明や答弁要求のみしか質問箇所に含まれていない場合）には,「不良」とする．

総合判定基準

- 答弁箇所と質問箇所の両者が「完全」の場合，

☆ 複数の復唱文がある場合には，先頭の復唱文のみが正解．
☆☆ この例は，羽田議員の質問に対する橋本首相の答弁の一部．
☆☆☆ 切り出し精度の判定は筆者 1 名で行った.「関連しない質問項目か」などの判定が難しい場合があり，判定結果には，必ずしも客観的といえない部分も残っている（付録 A.2 参照）．

表 4 質問–答弁対応箇所切り出し精度
Table 4 Accuracy of extracted question-answer passage pairs.

文書 (質問議員)	区画数 質問	区画数 答弁	話題対 抽出数 †	判定別話題対数 (構成比) 完全	許容	不良
羽田	39	35	17 (1)	1 (6%)	13 (76%)	3 (18%)
加藤 (紘)	31	12	5 (1)	1 (20%)	3 (60%)	1 (20%)
小沢 (辰)	32	15	11 (0)	2 (18%)	7 (64%)	2 (18%)
1998 年 2 月 18 日小計	102	62	33 (2)	4 (12%)	23 (70%)	6 (18%)
鳩山	30	20	10 (1)	2 (20%)	6 (60%)	2 (20%)
水島	15	7	7 (1)	5 (71%)	2 (29%)	0 (0%)
山岡	26	13	8 (1)	1 (13%)	6 (75%)	1 (13%)
不破	27	14	8 (0)	4 (50%)	4 (50%)	0 (0%)
土井	14	11	9 (1)	1 (11%)	6 (67%)	2 (22%)
小里	22	19	8 (0)	4 (50%)	3 (38%)	1 (13%)
神崎	12	15	6 (1)	1 (17%)	1 (17%)	4 (66%)
野田	10	3	2 (1)	0 (0%)	1 (50%)	1 (50%)
2000 年 8 月 1 日小計	156	102	58 (6)	18 (31%)	29 (50%)	11 (19%)
合計	258	164	91 (8)	22 (24%)	52 (57%)	17 (18%)

† 文書全体どうしからなる話題対は括弧内に分離して集計.

「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.

表 4 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．残り 2 割弱の「不良」と判定された話題対も，前述のように，1 文ずつは正しく対応する文を含んでおり，質問・答弁箇所のいずれかにおいて，関連の薄い質問項目が混在していたことが「不良」判定の原因である．すなわち，関連話題対の抽出精度をこれ以上にあげるには，話題境界の認定精度の向上が必要と考えられる．

3.4 抽出話題の網羅性と簡潔性の評価

本節では，提案手法の関連話題抽出部に関し，文書対に含まれる関連箇所を漏れなくかつ簡潔に抽出できるかという点について評価する．提案手法の趣旨は，前段の話題階層認定処理において，それぞれの文書に含まれる話題を漏れなく抽出し，後段の関連話題抽出処理において，文書対に含まれる関連話題を漏れなくかつなるべく少ない数で抽出することにある．そこで，関連話題抽出処理の評価としては，まず，答弁文書が，すべての話題（質問項目）を漏れなく，かつ，無駄なく含むという性質を利用して，固定閾値を用いて抽出する場合に比べて，抽出話題の網羅性と簡潔性が向上しているかという観点からの評価を行った．また，主要な話題の網羅性に関して，新聞掲載の要旨を利用して，要旨に取り上げられた主要な話題が，提案手法による関連箇所の切り出し結果から読み取れるかの評価を行った．

表 5 質問–答弁対の関連度分布
Table 5 Histgram of relevance scores of question-answer passage pairs

関連度	候補話題対 個数	候補話題対 答弁区画	抽出話題対 個数	抽出話題対 答弁区画
.6–.7	3 (0)	1.3	2(0)	1.5
.5–.6	13 (4)	4.5	6(1)	1.8
.4–.5	96 (40)	4.8	23(5)	2.2
.3–.4	256	5.3	25(7)	1.7
.2–.3	628	5.2	30(4)	1.3
.1–.2	1,604	5.1	5(0)	1.0
.0–.1	8,774	2.1	0(0)	0.0
合計	11,374	2.8	91(17)	1.7

() 内は不良話題対の内訳数

3.4.1 固定の閾値により抽出した場合との比較

表 5 は，前記の 11 対の質問・答弁文書における全話題対（候補話題対）から，上記手順により抽出した関連話題対（抽出話題対）の関連度別の分布を示している．表中,「個数」の欄の括弧内の値は，3.3 節の基準で「不良」と判定された話題対の個数である．ここで，候補話題対については，提案手法により抽出された関連話題対の総数以上の候補を含む関連度 0.4 以上の部分（表の区切り線より上）のみ，評価・集計した．また，「答弁区画」とは，各話題対によって対応づけられた答弁箇所に含まれる最小区画の個数の平均値である．

抽出話題対の関連度は，0.1〜0.7 と広く分布している．また,「不良」と判定された個数は，同数程度の話題対を関連度の大きい順に抽出した場合より，明らかに少ない．このことは，話題階層における話題区画の親子関係を利用して話題対を選別することが，固定の閾値で適切な話題対を選別するより，有効であることを示唆している．

表 6 抽出話題対による答弁区画の被覆率と重複出現率
Table 6 Coverage and redundancy rate of extracted topic passage pair concerning answer passages.

