話題の継続に着目した国会会議録要約

川端正法, 山本和英
長岡技術科学大学　電気系
E-mail:{kawabata,ykaz}@nlp.nagaokaut.ac.jp

1　はじめに

近年、情報化社会の中で文書の電子化が進んでいる。情報過多の中でユーザーが目的の文書を探すためには大量のテキストに目を通さなければならない。短い文書であれば全てを読むことも可能であるが、長い文書になるとユーザーの負担は大きくなる。そこで内容を判断するための指示的な文書要約が必要になり、盛んに研究されている。電子化された文書の種類は技術文書や会議録など多岐に渡る。それらの文書全てに対応できる自動要約は文書の種類によって特徴が異なるため困難である。文書の種類を大別すると文語と口語がある。新聞のような文語の自動要約手法は既に多く提案されている。しかし口語を対象とした自動要約は少ない。我々は以前国会会議録を対象に、話し言葉に特徴的な表現を用いて換言を行う報知的な自動要約手法を提案した [1]。しかし報知的要約は内容を判断するための要約ではない。そこで本稿では報知的な要約ではなく内容を判断するための要約として国会会議録を対象に 1000 字以内に収める指示的な要約手法を提案する。

2　関連研究

まず、比較的長い文書の自動要約手法として仲尾 [2] がある。この研究は TextTiling[3] を用いて本の内容を話題ごとに分割する。そして、要約率に応じた話題のまとまりから導入部を探して要約文とし、一頁程度に要約するというものである。一般に本は文語で書かれており、その内容別に章などの単位でまとめられているため TextTiling による話題の分割が有効である。しかし、仲尾の手法では口語である国会会議録に対して分割した話題のまとまりが信頼できないという結果であった。

このように国会会議録のような口語で長い文書を対象にした指示的要約は提案されていない。

3　国会会議録とその要約

国会会議録は国会で行われた衆議院および参議院の本会議、委員会等を記録した口語の文書であり、多くの議論がやり取りされている。議論は議題に沿ったものだが、その多くは議題についての細かい発言である。よって要約として必要な部分はわずかである。そこで本稿ではこのわずかな部分を話題の導入部分と結論部分と考えた。従来から多く提案されている手法として、単語の出現頻度などを使った重要文抽出がある。しかし、従来法では話題に関係なく文が抽出され、同じような話題の文が抽出されることがある。本手法では会議録内の各話題の継続に着目し、その導入部分と結論部分を取り出すことで要約を行う。

4　語句の定義

本稿で使用する語句の定義を以下に示す。

4.1　話題の手がかり

国会会議録中で “N_1 の N_2” になるような助詞ノで繋がっている名詞句を手がかりに話題を探す。本稿で “N_1 の N_2” に着目した理由は単名詞や複合名詞だけでは話題の手がかりとして意味が広すぎると考えたためである。助詞ノには限定や所有の用法が含まれているため、より明確に話題の内容を表すことができる。本稿では “N_1 の N_2” の事を話題の手がかりと呼ぶ。

4.2　継続段落及び継続数

本稿では話題の手がかり (N_1 の N_2) が含まれる段落以降で同じ名詞句が含まれている段落を継続段落、その段落数を継続数と呼ぶ。これは話題の手がかりによって表される話題の始点と終点を決定する際に使用する。また本稿で使用する継続段落には「順方向継続段落」と「逆方向継続段落」の 2 つがあり、これらについては 5.3 節で計算方法を述べる。

4.3　導入段落と結論段落

話題の導入部分を「導入段落」、結論部分を「結論段落」呼ぶ。この導入段落と結論段落について述べる。国会会議録では議論の話題が変わるときには、必ず次の話題についての発言がある。例えば「次に、〜についてですが」のような発言である。これはこれから話す話題についての発言である。本稿では、これから話す話題についての発言が含まれる段落を「導入段落」と呼ぶ。

また話題の中には「〜については〜と考える」のような結論を述べている発言がある。本稿では話題の中で結論を述べている発言が含まれる段落を結論段落と呼ぶ。以下に国会会議録における導入段落と結論段落の例を示す。

例 1)【導入段落】本日は基本計画の変更ということに当たる委員会でございますが、基本認識をまずお尋ねしたいと思います。

例 2)【結論段落】我が国としましては、民主的で安定した国づくりに懸命に取り組むイラク政府の努力を、国際社会と連携しながら積極的に支援をしていくことが重要であると考えております。

