概要

情報化社会の発展によりテキストが電子化され、テキストの大量管理・転送・加工・共有が可能となった。それと同時に、人間が処理しなければならない情報量も増加している。Web 上の情報を自動検索する技術が盛んに研究されているが、検索した情報の要旨を判断する作業は人手に頼らざるを得ない。

近年、人手による文書要旨判断の支援を目的とした、文書要約手法の研究が盛んに行なわれており、その多くは、重要と見なせる文を抽出した後、冗長要素を削減する手法 [1] を採用している。しかし、抽出文同士に意味的な繋がりがないため要約文の可読性が低いという問題があり、この問題を解決するため、単語の意味的連想関係ネットワークを使用する手法 [2] 等が提案されている。

本研究は、前述の問題点を解決する新しい手法として、手法 [2] の意味的関係ネットワークの概念をヒントにして、結合価パターン [3] と意味属性関係[1][4] で構成する「レベルネットワーク要約手法」を提案する。

「レベルネットワーク要約手法」とは、文書のすべての構成要素同士を意味的関係のリンクを張ったネットワーク上に再配置し、文意の中心からの距離によって重要度領域を区切ることで、各要素に重要度を付与する手法である。これにより、冗長要素を拾わない重要要素選択と、意味的連続性を持った要約文の生成が可能になる。

そして、今回の実験の結果、機械要約の観点から「レベルネットワーク要約」手法は非常に高い要約能力を持つことが実証されたが、人間特有の要約観点からは十分に満足できる結果とは言えなかった。

機械要約と人間特有の要約は根本的に方法論が異なっており、この両者を満足させる要約文を作成するのは困難なことである。しかし、機械要約を一般使用に耐えうる技術として確立させるためには、いかに人間の要約に近づけるかが重要となる。「レベルネットワーク要約」手法に、研究期間の都合上採用しなかった「意味的特異点となる直感的重要語」情報をレベルネットワークに組み込む事で、今後より一層の要約能力向上が期待できる。

[1] 一般名詞の意味的用法を表す 12 段 2710 の意味属性の上位下位関係、全体部分関係を木構造で示したもの

目次

1 はじめに 1

2 レベルネットワーク要約概要 2

3 レベルネットワーク要約アルゴリズム 3
3.1 文単位ネットワークの定義・・・・・・・・・・・・・・・・・・・・・・4
3.2 文意中枢の選定・・・・・・・・・・・・・・・・・・・・・・・・・・5
3.3 意味属性連結・・・・・・・・・・・・・・・・・・・・・・・・・・・6
3.4 修飾句の優位関係・・・・・・・・・・・・・・・・・・・・・・・・・9
3.5 要約文生成・・・・・・・・・・・・・・・・・・・・・・・・・・・・10

4 要約評価実験
4.1 実験方法・・・・・・・・・・・・・・・・・・・・・・・・・・・・・11
4.2 評価サンプル・・・・・・・・・・・・・・・・・・・・・・・・・・・12
4.3 要約評価実験結果・・・・・・・・・・・・・・・・・・・・・・・・・16
4.4 結果考察・・・・・・・・・・・・・・・・・・・・・・・・・・・・・17

5 おわりに 18

6 参考文献 19

図目次

1 レベルネットワーク要約概略図・・・・・・・・・・・・・・・・・・・・・・2
2 文単位ネットワーク具体例・・・・・・・・・・・・・・・・・・・・・・・4
3 TF*IDF 加算具体例・・・・・・・・・・・・・・・・・・・・・・・・・・5
4 意味属性による連結関係図 (1)・・・・・・・・・・・・・・・・・・・・・6
5 意味属性による連結関係図 (2)・・・・・・・・・・・・・・・・・・・・・7
6 意味属性による連結関係図 (3)・・・・・・・・・・・・・・・・・・・・・8
7 意味属性による連結関係図 (4)・・・・・・・・・・・・・・・・・・・・・8
8 修飾句の優位関係図・・・・・・・・・・・・・・・・・・・・・・・・・・9
9 優先順位入れ替え図・・・・・・・・・・・・・・・・・・・・・・・・・・10
10 文単位ネットワーク構成具体例・・・・・・・・・・・・・・・・・・・・13
11 意味属性連結具体例・・・・・・・・・・・・・・・・・・・・・・・・・14

表目次

1 TF*IDF 計算例・・・・・・・・・・・・・・・・・・・・・・・・・・・5
2 意味属性距離計算例・・・・・・・・・・・・・・・・・・・・・・・・・6
3 要約評価実験結果・・・・・・・・・・・・・・・・・・・・・・・・・15

1　はじめに

従来の文書要約手法は、重要と見なせる文を文単位で抽出後、文の冗長要素を削除できないため、不必要な要素を含めた要約文生成を行なう事になる。そのため、従来の多くの要約手法には、冗長要素を削除するアルゴリズムを後付けで実行する手法が採用されているが、抽出文同士に意味的関連性が考慮されていないため可読性が低く、抽出文同士を結び付けるため、再加工アルゴリズムが必要になるという問題もある。

そこで、本研究は、文書のすべての構成要素を単語単位に解析し、各単語の意味的連結によるネットワークを作成して文書構成要素を再配置する。そして、文意の中心からの距離によってを重要度領域で区切り、文書構成要素に重要度順位付けを行ない、意味的に必要な部分のみを選択して文を再構成する。これにより、冗長要素をあらかじめ除外し、かつ意味的な繋がりを持った要約文構成要素の抽出が可能となり、より人間の要約能力に近い要約文生成を実現する。