抽出法	抽出数	答弁区画 被覆率	抽出答弁区画 重複出現率	抽出答弁区画 平均出現数
提案手法	91	73%	27%	1.3
(許容以上)	(74)	(60%)	(15%)	(1.2)
上位 91 対	91	77%	65%	3.6
(許容以上)	(52)	(31%)	(53%)	(1.9)
.4 以上	112	86%	61%	3.7
(許容以上)	(68)	(33%)	(59%)	(2.4)

表 6 は，抽出話題の網羅性と簡潔性（冗長性）に関し，抽出話題対による答弁区画の被覆率と重複出現率を示している．表中,「被覆率」とは，それぞれの手法で抽出した話題対によって対応づけられた答弁箇所が，全体として，答弁文書中の最小区画をどのくらいカバーしているかを求めた値である．また,「重複出現率」と「平均出現数」は，最小の答弁区画が，いずれかの話題対によって対応づけられた答弁箇所に含まれることを「出現」とし，出現最小区画に対して，別の話題対によって対応づけられた答弁箇所にも含まれている区画の比率（重複出現最小区画/出現最小区画）と，平均出現数をそれぞれ求めたものである．たとえば，提案手法では，91 対の抽出話題対によって，164 個の最小答弁区画のうち 120 区画（異なり数）が，いずれかの答弁箇所の部分として抽出されていたので，被覆率は，73%（120/164）となっている．また，出現最小区画のうち 32 区画が，複数の話題対により答弁箇所の部分として重複抽出されていたので，重複出現率は，27%（32/120），最小区画の延べ出現数は，152 回であったので，平均出現数は，1.3（152/120）となっている．

表では，これらの値を，提案手法で抽出した場合と，関連度の大きい順に同数の話題対（91 対）を抽出した場合，および，固定の閾値（0.4 以上）で抽出した場合について示した．また，括弧内の値は，3.3 節の基準で,「完全」「許容」と判定された話題対（「不良」以外）に対して，上記の値を集計したものである．

許容以上と判定された話題対の比率（括弧内の値とその上の値の比率）についてみると，許容提案手法では，8 割強（74/91）が許容以上と判定されたのに対し，その他の手法では，6 割前後（52/91，68/112）にとどまっている．そのため，答弁区画の被覆率は，不良話題対も含めると，関連度順・固定閾値で抽出した方が高いが，許容以上の話題対では，提案手法が最も高くなっている．

この結果は，表 6 の「平均出現数」，また，表 5 の「答弁区画」（1 話題対あたりの最小答弁区画の数）が示しているように，候補話題対には必要以上に上層の（大きな）話題区画に関する関係が含まれていることに由来する．すなわち，関連度順・固定閾値で抽出した場合，話題階層の上層に対応する大きな答弁箇所が，提案手法による場合より多く，また，繰り返し抽出される傾向がみられた．そして，そのような答弁箇所には，関連の薄い質問項目に対する答弁の寄せ集めになっているものが多くみられ，結果として，不良と判定された話題対が多くなっている．また，質問文書に関しても，同様に，必要以上に大きな質問箇所が繰り返し切り出される傾向がみられ，結果として，対応づけられた答弁箇所に含まれる答弁とは無関係の質問項目の寄せ集めになっているという理由により，不良と判定された話題対が多くなっている．

最小区画の重複出現率についてみると，提案手法の値が，関連度順・固定閾値で抽出した場合の値より明らかに小さい．また，抽出話題対全体と許容以上の話題対とに対する値とを比較すると，許容以上の話題対の方が重複出現率が低くなっていることが分かる．

この結果をもたらした主な原因は，関連の薄い質問項目群を含む大きな質問・答弁区画に対応する不適切な話題対の混在である．つまり，いずれの抽出手法でも，このような不良話題対が抽出されており，そのため，必要以上に最小区画の重複出現率が高くなっている．それだけでなく，関連度順・固定閾値で抽出した場合には，関連の薄い質問項目群を含む不適切に大きな質問・答弁区画が何度も繰り返し抽出される傾向がみられた．そのため，最小区画の平均出現数が，提案手法に比べ大きく，かつ，不良話題対を除去した場合の減少率が大きくなっている．