5　提案手法

5.1　手法概要

本稿では国会会議録の段落中に含まれる話題の手がかりに対し、順方向及び逆方向の継続数をその話題の大きさとして付与する。その話題に対し、導入段落と結論段落を決定し、大きな話題の順にそれらを要約として出力する。大きな話題程国会会議録の中で主要な内容である。

ここでは 1 つの話題の手がかりに着目して行う。最終的には、文書に含まれる全ての話題の手がかりについて行い、各話題の段落継続数を用いて出力する話題を決定する。以下に本手法の流れを示す。また各 Step の詳細については次節以降で述べる。

Step 1 話題の手がかりを抽出
要約対象の国会会議録から話題の手がかり (N_1 の N_2) となるもの全てを抽出する。

Step 2 話題の同定
話題の手がかりに対し順方向の継続段落及び逆方向の継続段落を求め、同時に継続数を付与することで話題を同定する。

Step 3 導入段落候補の抽出
逆方向の継続数が 0 の話題の手がかりから、順方向の継続数が大きい順に導入段落候補を抽出する。

Step 4 結論段落候補の抽出
導入段落が含まれている話題の手がかりから、逆方向の継続段落数の大きい順に結論段落の候補を抽出する。

5.2 話題の手がかりの抽出

形態素解析器 MeCab(1) の結果を用いて、国会会議録に含まれる話題の手がかり (N_1 の N_2) をすべて抽出する。話題の手がかりに含まれる名詞は単名詞のみではなく複合名詞も含む。ここで複合名詞とするものは国会会議録中で連続した名詞のことである。複合名詞の場合は以下の例に示すように形態素で分割し保持する。以下 N_2 が含まれると言う表現は、N_2 のいずれかの名詞が含まれている場合を示す。

例 3)
イラクの治安状況
N_1=イラク, N_2={ 治安, 状況 }

ここで、“N_1 の N_2” となっても話題の手がかりとしない場合がある。この条件は大きく 2 つに分けられる。1 つは N_1,N_2 に関する条件で、もう 1 つは段落に関する条件である。これについて述べる。

まず、N_1,N_2 に関する条件は以下の 3 種である。

1. “N_1の N_2” の直後の 2 形態素目が動詞または名詞-サ変接続
2. N_1または N_2 に非自立の名詞 (「〜の中」,「〜のこと」) が含まれる
3. N_2 に名詞-形容動詞語幹 (「〜の着実」,「〜の必要」) となる語が含まれる

国会会議録を観察した結果、“N_1の N_2” の直後の 2 形態素目には動詞または名詞-サ変接続が出現しにくいことから条件 1 を導入した。名詞の非自立や、名詞の形容詞語幹は話題の手がかりとして相応しくないと考え、条件 2,3 を導入した。この理由は文章中に良く出現するために、どのような話題においても良く出てくるためである。

次に、段落に関する条件について述べる。

1. 段落が過去形で終わる

国会会議録を観察した結果、過去形で終わる導入段落の出現数は少なかったためこの条件を導入した。

5.3 継続段落数の算出

ここでは、継続段落数の算出方法について述べる。継続段落数の算出は話題の手がかり (N_1 の N_2) に含まれる名詞を用いて行う。ここで、1 種類の話題の手がかりに着目して説明する。

まず、話題の継続する条件は、N_1 と N_2 が両方含まれた段落とする。ここで N_2 が含まれるというのは 5.2 節で示したように、N_2 を構成する名詞が 1 つでも含まれていればよい。この段落を話題が継続している段落と呼ぶ。ここで継続していない段落が多く続くと話題が継続しているとは言えない。そこで、話題が継続していない段落が S 段落続いたときそれ以降の段落は話題が継続していないとする。本稿ではこの段落数を 20 とした。

この話題の継続は、話題の手がかりを起点に順方向、逆方向に対して行う。順方向に継続する全ての段落を「順方向継続段落」、逆方向に継続する全ての段落を「逆方向継続段落」と呼ぶ。

順方向継続段落の数を「順方向継続段落数」と呼び C_f と表す。逆方向継続段落の数を「逆方向継続段落数」と呼び C_b と表す。C_f と C_b の和を継続数と呼び C_n と表す。