本研究はこの手法を「レベルネットワーク要約」として提案する。

2　レベルネットワーク要約概要

文書の最小構成単位である単語は、結合価パターンによって文意の最小単位である節（本研究ではこれを結合価キーワードと呼ぶ）に区切る事ができる。

この結合価キーワード同士を意味属性関係による意味的類似性によって連結したネットワークを作成ことにより、文書がどのような意味的な連結関係で構成されているのかが明らかになる。

文意の中心をネットワークの中心とし、中心からの意味的な距離により領域と領域内部の構成要素にレベル付けを行なう。そして、そのレベルを基準にして文を再構成したものを要約文とする。これがレベルネットワーク要約の概要である。

図 1 に、レベルネットワーク要約の概略図を示す。

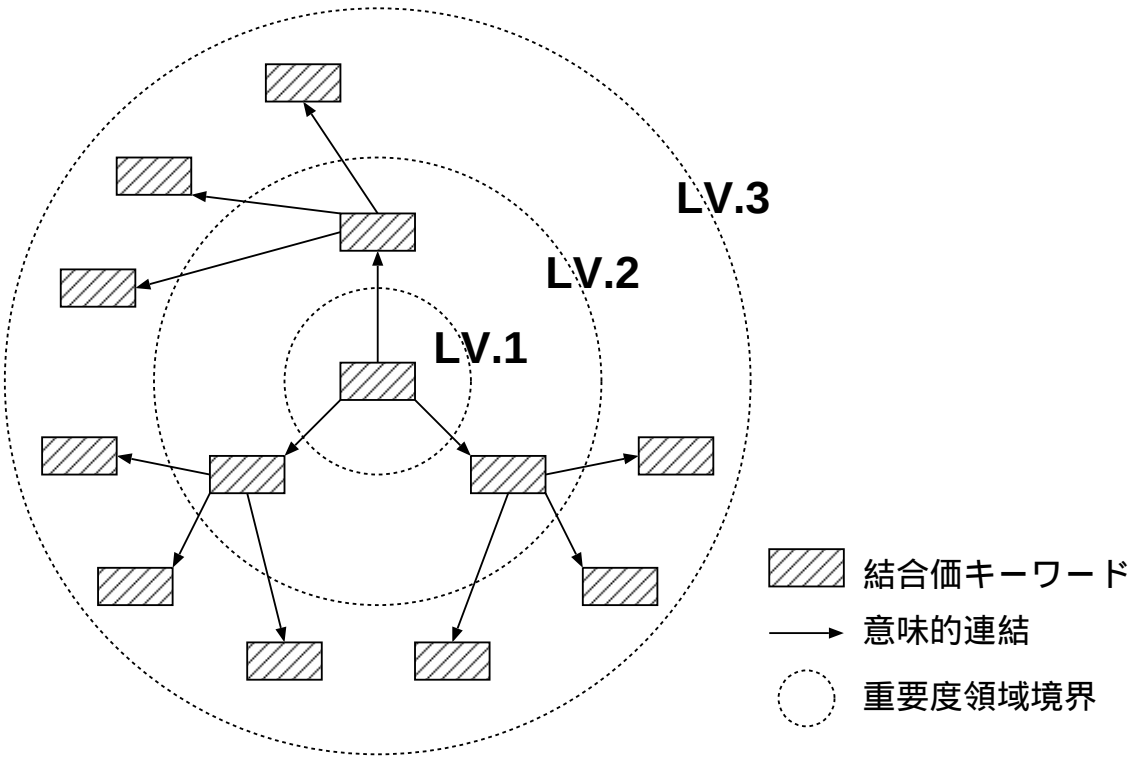

図 1 レベルネットワーク要約概略図

3　レベルネットワーク要約アルゴリズム

レベルネットワーク要約は、以下 5 つのアルゴリズムで構成する。

3.1 要約対象記事の解析

- 文書の形態素解析
- 結合価パターン適用による結合価キーワードの決定
- 文単位ネットワークの作成

3.2 文意中枢の選定

- TF*IDF の算出
- TF*IDF による文単位ネットワークの重要度付け

3.3 意味属距離による連結

- 各結合価キーワードの意味属性距離比較
- 最短意味属性連結と絶対・相対レベル決定
- 補完意味属性連結によるサブネットワーク連結

3.4 修飾要素の優位関係の決定

- 修飾句の重要度別分類

3.5 レベル選択要約文生成

- 必須修飾句の追加
- 前後関係・接続関係を考慮したレベル調整と語尾変換
- 指定要約率に対応した具体修飾句と接続的修飾句の追加

なお、3.4、3.5 の手順は計算機処理ではなく、著者の主観による手作業で行なう。

3.1 要約対象記事の解析

形態素解析[2][5] により、与えられた文を形態素・語の並びに分解し、それぞれの形態素・語の品詞を決定するとともに、名詞の意味属性コードを調べる。

形態素解析の結果決定した文書中の動詞について、結合価パターン[3]を適用して、文を文意の最小構成単位である結合価キーワードに区切る。

結合価キーワードを図式化すると、図 2 の例ような、文の中心動詞と結合価キーワードを結ぶ階層的構造となる。これを「文単位ネットワーク」として定義する。

具体例

原文 :

政府が記者団に「内閣改造が JCO 事故によって延期された」と発表した。

結合価パターン :

延期する ： N [humain] が + N [action] を + V
政府 (が) 内閣改造 (を) 延期する

発表する ： N [humain] が + N [humain] に + S と + V
政府 (が) 記者団 (に) S (と) 発表する

付属語による格助詞対応の変化 :