たとえば，上位 91 対を抽出した場合には，不良話題対を除去することで，平均出現数が 3.6 から 1.9 と半分近くまで減少しているのに対して，提案手法の場合には，1.3 から 1.2 と減少率が比較的小さい．これは，提案手法が同一の話題区画を 2 度以上抽出することを明に抑制していることに由来する．すなわち，提案手法で，同一の最小区画が重複して抽出されるのは，包含関係にある複数の話題区画のそれぞれが，別の話題区画との対の形で抽出される場合に限られるのに対し，関連度順・固定閾値で抽出した場合には，同一の話題区画そのものが複数の話題対の構成要素として繰り返し抽出されることがある．よって，後者の手法では，たとえば，同じ答弁区画に対して，包含関係にある複数の質問区画との関係が，別の話題対として抽出されることがあり，結果として許容以上の話題対に関

○水島広子君 [9-10q]	○内閣総理大臣（森喜朗君）[4-5a]
¶^{9q} 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．子供たちは**大人**のまねをして成長します．**大人社会**のモラルがこれほど低下した今の日本で，子供たちのモラルだけが高まったら，むしろおかしなことだと思います．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．子供たちを批判する前に，総理**御**自身も含めて，私たち**大人**がまず反省すべきではないでしょうか．	¶^{4a} **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．青少年を取り巻く**社会**環境の改善のため，**社会**が一体となった取り組みを進めることが極めて重要であると考えております．
¶^{10q} 子供たちの問題行動とメディアによる**有害情報**の関係を指摘する専門家はたくさんいます．仮に犯罪に直結しなくても，幼いころから**有害情報**に当たり前のように触れることが子供たちの精神面の発育に及ぼす影響は無視できません．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．これはもちろん，国家による検閲というような形をとるべきではありません．例えば，子供にとって**有害**な**情報**であるか否かを親が判断して選べるようなシステム，また，町中でも子供が**有害情報**に触れるのを防ぐような社会的なバリアをつくるなど，地域社会の大人たちが子供たちを守るようなシステムをつくるべきだと思います．子供を**有害情報**から守るための立法の必要性について，森総理はいかがお考えでしょうか．	¶^{5a} 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．私は，かねてから，少年非行対策は与野党対立案件にあらずと考えておりますが，御指摘の点については，まさに議員と意見を一にするものであります．しかしながら，この種**法律**の制定につきましては，青少年をめぐる環境の浄化の基本的なあり方や表現の自由とのかかわりなど，国民的な合意の形成が必要であると考えられ，関係方面の幅広い議論を重ねていきたいと考えております．

図 5 重複して抽出された関連箇所の例
Fig. 5 Example of related passages extracted twice.

しても重複出現率が高くなっている☆．

これらの実験結果は，提案手法は，少なくとも関連度順・固定閾値で抽出する場合より，冗長な話題対の抽出を抑制できることを示唆している．また，関連度 0.4 以上の話題対において「完全」と判定されたものを調べたところ，全 10 対のうち 8 対は，提案手法で抽出されていた．すなわち，包含関係にある複数の話題区画の対応づけ候補の中から，いずれか 1 つを選ぶという操作に関し，提案手法による関連話題抽出処理は，8 割の精度で正しい話題区画を選別できたことになる．

図 5 は，提案手法で最小区画が重複抽出される場合の典型例を示している．ここでは，図 4 で話題対の構成要素として二重に抽出された，ノード 9q，10q および，ノード 4a，5a の部分の内容を出力している．図中，太字は，関連箇所抽出における主要な手がかりとなったと考えられる語である．また，下線を付与した語は，話題階層認定において，ノード 9q と 10q，および，ノード 4a と 5a が，上位階層において 1 つの話題として認定された主要な要因と考えられる語である．具体的には，対応する質問–答弁箇所（あるいは，質問区画間，答弁区画間）をまたがり出現し，かつ，式 (2) で得られる値が大きい語群☆☆である．

この図から読み取れるように，ノード 9q に対する答弁はノード 4a であり，ノード 10q に対する答弁はノード 5a である（太字の語を参照）が，ノード 9q とノード 10q との間，および，ノード 4a とノード 5a との間にも，強い関連性が読み取れる（下線の語を参照）．すなわち，ノード 9q は，マスメディアによる子供への悪影響を指摘した部分であり，ノード 10q は，それをふまえて有害情報から子供を守る法律の策定を提案した部分である．ノード 4a は，ノード 9q の発言に対して，同感である旨を表明した首相の答弁に，ノード 5a は，ノード 10q に対して，有害情報に関する法律制定へ向けた政府の取組みを説明した答弁である．このように，（ノード 9-10q，ノード 4-5a）は，質問・答弁文書における大きな話題の間の関係に対応し，（ノード 9q，ノード 4a）と（ノード 10q，ノード 5a）は，より小さい話題の間の関係に対応しているので，これらの 3 つの話題対は，冗長ではない．逆に，このように同一の箇所からも，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 1 つの利点とも考えられる．