1 種類の話題の手がかりを用いても、話題の手がかりは複数の段落に存在することがある。この全ての話題の手がかりについて継続段落を求める。

5.4 導入段落候補の抽出

ここでは導入段落候補の抽出を行う。これは導入段落らしさを定義し、これを用いて導入段落候補を抽出する。導入段落はこれから議論される話題についての導入部である。

導入候補らしさには以下の指標を用いる。

- 話題が長く継続している
- 話題がその段落以前で出現していない

この指標をから、導入段落候補は $C_n \geq 1$ かつ、$C_b = 0$ のものである。これは C_n が大きいものほど優先して扱う。

5.5 結論段落候補の抽出

ここでは、結論段落候補の抽出を行う。結論段落候補は導入段落候補と対となる。つまり、5.4 節で抽出した導入段落候補 1 段落に対して、結論段落候補も 1 段落存在する。

結論段落は以下の条件全てに当てはまる段落である。この段落を 5.4 節で求めた導入段落候補の対になる結論段落候補とする。

- 対象としている話題が継続している段落である
- “N の N” が含まれる段落
- この “N の N” の逆方向継続段落数が最も大きい段落

5.6 要約文の生成

ここでは抽出された導入段落候補と結論段落候補の組から要約文の生成を行う。導入段落候補に付与されている継続段落数 C_n が大きい程長く議論されている。よってこの継続段落数の大きいものから順に、導入段落候補と結論段落候補を出力する。この処理は 1000 文字を超えた時点で終了する。1000 文字を超えた段落は削除する。

6 評価実験

国会会議録を用いて提案手法の評価を行った。評価では提案手法によって、導入段落および結論段落の適合率および再現率を調べた。使用したデータを次に示す。

6.1 実験データ

実験データは国会会議録 (2) から無作為に選んだものに人手で導入段落と結論段落の正解をつけたものを使用した。データの大きさに幅があり国会会議録では会議によって、文書の長さは異なっている。評価で使用した導入段落の正解データの段落数は、147 段落から 748 段落である。また、文字数は 9 千字から 14 万字程度である。なお、今回使用した国会会議録は人手で議論のまとまりごとに分割してある。

6.2 導入段落の評価

本手法では 5.3 節で段落に含まれる N_1 の N_2 を手がかりに継続段落数を求めた。しかし、実際の導入段落には “N_1 の N_2” が含まれていないものも存在する。そこで、導入段落すべてに正解を付けたものと “N_1 の N_2” が含まれている導入段落だけに正解をつけたもの両方について精度を求めた。使用した国会会議録は無作為に選択した 7 セットである。表 1 に適合率および再現率を示す。表中の（手がかりのみ）は、話題の手がかりのみを正解とした場合の再現率である。

表 1: 導入段落の評価結果

会議録名	適合率	再現率	再現率 (手がかりのみ)
国会会議録 1	0.22	0.22	0.29
国会会議録 2	0.08	0.07	0.11
国会会議録 3	0.47	0.22	0.28
国会会議録 4	0.18	0.15	0.18
国会会議録 5	0.15	0.12	0.13
国会会議録 6	0.27	0.16	0.18
国会会議録 7	0.17	0.12	0.14
平均	0.22	0.15	0.19

表 1 より、導入段落の抽出精度の平均は 0.22 であった。すべての導入段落を正解とした場合の再現率の平均は 0.15 であった。また、話題の手がかり “N_1 の N_2” が含まれている導入段落とした場合の評価では再現率の平均は 0.19 であった。これは、すべての導入段落を正解とした場合よりも 0.04 上回った。

6.3 結論段落の評価

国会会議録よりランダムで選んだ 3 セットに導入段落の正解を人手で作成した。これを利用して結論段落の抽出精度を求めた。抽出精度は結論段落の 1 位のみが正解だったときと、3 位までに正解が含まれていた場合の 2 つの精度を求めた。結論段落の抽出精度を表 2 に示す。表中の括弧内は 1 位のみの精度なら 1 位に、3 位までの精度なら 3 位までに正解が含まれていた結論段落候補の数を示す。

表 2: 結論段落候補の精度

会議録 ID	導入段落候補数	1 位のみ精度	3 位までの精度
002	20	0.00(0)	0.20(4)
003	18	0.11(2)	0.50(9)
1009	48	0.15(7)	0.37(18)

7 考察

7.1 導入段落と結論段落の抽出

本手法における国会会議録を対象に導入段落および結論段落の抽出精度を評価した。その結果、22%の精度で導入段落を抽出することができた。。また、“N_1 の N_2” が含まれる導入段落のみを正解とした導入段落候補の再現率は 0.19 であった。例えば、評価に用いた “国会会議録 2” の導入段落の全正解数（手がかりのみ）は 9 段落であった。導入段落候補に含まれていた正解は 1 段落のみで、再現率 0.11 である。この会議録の話題の継続段落数を調べたときに C_n が正の値となる正解段落は 3 段落である。また、本稿では導入段落候補とする段落は逆順継続段落数が 0 段落であるもののみとしている。先程の 3 段落の正解段落のうち 2 段落は話題以前の継続段落数が 1 以上であるため最終的な導入段落候補は 1 段落のみとなる。このことから、導入段落候補の減少は次の 2 つの要因が考えられる。