延期された ： N [action] が + V
内閣改造 (が) 延期された

結合価パターンの合成 :

政府 (が) 記者団 (に) 内閣改造 (が) 延期された (と) 発表した

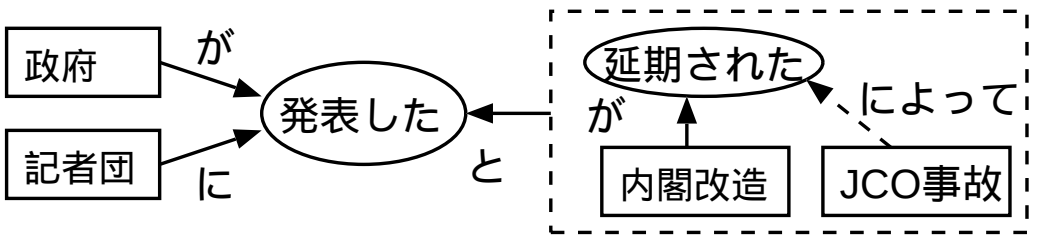

図 2 文単位ネットワーク具体例

[2] sample_jaws(C)1995,1999,NTT を使用
[3] 用言を「体言 + 格助詞」との結合関係で捉え体言の意味特徴と格助詞の種類により記述する型

3.2 文意中枢の選定

結合価キーワードの重要度には TF*IDF[4]を使用する。TF 値とは、要約対象文書の対象単語頻度であり、IDF 値とは、94 年毎日新聞 109300 記事、213026 単語から算出した対象単語の出現頻度である。

TF*IDF 計算式

$$\text{TF*IDF} = \text{単語の頻度} \cdot \log\left(\frac{\text{総文書数}}{\text{単語を含む文書数}}\right)$$

表 1.TF*IDF 計算例

名詞	TF	IDF	TF*IDF
政府	2	2.48	4.96
記者団	1	3.79	3.79
内閣改造	1	7.40	7.40
JCO 事故	3	6.93	20.80

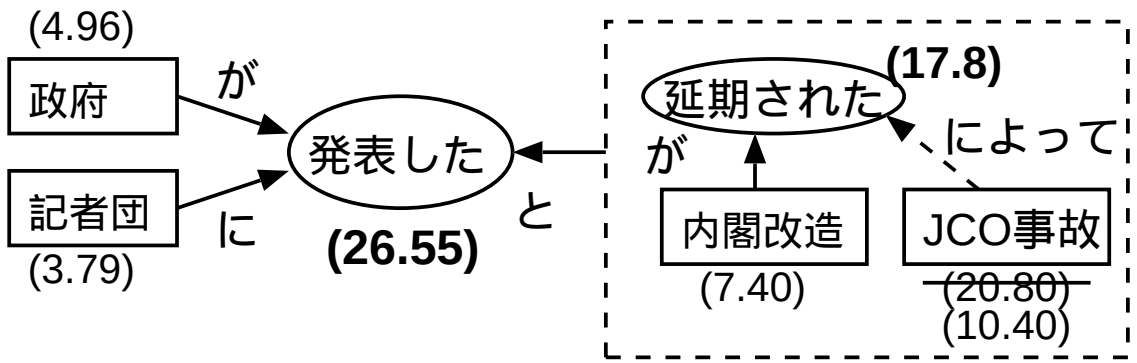

図 3 TF*IDF 加算具体例

各結合価キーワードの TF*IDF 値を中心動詞に加算する。この時、中心動詞との階層的距離に応じて修飾句の TF*IDF 値に修正を加える。図 3 の例ならば、中心動詞と距離 2 の修飾句「JCO 事故」の TF*IDF 値は 1/2 になる。

[4]文書頻度の逆数による牽引語の重要度評価

3.3 意味属性連結

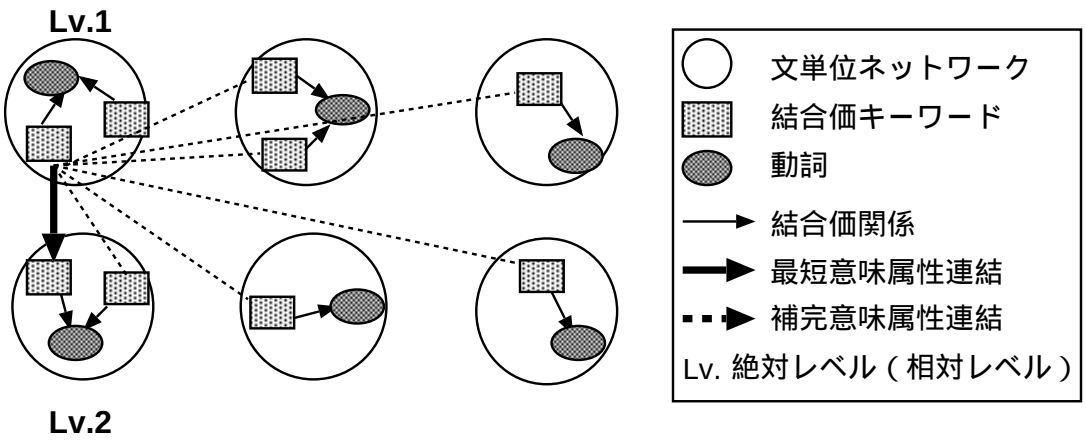

図 4 意味属性による連結関係図 (1)

文意中枢となる文単位ネットワークを絶対レベル Lv.1 と定義し、Lv.1 に含まれる結合価キーワードと、他の文単位ネットワークの結合価キーワードとの意味属性距離を比較し、最短となるものを選択する。こうして結ばれた結合価キーワードを含む文単位ネットワークを、Lv.2 と定義する。