以上の結果は，提案手法が，少なくとも，関連度順・固定閾値で抽出する場合に比べ，冗長性が少なく，かつ，正確な話題対を抽出できるという性質を持つことを示唆している．また，話題の網羅性に関しても，答弁区画の被覆率（許容以上の場合）が，関連度順・固定閾値で同数程度の話題対を抽出する場合より高くなっていることから，提案手法は，固定閾値で抽出するより効率的に関連話題を抽出できるといえる．

☆ 3.3 節の判定基準は関連のない質問項目の記述箇所が混在しても，その量が少なければ許容と判定していることに注意．
☆☆ 直観的には，関連度の半分程度をもたらしていそうな語群に相当．具体的手順は，付録 A.1 を参照．

3.4.2 新聞掲載の要旨との比較

上記の評価は，提案手法で抽出した結果が正しいかという観点から評価したものであるため，本研究の目的において抽出すべき主要な話題が抽出結果から読み取れるのかという点が明らかではない．利用者の視点から考えると，個々の細かい話題が抽出できることより，大きくとらえた共通の話題，すなわち，関連のある話題はまとめて抽出できることの方にメリットがある場合が考えられる．たとえば，いずれの比較文書の内容もまだ把握していない段階では，最初に分析すべき箇所の見当をつける意味で，後者の機能のニーズが高いと考えられる☆．

そこで，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．比較対象とした要旨は，新聞記者が，会議録中の発言を部分的に抜粋し，一問一答形式にまとめ直したとみられるものであり，記者が報道すべき内容として選択したという点で，客観性のある重要話題の例と考えられる．

表 7 は，前記の 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」「完全」とした対）に対応する質問箇所・答弁箇所を比較した結果である．表中,「完全」は，記事中の一問一等形式の要約に抜粋された発言の部分をすべて含む話題対が存在した場合に対応し,「一部（対）」は，抜粋された質問・答弁の部分双方を部分的に含む話題対が存在した場合に対応する．また,「一部（孤立）」は，質問あるいは答弁のどちらかのみ，抜粋された発言内容の一部/全部を含む話題対が存在した場合に対応する．

表によれば，全 55 組の一問一答形式の要約のうち，要約に抜粋された発言の部分が一部ずつでも質問箇所・答弁箇所に含まれていたものは 33 組（60%）であり，うち，要約に抜粋された発言部がすべて含まれていたものは，23 組（42%）である．60%という数値は，前節の評価により得られた許容以上の話題対に関する網羅性（答弁区画の被覆率）と一致しており，このことは，提案手法は，話題階層を利用することにより，主要な話題についても，3.4.1 項，表 6 の評価結果（「許容」以上の被服率）と同程度の精度で抽出できることを示唆している．

☆ この考え方は，本研究において，文書に固有の話題構成を重視した 1 つの理由である．利用者の視点などにあわせて動的にパッセージを切り出すこと（たとえば，文献 5)，7)）も考えられるが，本研究では，比較文書の部分間の関係を，文書固有の話題階層上の関係として，なるべく簡潔に表現することを目標としている．

表 7 日本経済新聞掲載の要旨との比較による評価
Table 7 Evaluation of the extracted passage pairs based on the gist of interpellations in articels of Nihon Keizai Shinbun.

文書 (質問議員)	掲載数	判定別掲載要約数 (構成比) 完全	一部 (対)	一部 (孤立)
羽田	10	1 (10%)	4 (40%)	1 (10%)
加藤 (紘)	6	3 (50%)	0 (0%)	1 (17%)
小沢 (辰)	10	5 (50%)	2 (20%)	2 (20%)
小計	26	9 (35%)	6 (27%)	4 (15%)
鳩山	4	2 (50%)	1 (25%)	2 (10%)
水島	1	1(100%)	0 (0%)	0 (0%)
山岡	3	2 (67%)	0 (67%)	0 (0%)
不破	4	2 (50%)	0 (0%)	1 (25%)
土井	4	3 (75%)	1 (25%)	0 (0%)
小里	7	3 (43%)	1 (14%)	1 (14%)
神崎	4	1 (25%)	1 (25%)	0 (0%)
野田	2	0 (0%)	0 (0%)	1 (50%)
小計	29	14 (48%)	4 (14%)	5 (17%)
合計	55	23 (42%)	10 (18%)	9 (16%)

表 8 部分的対応がみられる文書対に対する関連箇所認定精度
Table 8 Accuracy of related passage extraction across partially related documents.