表 3: 継続先段落の下限を変化させた場合の適合率

C_h 下限	正解	不正解	適合率	再現率	F 値
0	7	40	0.15	0.14	0.15
1	7	35	0.17	0.14	0.15
2	7	35	0.17	0.14	0.15
3	6	30	0.17	0.12	0.14
4	6	27	0.18	0.12	0.15
5	6	24	0.20	0.12	0.15
6	6	20	0.23	0.12	0.16
7	6	16	0.27	0.12	0.17
8	6	14	0.30	0.12	0.17
9	6	13	0.32	0.12	0.18
10	6	13	0.32	0.12	0.18
11	6	12	0.33	0.12	0.18
12	6	11	0.35	0.12	0.18
13	6	10	0.38	0.12	0.18
14	6	9	0.40	0.12	0.19
15	6	7	0.46	0.12	0.19
16	6	5	0.55	0.12	0.20
17	5	5	0.50	0.10	0.17
18	5	4	0.56	0.10	0.17
19	5	4	0.56	0.10	0.17
20	5	4	0.56	0.10	0.17

まず 1 つ目の要因は、導入段落であっても話題がその後の段落で共起しない場合である。本手法では、話題の名詞 N_1 が複合名詞であっても名詞ごとには分けずに 1 つの名詞として扱った。これは “N_1 の N_2” のような表現では助詞ノの後ろの名詞よりも前の名詞がその表現の中心であると考えたからである。しかし、問題は複合名詞に略語がある場合である。導入段落のような話題の始まりの発言では、略語はあまり略さずに使われる。そして一度発言されたものは、その後では略して発言されることが多い。例えば “航空自衛隊” などが N_1 であっても、その後の段落では “空自” と略される。そのため、その後の段落で話題が共起せず、継続段落数は低い値となったと考える。

2 つ目の要因としては導入段落の抽出時に C_b が 0 のものしか抽出しなかったことである。実際には C_b が 0 でない場合にも導入段落候補があるため、$C_b = 0$ という制約を与えたことで出力される導入段落候補は減少したと考える。この C_b の値はいくつが最適であるかは本稿では触れなかったが、C_b の値の調整は今後の課題である。

7.2 閾値の決定による精度の向上

本稿で提案した手法における導入段落の抽出は $C_b = 0$ であるものを全て導入段落候補としている。しかし出力された導入段落には多くの不正解が存在している。そこで 5.4 節で定義した導入段落らしさより、話題組のスコア C_n に制約を与えることで再現率は犠牲になるが適合率の向上が見込めることが分かった。ここでは C_b と C_n との差を C_h とし、この値を調整する。導入段落候補として出力する導入段落のスコア C_h の下限を 0 から 20 まで変化させたときの適合率および再現率を表 3 に示す。

表 3 より C_h の値が段落継続数以上の話題を導入段落候補として出力した場合、段落継続数が 0 段落のときは適合率は 0.15 である。段落継続数を変化させて 20 段落にした場合の適合率は 0.56 である。出力された候補中の正解数は継続段落数が 0〜20 段落に変化すると 2 個減少した。また、不正解数は 40 個から 4 個へと 36 個減少している。このことから段落に含まれる話題の継続段落数 C_h が高いほど導

入段落らしいといえる。

国会会議録の導入段落の正解データを大量に用意できれば、適合率および再現率から F 値を求めて F 値が最高となる継続段落数の閾値を決定することができる。閾値を決定することで、本手法で抽出した導入段落候補の精度よりも高い精度が期待できる。

7.3 導入段落らしさについて

上記で継続段落数を変化させた場合の導入段落候補の適合率および再現率を示している。表 3 より、継続段落数を上げたときの正解数の減少率は不正解数の減少率よりも緩やかに減少していることが分かる。これは本稿で定義した導入段落らしさが正しかったことを示している。