例

「東京都千代田区の国立劇場」と「陛下」の意味属性距離計算例

表 2. 意味属性距離計算例

	名詞	陛下	陛下	陛下	陛下
名詞	意味属性コード	43	320	163	173
東京都	464	5.00	5.50	6.00	6.00
東京都	364	2.75	3.00	3.25	3.25
千代田区	464	5.00	5.50	6.00	6.00
千代田区	364	2.75	3.00	3.25	3.25
国立劇場	409	6.00	6.25	7.00	7.00

名詞の意味属性距離 = 比較対象となる両者の意味属性コードのノードの深さの和を、共通の親のコードの深さの倍数で除算し、逆数を取る

結合価キーワードの意味属性距離 = 結合価キーワードに含まれる名詞の意味属性距離の総平均

(表 2 の場合、平均意味属性距離=4.79)

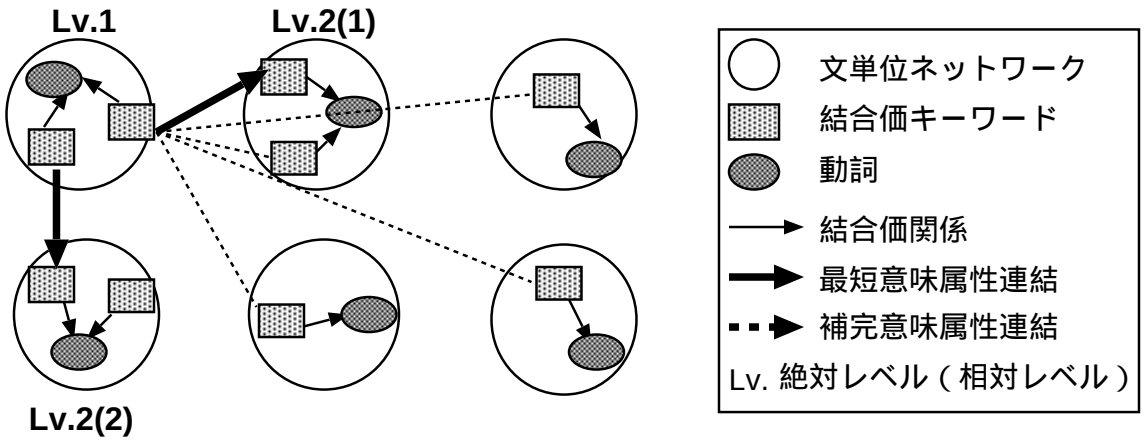

図 5 意味属性による連結関係図 (2)

同様に、Lv.1 の文単位ネットワークのもうひとつの結合価キーワードと他の文単位ネットワークとの意味属性距離を比較し、最短となるものを選択する。

この際、Lv.2 の領域に属する文単位ネットワークが複数存在するため、文単位ネットワークの中心動詞の TF * IDF 値を比較し、相対レベルを付与する。絶対レベルと相対レベルの組み合わせによって、文単位ネットワークの重要度順位は図 5 の場合、重要度の高い順に、Lv.1、Lv.2(1)、Lv.2(2) となる。

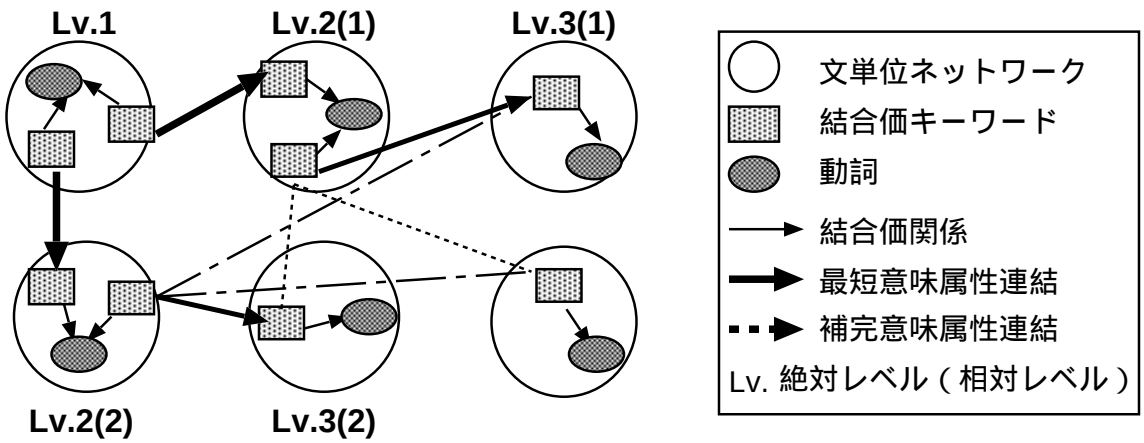

図 6 意味属性による連結関係図 (3)

同様に、Lv.2 に含まれる結合価キーワードと、未選択の文単位ネットワークの結合価キーワードとの意味属性距離比較を行ない、最短となるものを選択し、Lv.3 の文単位ネットワークを定義し、相対レベルを付与する。