比較対象	抽出箇所	正解	準正解
演説	75	51 (68%)	13 (17%)
別演説	57	31 (54%)	7 (12%)

3.5 弱い対応関係を持つ文書対を対象とする場合の抽出精度

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，簡単な補足実験を行った．この実験では，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．抽出した関連箇所の妥当性の評価結果を**表 8** に示す．

表中,「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．たとえば，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，関連に疑問が残ったものを「準正解」として分離し，残りを「正解」とした．完全に対応がないと判定した抽出箇所は，**図 6** の例のように，あまり重要でない語（この場合は「経験」1 語）の出現により関係づけられたものなどである．関連に疑問が残ったものの例としては，冒頭の挨拶どうしが共通の

話題（「有珠山（の噴火）」など）を含むために関係づけられたもの,「別演説」との比較において「与党三党」「補正予算」など別の実体を指す語によって関係づけられたものなどがある．

準正解を含めれば，対応のある演説と比較する場合で 85%，対応のない演説と比較する場合でも 66%の抽出箇所に，意味のある対応関係が見いだされた．正解と判定した例として**図 7** に森首相の所信方針演説と水島議員・土井議員の代表質問から抽出された関連箇所の要約を示す．これは，森首相の演説中の区画 22–24 に関して抽出された 2 つの関連箇所対を，付録 A.1 の手順で要約したものである．太字は，図 5 と同様に，関連話題対の抽出の主要な手がかりとなったと考えられる語群である．

○内閣総理大臣（森喜朗君）[20]	**○水島広子君 [11]**
七十歳まで働くことを選べる社会の実現に向けて，意欲と能力のある高齢者や障害者の働く場を確保するための条件整備を図るとともに，住宅や交通，公共機関のバリアフリー化の促進，歩いて暮らせる街づくりの推進など，いつまでも元気で生きがいを持って暮らせる環境を整備してまいります．高齢者や障害者の方々も，その技能や**経験**を生かして，社会の重要な担い手として活躍いただける世の中にいたします．	私は，精神科医としての**経験**の中から，人生の質を決めるものはコミュニケーション能力であると思っています．自分の意見を言い，相手の意見を聞き，お互いに納得のできる結論まで話し合える能力，それが人生のあらゆる場面で必要とされる能力であり，相手への思いやりや社会のルールを学ぶために必要とされる能力であり，また，国際社会の中で日本人に特に欠けている能力でもあります．

図 6 不適切な関連箇所の抽出例
Fig. 6 Example of incorrect related passage pair.

4. 結　　論

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．語彙的結束性に基づき認定した話題階層を利用する提案手法の特徴は，1）様々な大きさの話題区画を体系的に比較している点，2）関連話題抽出用の閾値を話題階層における話題区画の親子関係に基づき自動設定している点にある．

話題区画の体系的認定に関しては，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[10]によって，文書を 1/2 程度の大きさに分割する大きな区画から，最小窓幅（今回は 40 語）程度の大きさの小さな区画まで，様々な大きさの話題区画を求めた．これは，それぞれの文書が固有に持つ話題のまとまりに相当する区画を，ほぼ 1/2 ずつ大きさを縮小する形で求めていることになる．すなわち，語彙的結束性の分