7.4 要約結果について

自動要約の手法として重要な語の頻度などで文にスコアを付与する重要文抽出がある。国会会議録は 8000 字を超える文書が多く存在する。このような文書を 1000 字以内で要約するには、本当に必要な部分を抜き出す必要がある。しかし、語の頻度による重要文抽出では頻度の高い語が含まれる文を文書中から所々抜き出すだけである。それを並べただけでは、内容を判断することは難しいと考える。

本手法では、話題の継続によって導入段落と結論段落を要約として抽出する手法を提案した。国会会議録中の導入段落は、それ自体がこれから議論する内容の要約といえる。結論段落もこれまでに議論された話題についての要約と考えることができる。このように、導入段落および結論段落をを抽出することで、所々から文を抽出して並べるよりも、内容を判断しやすいのではないかと考える。付録に実際に出力した要約結果を示す。

8 おわりに

国会会議録の要約として相応しいのは会議録中で多く議論されている内容であると考えた。そこで話題の継続に着目して導入段落と結論段落を抽出することで指示的な要約を行った。本稿では可変の要約率に対応できる手法であるため 1000 字の自動要約を行い、評価を行った。導入段落の抽出精度は 22%であった。また、結論段落の抽出精度は 36%であった。導入段落は継続段落数の閾値を決定することで再現率は犠牲となるが、精度を向上させることができる。今後の課題としては、段落からの導入段落および結論段落の特定や再現率の向上がある。

謝辞

本研究は (株) ジムコとの共同研究として行った。

使用した言語資源及びツール

(1) 形態素解析器「MeCab」,Ver.0.9.1, http://mecab.sourcefoge.jp/

(2) 国会会議録検索システム, http://kokkai.ndl.go.jp/

参考文献

[1] 山本 和英, 安達 康昭: 国会会議録を対象とする話しことば要約, 自然言語処理 Vol.12, No.1, pp.51-78, 2005.

[2] 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, 情報処理学会研究会報告 NL132-7, pp.49-56, 1999.

[3] Hearst, M. A.: Multi-paragraph segmentation og expository text, *Proc. of ACL-94*, pp.9-16, 1994.

付録

以下に、国会会議録を本手法によって 1000 字程度に要約した結果の一例を示す。導) は導入段落を表し、結) は結論段落を表している。

【要約文】(1000 字程度)

導) 次に、森林の有する多面的機能の発揮、林業の持続的かつ健全な発展を基本とする森林・林業政策の展開についてであります。

結) 平成十五年度の農林水産予算は、食の安全と安心の確保、農業の構造改革の加速化、都市と農山漁村の共生、対流を推進するとともに、地球温暖化防止等に資する森林整備の推進を中心とした森林・林業政策や、安全で安心な水産物供給体制の整備等の水産政策を展開するとの観点から、重点施策に思い切った予算配分を行うなど、新たな政策展開が図られるよう編成いたしました。

導) 地球温暖化の防止に向けて、京都議定書で我が国が約束した温室効果ガスの削減目標を達成するため、二酸化炭素の吸収源としての森林の果たす役割の発揮に向けて、昨年十二月、地球温暖化防止森林吸収源十カ年対策を策定しました。これに基づき、緑の雇用の推進などを通じて担い手の育成を推進しつつ、多様で健全な森林の整備保全等をより重点的に推進していく考えであります。

結) 第四に、森林・林業政策については、特に、京都議定書に定められた二酸化炭素などの温室効果ガスの削減目標を達成していく上で、我が国の森林を適切に整備保全していくことが極めて重要となっていることから、多様で健全な森林の整備保全を積極的に推進してまいります。

導) 次に、森林の有する多面的機能の発揮、林業の持続的かつ健全な発展を基本とする森林・林業政策の展開についてであります。

結) 平成十五年度の農林水産予算は、食の安全と安心の確保、農業の構造改革の加速化、都市と農山漁村の共生、対流を推進するとともに、地球温暖化防止等に資する森林整備の推進を中心とした森林・林業政策や、安全で安心な水産物供給体制の整備等の水産政策を展開するとの観点から、重点施策に思い切った予算配分を行うなど、新たな政策展開が図られるよう編成いたしました。

導) 次に、資源の持続的利用の確保を基本とした水産政策の展開であります。

結) 平成十五年度の農林水産予算は、食の安全と安心の確保、農業の構造改革の加速化、都市と農山漁村の共生、対流を推進するとともに、地球温暖化防止等に資する森林整備の推進を中心とした森林・林業政策や、安全で安心な水産物供給体制の整備等の水産政策を展開するとの観点から、重点施策に思い切った予算配分を行うなど、新たな政策展開が図られるよう編成いたしました。