ここで、Lv.3 には未選択の結合価キーワードがなくなるが、未選択の文単位ネットワークは残っている。その残った文単位ネットワークを最下位 Lv.4 とする。

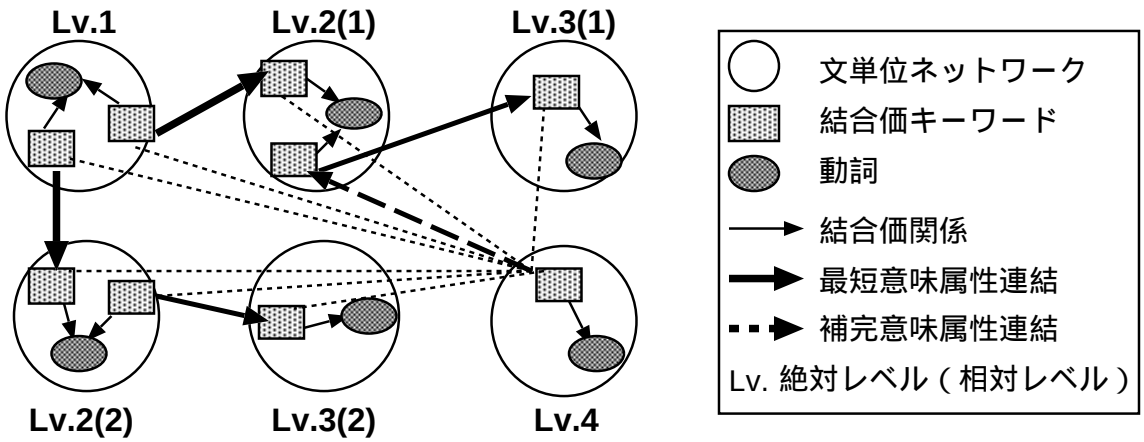

図 7 意味属性による連結関係図 (4)

最下位 Lv.4 の結合価キーワードと、上位レベルの結合価キーワードとの意味属性距離を比較し、最短となるものを選択する。これを、補完意味属性連結と呼ぶ。最下位レベルは、補完意味属性連結によって連結される、上位の文単位ネットワークのレベルに依存するサブネットワークとして取扱う。

3.4 修飾句の優位関係

次に、修飾句が所属している文単位ネットワークのレベルに基づいて修飾句の優位関係を決定し、修飾句を図 8 のように必須修飾句、具体修飾句、接続的修飾句の 3 つに分類する。 この分類は、要約文生成時の文字数調整に使用する。

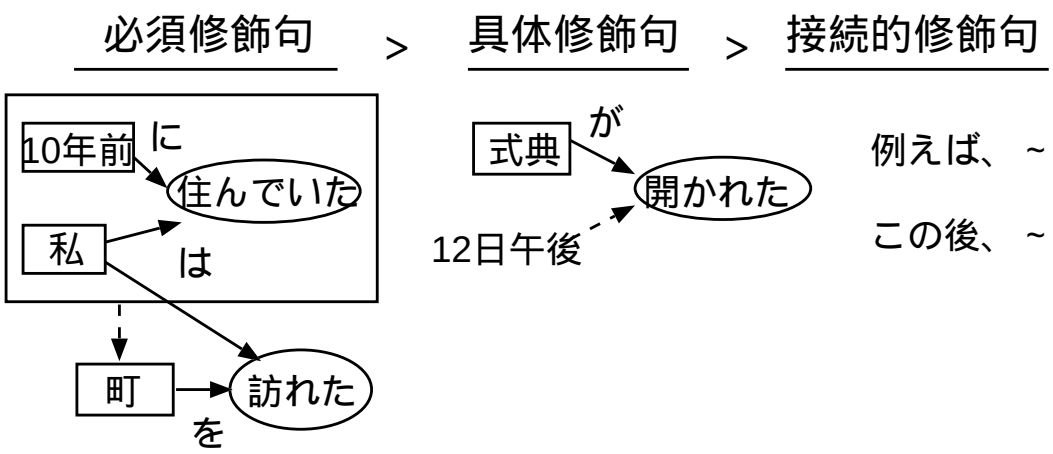

図 8 修飾句の優位関係図

必須修飾句は、図 8 の例では「私は町を訪れた」という最小単位の文では掴むことのできない事象 (または行為) を修飾するものである。「10 年前に住んでいた町」という情報が、「訪れた」という中心動詞をより具体化することができる、意味的に省略できない修飾句である。

具体修飾句とは、動詞を修飾する時制・位置・数量関係である。 具体修飾句の重要度は修飾している動詞の重要度に依存する。

接続的修飾句とは、文頭で接接続詞的に使用される修飾句である。要約文の前後関係を表し、可読性調整において使用する。

3.5 要約文生成

(1) 文単位ネットワークに、3.4 節で定義した必須修飾句を加える。

(2) 3.3 節で重要度順位付けした文単位ネットワークについて、文末が句点であり、かつ原文で一文関係にある文単位ネットワークの要素が、1 つ下の絶対レベルに存在する場合、両者を結合して下位要素の絶対レベルに降格する。それ以外の場合は結合対象とせず、両者の文末表現を修正し、句点を読点に変換する。

(3) 未選択の最上位レベルの文単位ネットワークを選択。この時点で指定要約率を越えていなければ、(1) で除外した具体修飾句を追加する。この処理を未選択の最上位レベルに移行させ、指定要約率を越えない範囲で繰り返す。また、接続的修飾句は指定要約率範囲内で追加できる具体修飾句が無い場合に追加する。