○内閣総理大臣（森喜朗君）[22-24]	**○水島広子君 [1-3]**	**○土井たか子君 [17-22]**
日本**新生**プランの第三の柱は，**教育**の**新生**，すなわち**教育**改革であります．悪質な**少年**犯罪の続発や不登校，学級**崩壊**などの深刻化は，まことに心痛むものがあります．･･･**命**を**大切**にし，**他人**を思いやる心，**奉仕**の精神，日本の文化，伝統を**尊重**し，国や地域を愛する気持ちをはぐくみ，二十一世紀の日本を支える**子供**たちが，創造性豊かな立派な**人間**として成長することこそが，心の豊かな美しい国家の礎と言えるのではないでしょうか．私は，かねてから体育，徳育，知育のバランスのとれた全人**教育**を**充実**するとともに，世界に通用する技術，能力を備えた人材を育成するため，世界トップレベルの**教育**水準の確保が**必要**であると考えてきました．阪神・淡路大震災やナホトカ号重油流出事故のとき，全国津々浦々から**若者**たちが集まり，献身的に**ボランティア活動**をしていた姿を見て，さすが日本の**若者**と感動したことを思い出します．･･･また，制定して半世紀となる**教育基本法**についても，抜本的に見直す**必要**があると考えております．**教育**改革国民**会議**においても，九月の中間報告に向けて，我が国の**教育**各般にわたり議論が行われているところであります．私は，**学校**の運営体制を整備するとともに，教師が，**人間**が**人間**を教えるというとうとい使命感に燃えて**教育**に携わることが何よりも**大切**であり，IT **教育**や中高一貫の**教育**の推進，大学九月入学の推進，教員や**学校**の評価システムの導入，**教育**委員会のあり方なども重要な**課題**であると考えております． (原文との字数比 90%)	･･･**少年**犯罪についても，加害者に対する更生システムを専門化し徹底すると同時に，被害者のケアを**充実**するといった**課題**に目を向けずに，**少年**法を改正することで安易に厳罰化を図ろうとするような政治の姿勢には大きな危惧を抱いております．総理は，所信表明演説の中で**教育**の**新生**について述べておられました．しかし，その具体的内容を見ると，余りにも形式的，表面的なことばかりに思え，今**子供**たちの**教育**の場に最も**必要**とされている視点が欠けているように思えてなりません．･･･いじめの問題を根本的に解決するには，**人間**の多様性を**尊重**して，自分も**他人**も**大切**にできる**子供**を育てる**教育**が不可欠です．･･･**教育基本法**の見直しにしても，本来，他者との触れ合いを通して自発的に育てるはずの**奉仕**の精神や道徳心といったものを法改正によって一方的に押しつけようとするのであれば，逆効果となり，取り返しがつかないことになると思います．･･･ (原文との字数比 36%)	･･･まさに全国民的な**課題**と言ってよいでしょう．･･･むしろ，**教育基本法**改正や**少年**法改正の論議に深く踏み込むことによって問題は放置されてしまうと思えてなりません．･･･**子供**たちの荒れるのも，クラスの**崩壊**も，暴力も，**子供**たちが**人間**として**尊重**されていないところから始まっているように私には思えます．自分自身の**命**と人生を何よりも**尊重**され，**大切**に思える**子供**は，**他人**の**命**と人生も**尊重**し，**大切**にするものです．･･･**奉仕活動**を**学校**で強制するなど，無意味どころか反発を呼んで逆効果であろうと私は思います．･･･**ボランティア**の強制など，言葉の矛盾であるばかりでなく，ようやく日本にも育ちつつある**若者**たちの本来の**ボランティア**精神をも押し殺してしまうことになるでしょう．･･･総理，あなたは所信表明で，**教育基本法**を抜本的に見直す**必要**を言われました．･･･**教育基本法**の改正を提起しようとしている**教育**改革国民**会議**についてもお尋ねいたします．･･･ (原文との字数比 24%)

図 7 3 文書に対する関連箇所の要約例
Fig. 7 Summary of related passages across three documents.

析において意味のある境界が存在する範囲で，ほぼ2分木の形の話題階層を求めていることになる．よって，最小窓幅（40語）以上の話題区画については，網羅的に比較対象に含められることになる．そして，1対の比較対象文書について，それぞれの文書で固有の意味を持つ話題区画の間の関連度を網羅的に求めているので，原理的には，文書間にまたがり現れる様々な粒度の関連話題を網羅的に抽出できると期待される．

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，要素的話題に相当する最小の答弁区画の約6割は，正しく抽出できた関連箇所の組でカバーされることが分かった．これらの値が，同数程度の関連箇所を固定の閾値によって抽出する場合より明らかに高くなっていることは，提案手法が，文書間の話題を網羅的かつ簡潔に抽出するうえで有効であることを示唆している．また，新聞に要旨として掲載された内容の約6割は，この手法で抽出された関連箇所の対から読み取れることも確かめられた．このように，少なくとも対応が保証された文書であれば，様々な粒度の話題に対する関連箇所を効率的に検出できる見込みが得られた．

今後の課題としては，1）関連箇所抽出手法をより広い範囲の関連文書に対して適用すること，2）関連話題抽出手法を使って関連文書の比較作業を支援する方法を確立することの2つがあげられる．

前者に関しては，今回実験対象とした文書対は，最も弱い対応を持つ文書対でもここ数年の政治課題に関する文書であるという意味で，比較的強い対応を持った文書対であると考えられ，もっと弱い対応関係しか持たない文書対に対しても提案手法が有効であるかを確かめることが課題となる．そのような文書の場合用語法の違いなどが想定されるので，シソーラスなどを利用して認定した関連語彙の連鎖を内容比較に利用する手法（たとえば，文献3)や7)）を応用することも，検討の余地がある．

後者に関しては，話題階層表示により関連箇所の構造上の対応関係を分かりやすく提示すること，関連箇所の内容を簡潔にまとめた要約を提示して，多くの関連箇所を一覧できるようにすることなどを検討中である．

謝辞 本研究を進めるにあたり，東京大学情報理工学系研究科の辻井潤一先生にご指導いただきました．また，(株）富士通研究所ドキュメント処理研究部の有志の方々には補足実験にご協力いただきました．ここに記して感謝いたします．