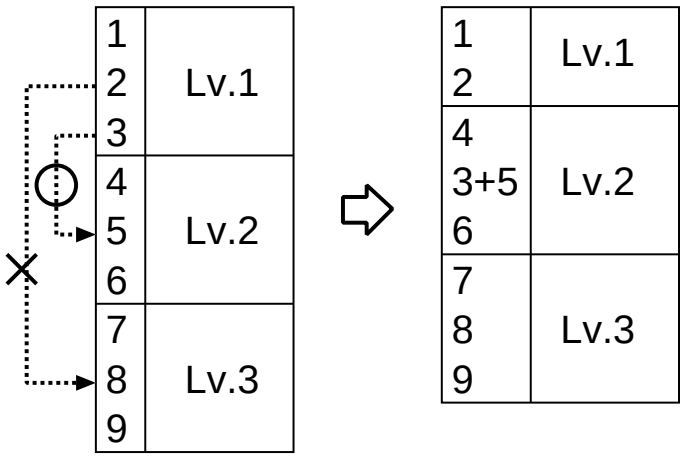

図 9 優先順位入れ替え図

4　要約評価実験

レベルネットワーク要約手法は新提案であるため、他の要約手法と単純に比較する事はできない。
よって要約評価実験は、自己相対評価形式で行なう。

4.1　実験方法

要約評価は、以下の 3 種類の文書について要約率を 50%、25%に設定して、7 人の被験者を要約手法知識のある被験者と要約手法知識のない被験者とにグループ分けし、文の読み易さを「人間と同等の要約、良い要約、ある程度の要約、悪い要約、使い物にならない要約」の 5 段階で評価する。

- 文書 1 見出し記事を持つ新聞記事　　12 文 667 文字
- 文書 2 読者投稿書評　　9 文 341 文字
- 文書 3 見出し記事を持たない新聞記事　　8 文 662 文字

次項以降に、評価サンプルの原文、文単位ネットワーク構成図、意味属性連結関係構成図、要約率 50%、要約率 25%の例を示す。

4.2　評価サンプル

原文（662文字）

政府主催の「天皇陛下御在位十年記念式典」が十二日午後、天皇、皇后両陛下を迎えて東京都千代田区の国立劇場で開催され、小渕恵三首相、衆参両院議長、最高裁判長官ら約千三百人が参加した。陛下はお言葉で、阪神大震災やソ連崩壊などに触れながら十年間を振り返り「今日、わが国の経済や社会は大きな課題に直面しています。国民の英知と努力によってこれらの困難が克服されることを確信しています」と述べられた。同日は宮内庁庁舎前で一般参賀が行われ、午後四時すぎの終了時までに約八千五百人が記帳した。各地の郵便局では記念切手を発売、金融機関では記念の五百円白銅貨の引き換えも始まった。

式典では君が代斉唱の後、小渕首相が「人々の幸せと平和を祈られ、国民の悲しみをご自身の悲しみとしてお受け止めになる両陛下のお姿を拝見して私どもは強い感銘を受け、勇気づけられました」と式辞を述べ、続いて伊藤宗一朗衆議院議長らがあいさつした。

お言葉先立ち、世界的チェロ奏者ヨーヨーマさんの演奏や児童合唱団らによるコーラスが披露され、陛下が一九七五年に沖縄県を初訪問した際に詠んだ琉歌に皇后さまが曲を付けた「歌声の響き」が演奏されると会場から拍手がわいた。

この後、皇居前広場を中心に、民間や超党派の国会議員団体が主催する「天皇陛下御即位十年をお祝いする国民祭典」が開かれた。

各界著名人が出席して開かれた祝賀式典では、両陛下もちょうちんを手に二重橋に立ち、約二万五千人の参加者を前に「これからの日々が、少しでも平和で希望に満ちたものとなるように願っています」とお言葉を述べた。