参 考 文 献

1) Allan, J.: Automatic Hypertext Construction, Ph.D. Thesis, Cornell University (1995).
2) Callan, J.P.: Passage-level Evidence in Document Retrieval, *Proc. 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '94*), pp.302–310, ACM (1994).
3) Green, S.J.: Automatically generating hypertext by computing semantic similarity, Technical Report 366, University of Tronto (1997).
4) Hearst, M.A.: Multi-paragraph segmentation of expository text, *Proc. 32nd Annual Meeting of the Association for Computational Linguistics*, pp.9–16, ACL (1994).
5) 黒橋禎夫，白木伸征，長尾 眞：出現密度分布を用いた語の重要説明箇所の特定，情報処理学会論文誌，Vol.38, No.4, pp.845–854 (1997).
6) Mani, I. and Bloedorn, E.: Summarizing Similarities and Differences among Related Document, *Advances in Automatic Text Summarization*, chapter 23, Mani, I. and Maybury, M.T.(Eds.), pp.357–379, The MIT Press, London (1999). (Reprint of *Information Processing and Management*, Vol.1, No.1, pp.1–23, 1999).
7) 望月 源：語彙的連鎖を用いたパッセージ抽出とその応用に関する研究，博士論文，北陸先端科学技術大学院大学 (1999).
8) 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，情報処理学会研究報告 NL-117-17 (1997).
9) 仲尾由雄：文書の話題構成に基づく重要語の抽出，情報処理学会研究報告 FI-50-1 (1998).
10) 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).
11) 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，情報処理学会研究報告 NL-132-7 (1999).
12) Nakao, Y.: An Algorithm for One-page Summarization of a Long Text based on Thematic Hierarchy Detection, *Proc. 38th Annual Meeting of the Association for Computational Linguistics*, pp.302–309, ACL (2000).
13) 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, *Proc. 2nd ACM Conference on Hypertext* (*Hypertext '89*), pp.319–341, ACM (1989).
14) 西野文人：日本語テキスト分類における特徴素抽出，情報処理学会研究報告 NL-112-14 (1996).
15) 大森信行，岡村 潤，森 辰則，中川裕志：情報

検索手法を利用した関連マニュアル群のハイパーテキスト化，情報処理学会論文誌，Vol.40, No.6, pp.2776–2784 (1999).
16) Salton, G., Singhal, A., Buckley, C. and Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes, *Proc. 7th ACM Conference on Hypertext* (*Hypertext '96*), pp.53–65, ACM (1996).
17) Singhal, A. and Mitra, M.: Pivoted Document Length Normalization, *Proc.19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '96*), pp.21–29, ACM (1996).

付　　録

A.1 要約手法

図 7 の作成に使った要約手法について説明する．**図 8** はその基本手順である．この手順において，文献 8) の手法というのは，与えられたキーワードをすべて含み，かつ，なるべく少ない文数からなる要約を作成するという趣旨で考案した手法である．必ずしも最小文数であることは保障されないが，直観的にいえば，すべてのキーワードが 1 回ずつ現れるような要約が作成される．たとえば，図 5 をこの手順で要約すると，**図 9** のようになる．両図において太字で強調表示した語は，ステップ 2b で抽出したキーワードである．図 9 の要約にみられるように，この手順を使うと，少なくとも 1 回は，与えたキーワードが出現する要約が作成できる．ステップ 2b で抽出したキーワードは，本稿においては，関連度計算において主要な役割を果たしている語という位置づけであり，関連話題抽出処理の動作を示唆する意味で，図 7 は，この手法による要約結果を示した．

なお，図 7 の作成においては，基本的にこの手順に従い，森首相の演説–水島議員の質問，森首相の演説–土井議員の質問の 2 つの文書対を要約したが，森首相の演説箇所については，水島議員の質問箇所，土井議員の質問箇所との比較により抽出された 2 セットのキーワードをマージして用いた．

A.2 質問–答弁対応関係の精度判定に関する一致率

3.3 節の表 4 に掲げた切り出し精度の判定は，筆者が 1 名で行ったので信頼性に疑問が残る部分がある．本節では，この点に関する補足実験の結果を紹介する．この補足実験では，5 名の被験者に精度判定対象話題対の 1 割程度を別に評価してもらい，被験者の判定が筆者の判定とどのくらい一致するかを調査した．

被験者は，筆者の職場で自然言語処理関連の研究や分析作業を行っている 20 代〜40 代の男女（男性 4 名，女性 1 名）であり，少なくとも 2 年間は，自然言語処理関連の研究・製品開発などに携わった経験を持つ者である．実験対象とした話題対は，鳩山議員の質問に関する質問–答弁対（10 対）であり，判定基準は 3.3 節のものとほぼ同じである．ただし，補足実験では，実験手順が煩雑になるのを避けるため，3.3 節の精度判定とは異なり，原文（切り出し箇所の前後の文脈）の参照を許さなかったため，以下の判定要素については，被験者の主観にまかせることとした．