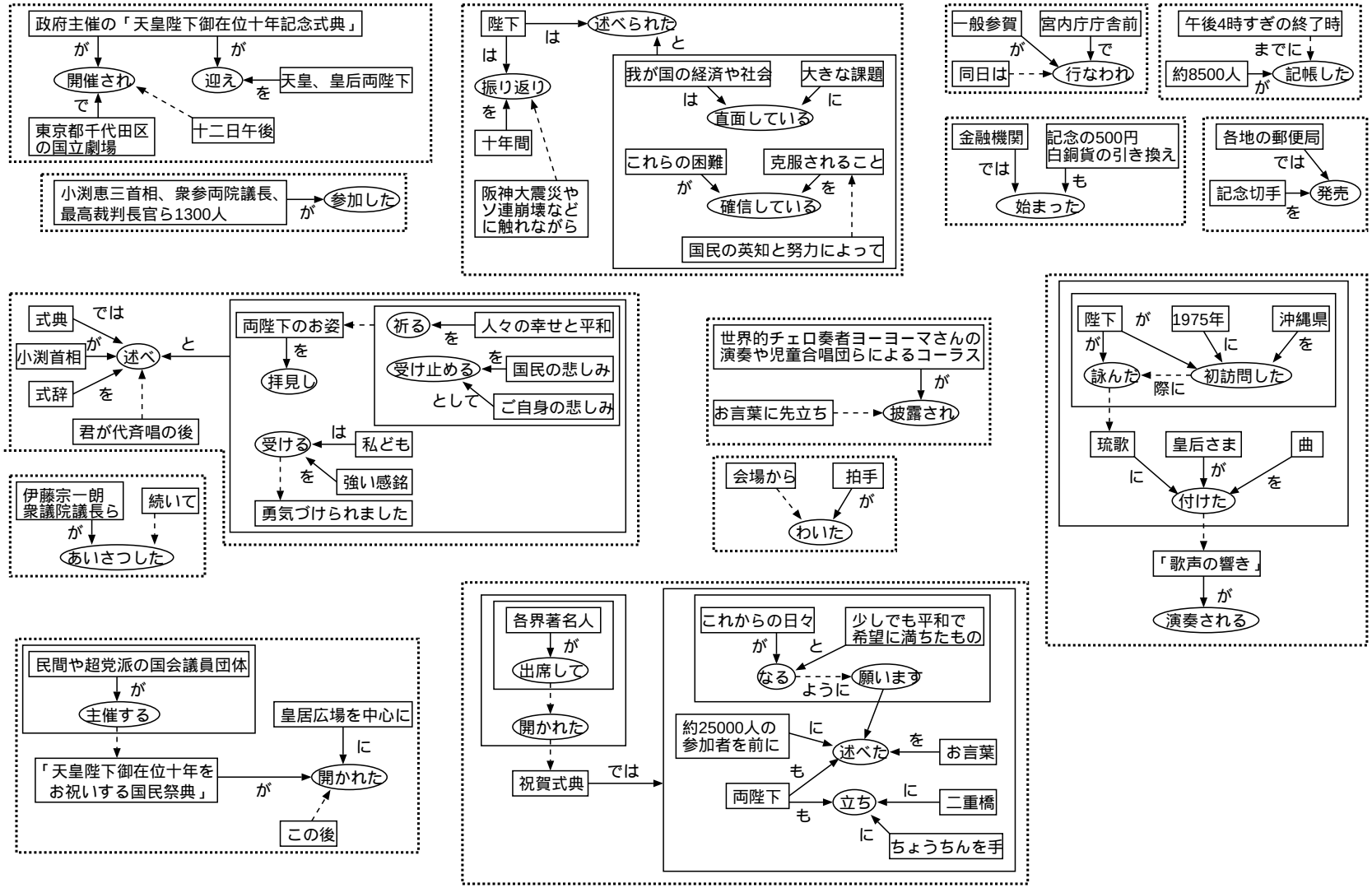

図10 文単位ネットワーク具体図

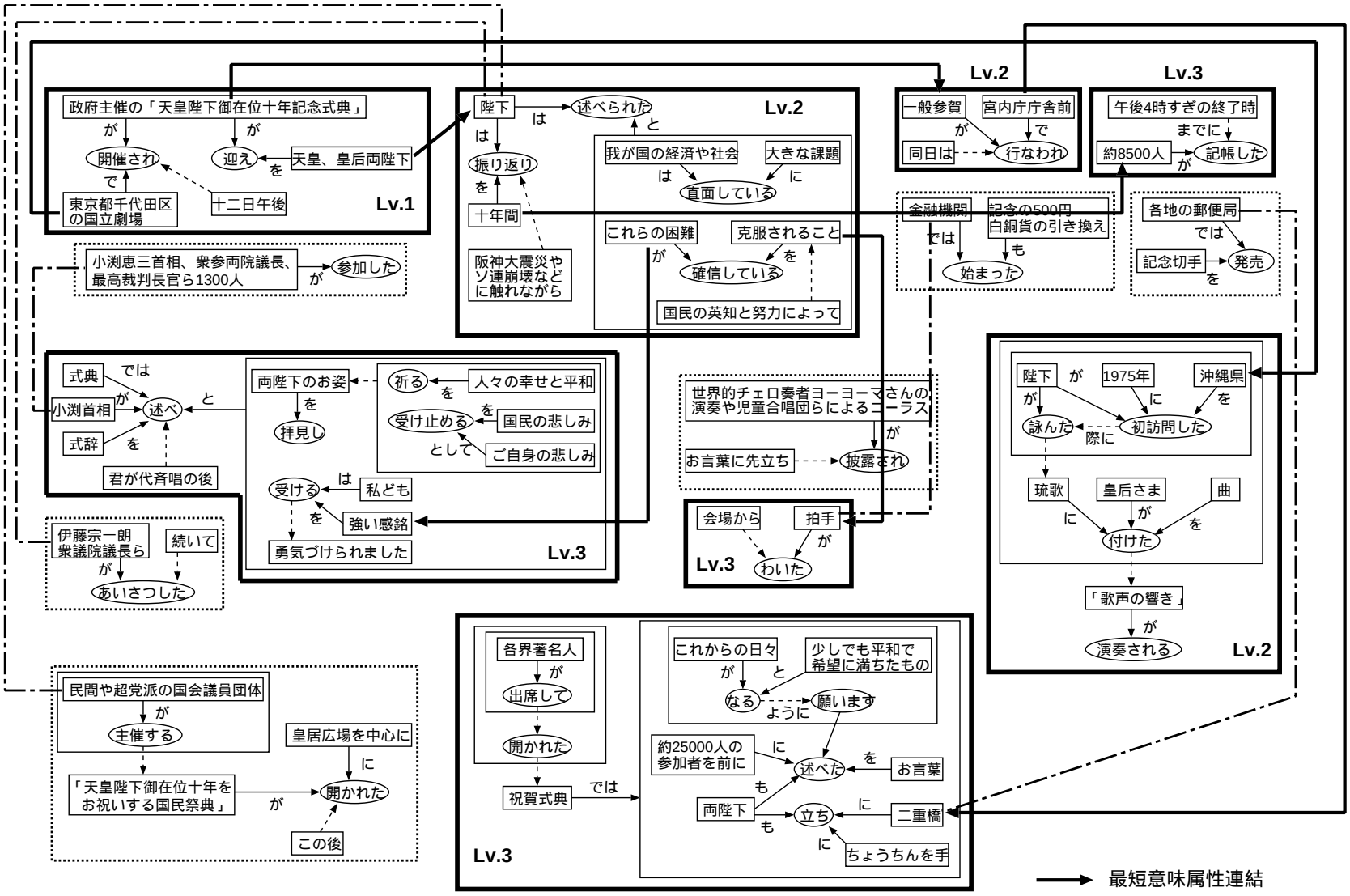

図11 意味属性連結関係具体図

要約率 50% (331 文字以下) (331 文字)