- 一連の答弁内容のすべてが含まれているか．
- 背景，質問点（提案点），答弁要求の一群が（存在するかぎり）すべて含まれているか．

(1) 第 1 の抽出箇所と第 2 の抽出箇所に共通に含まれる語を求め，共通語とする．
(2) 第 1 の抽出箇所と第 2 の抽出箇所のそれぞれに対して以下の操作を行う．
　(a) 補正エントロピーの総和の計算：抽出箇所に含まれる共通語について，補正エントロピーの総和 H を求める．ここで，補正エントロピーとは，式 (2) で得られる値のことである．
　(b) キーワードの抽出：抽出箇所に含まれる共通語を，補正エントロピーの大きい順に取り出し，抽出キーワードの補正エントロピーの和が $H/2$ を超えない範囲で，キーワードとして抽出する．
(3) 第 1 の抽出箇所と第 2 の抽出箇所から抽出されたキーワードを合わせたものを核として，文献 8) の方法で要約を作成する．

図 8 関連談話要約アルゴリズム
Fig. 8 Algorithm for summarizing a related passage pair.

○水島広子君 [9-10q]	○内閣総理大臣（森喜朗君）[4-5a]
¶^{9q} 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 1 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．・・・	¶^{4a} **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．・・・
¶^{10q} ・・・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．・・・	¶^{5a} 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．・・・
(原文との字数比 34%)	(原文との字数比 38%)

図 9 関連談話の要約例（水島議員の質問と森首相の答弁）
Fig. 9 Summary of related passage pair (Mizushima's question and Prime Minister's answer).

表 9 質問–答弁対応箇所切り出し精度に関する正誤判定の一致状況
Table 9 Agreement among judgments on accuracy assessment of extracted question-answer passage pairs.

話題対番号	筆者判定	多数意見	判定内訳		
			完全	許容	不良
1	許容	許容	0	**4**	1
2	許容	許容	0	**3**	2
3	**許容**	**不良**	1	1	**3**
4	不良	不良	0	2	**3**
5	不良	不良	0	2	**3**
6	完全	完全	**3**	2	0
7	許容	許容	0	**3**	2
8	**許容**	**完全**	**4**	1	0
9	許容	許容	0	**4**	1
10	完全	完全	**3**	2	0
合計			11	24	15

「判定内訳」の下線は筆者判定に一致．太字は多数意見．
その他の下線・太字は筆者判定と多数意見の不一致を示す．

実験は，次の手順で，個々の被験者に別々に進めてもらった．

(1) 初めに，別に用意した例題を使って 5 分程度作業手順を説明し，質問を受け付ける．
(2) 制限時間は設けず，また，必要に応じて参照できるよう作業手順書を渡した状態で，自由に作業をしてもらう．
(3) 被験者は，答弁文書，質問文書，総合判定の 3 段階で，それぞれの判定結果を「完全」「許容」「不良」の 3 カテゴリの中から選択する形で，アンケート用紙に記入する．

表 9 は，この補足実験の結果である．この表から読み取れる主な結果，および，補足情報を以下に示す．

- 筆者の判定と個々の被験者の判定との一致率は，全体平均で 56%（28/50）である．
- 筆者の判定と多数意見とを比較した場合には，80%の一致率となる．
- 1 件の例外（話題対 3）を除き，被験者の判定の揺れは，「完全/許容」「許容/不良」の違いに収まる．例外となった話題対は，代表質問の冒頭で議員が質問の形をとらずに指摘した事項とその答弁からなるもので，質問の形をとっていないことをどう判定するかの基準を設けなかったことが「完全」から「不良」まですべての判定が混じったことの原因と考えられる．
- 筆者の判定が「完全」のものは，本来，被験者すべてが「完全」と判定した方が自然だと思われるが，そうはなっていない．この原因に関し，被験者の意見を事後に確かめたところ，原文の参照を許さなかったため，答弁内容が他にもあるのではないかなどの疑念が生じていたことが判明した．

全体として，多数意見とは 80%の一致をみており，不一致の話題対は，前述の話題対 3 と，質問冒頭の導入部 2 文のみが欠落していた話題対 8（被験者判定は「完全」）であったので，おおむね筆者の判定は妥当であったと考えられる．ただし，被験者の意見が完全に一致した話題対がないなど，一致率が低いので，評価用データを多人数で作成する場合には，判定の基準と手順をより詳細化し，例題を通じて意識を統一するなどの改善が必要であると考えられる．

(平成 13 年 4 月 7 日受付)
(平成 13 年 6 月 27 日採録)

(担当編集委員 大山 敬三)

仲尾 由雄（正会員）

1962 年生．1986 年東京大学理学部物理学科卒業．同年（株）富士通研究所入社．1988 年〜1995 年（株）日本電子化辞書研究所へ出向し，自然言語処理用電子化辞書の研究開発に従事．現在（株）富士通研究所．自然言語処理技術を使った文書処理システムの研究開発に従事．言語処理学会，ACL 各会員．