政府主催の「天皇陛下御在位十年記念式典」が十二日午後、天皇、皇后両陛下を迎えて東京都千代田区の国立劇場で開催された。

陛下はお言葉で、阪神大震災やソ連崩壊などに触れながら十年間を振り返り「今日、わが国の経済や社会は大きな課題に直面しています。国民の英知と努力によってこれらの困難が克服されることを確信しています」と述べられた。

祝賀式典では、両陛下もちょうちんを手に二重橋に立ち、約二万五千人の参加者を前に「これからの日々が、少しでも平和で希望に満ちたものとなるように願っています」とお言葉を述べた。

式典では、小渕首相が「人々の幸せと平和を祈られ、国民の悲しみをご自身の悲しみとしてお受け止めになる両陛下のお姿を拝見して私どもは強い感銘を受けました」と式辞を述べた。

要約率 25% (165 文字以下) (157 文字)

政府主催の「天皇陛下御在位十年記念式典」が、天皇、皇后両陛下を迎えて東京都千代田区の国立劇場で開催された。

陛下はお言葉で、阪神大震災やソ連崩壊などに触れながら十年間を振り返り「今日、わが国の経済や社会は大きな課題に直面しています。国民の英知と努力によってこれらの困難が克服されることを確信しています」と述べられた。

4.3 要約評価実験結果

表 3 要約評価結果

要約手法知識グループ	要約手法知識（有）		要約手法知識（無）	
要約率	25%	50%	25%	50%
5（人間と同等の要約）		8%		
4（良い要約）	42%	84%		
3（ある程度の要約）	42%		22%	56%
2（悪い要約）	16%	8%	45%	33%
1（使い物にならない）			33%	11%

7 人の被験者によって 3 種類の文書で要約率 25%、50%で作成した要約文を 5 段階評価した結果を表 3 に示す。なお、被験者知識グループの分類は著者の主観によって行ない、サンプル文書と要約手法への慣れによる評価の揺れを防ぐため、実験は 1 回のみとした。

その結果、要約手法知識のある被験者グループにおいては、要約率 25%で「良い要約 or ある程度の要約」は 84%、要約率 50%では「人間と同等の要約 or 良い要約」が 92%に達し、非常に良好な結果を得ることができた。

これに対し、要約手法知識のない被験者グループにおいては、要約率 25%で「ある程度の要約」は 22%、要約率 50%では「ある程度の要約」は 56%に止まるという結果になった。

4.4 結果考察

要約手法知識のある被験者グループ

要約率 25%では 84%の被験者が「ある程度 or 良い要約」、要約率 50%では 92%の被験者が「良い要約 or 人間と同等」と評価した。

このように高い評価が得られた理由として、要約手法知識のある被験者は、機械要約がいかに複雑で困難なものであるかを理解しており、かつ文書構造と情報量の保持という、要約効率を重視する機械要約評価を基準にして評価した結果であることが挙げられる。

要約率については、要約率 50%で「良い評価」以上が 92%、要約率 25%では「良い要約」は 42%となり、機械要約の実用の目安となる要約率 30%でも、かなり高い精度が期待できる結果となった。

要約手法知識のない被験者グループ

要約率 25%では 22%の被験者が「ある程度の要約」、要約率 50%では 56%の被験者が「ある程度の要約」と評価した。

このように要約手法知識のある被験者と対照的な結果となった理由として、要約手法知識のない被験者にとって要約とは、人間が直感的な重要語の組み合わせで要約文を再構成して人間が作る要約文であり、被験者が作る理想的な要約文と実験結果の要約文とを比較するという、人間特有の要約評価を基準にして要約文を評価した結果であることが挙げられる。

要約率については、要約率 50%で「ある程度の要約」が 56%、要約率 25%では「ある程度の要約」は 22%となり、要約率 30%での実用には適さないと言える結果となった。その原因は、人間が判別する重要語は文関係に拘束されず、個々の繋がりに規則性が見い出しにくいという性質によるものである。

5　おわりに

結論

今回の実験結果から、文書構造と情報量の損失を最小限に抑えるという機械要約の観点において、本研究が提案するレベルネットワーク要約が有効であることが実証できた。

それと同時に、人間の要約能力との差異も明らかになったが、本研究には研究期間の関係上取り入れることのできなかった手法があり、その手法を今後発展させることにより、計算機の要約能力と人間の要約能力との差異を縮める事が可能である。

今後の課題

人手の介入が必要となる文単位ネットワークの階層構成、および、独自の分類法を使用している修飾句分類手順の自動化が必要となる。

そして、人間が直感的に判断している重要語を意味的特異点として識別し、レベルネットワークに組み込む手法を考案することで、より人間の要約能力に近い要約文生成の実現が期待できる。

6　参考文献

[1] 船坂、山本、増山 ： 冗長度削減による関連新聞記事の要約、
情報処理学会研究報告、Vol.96、No.65、pp39-46(1996)

[2] 窪田、山下、吉田 ： 要約文生成のための単語抽出方法、
情報処理学会研究報告、Vol.98、No.99、pp-143-149(1998)

[3] 水谷、石綿、荻野、賀来、草薙 ： 文法と意味Ｉ、朝倉書店 (1983)

[4] 池原、宮崎、白井、横尾、中岩、小倉、大山、林 ： 日本語語彙大系、
岩波書店 (1997)

[5] 長尾 ： 自然言語処理、岩波書店 (1996)