修 士 論 文

電子メールコミュニケーションにおける討議内容の要約と呈示法について

北陸先端科学技術大学院大学
情報科学研究科情報システム学専攻

渡邊 大貴

2002年3月

修 士 論 文

電子メールコミュニケーションにおける討議内容の要約と呈示法について

指導教官　落水浩一郎 教授

審査委員主査　落水浩一郎 教授
審査委員　篠田陽一 教授
審査委員　片山卓也 教授

北陸先端科学技術大学院大学
情報科学研究科情報システム学専攻

010128 渡邊 大貴

提出年月: 2002 年 2 月

概 要

本稿では、電子メールを用いたコミュニケーションにおける討議内容の要約手法を考案する。要約手法として、tf*idf 法を用いて各発話中の索引語に tf*idf 重み付けを行ない、その値を手がかりに重要発話を抽出する。また、 この要約手法をもとに、討議内容の自動要約を行なう討議内容要約エンジンを実現し、討議内容とその要約を視覚表示するシステムを設計し、実現する。

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

近年インターネットの普及により、地理的に分散した作業者同士が、電子メール等を利用し、ソフトウェア開発などの共同作業を行なう機会が増加している。電子メールを利用したコミュニケーションにおいては、対面の場合のような話者交代を行なうことができないため、複数の話題を同一メール中に記述する傾向にある。そのため、複数の議論が並行に進行しやすく、議論の流れを把握しづらいといった問題がある。この問題を解決するための情報呈示手法が提案されつつあり、本研究室においても、討議の流れを構造化するための討議構造モデルを提案し、討議構造を自動抽出するツール（討議構造木抽出エンジン）と討議構造を呈示するためのメールクライアントの開発が行なわれた [1]。しかし、現時点での討議構造の呈示手法では、討議全体の構造を呈示することが困難である。

自然言語処理の分野において、読み手にテキストの内容を容易に把握させるという目的で、テキスト要約技術に関する研究が行なわれている。重要文抽出による要約手法や、近年では、抽象化、言い換えによる要約手法が考案されている [2]。重要文抽出による要約手法では、テキスト中の文を１つの単位とし、それらに何らかの情報をもとに重要度を付与し、それをもとに重要な文を選択するという手法がとられている。この技術は、討議構造を呈示する際に必要な情報のみを抽出することに利用できるものと考えられる。例えば、重要箇所抽出法を用いてメール情報の抽出システムが実現されている。しかし、このシステムは、一つのメール文から情報を抽出し、個々のメール文の要約を呈示するシステムであり、討議の流れの理解を支援するものではない。

1.2　本研究の目的

本研究の目的は、全ての討議内容の重要箇所を表示し、また、討議の参加者が求める討議内容も表示できるような視覚表示するシステムの設計・実現である。具体的には、以下に示す要約抽出機能と、視覚表示システムを設計・実現する。

1. 複数の話題が並列に議論されている場合、議論の流れを呈示するだけでは、参加者が有用な議論のみを検索することは困難である。そこで、討議内容を要約の対象とし、話題が終結した討議内容からは議題とそれに対する結論となる情報を抽出し、未終結な討議内容からは議題と最終提案となる情報を抽出する手法を考案・実現する。

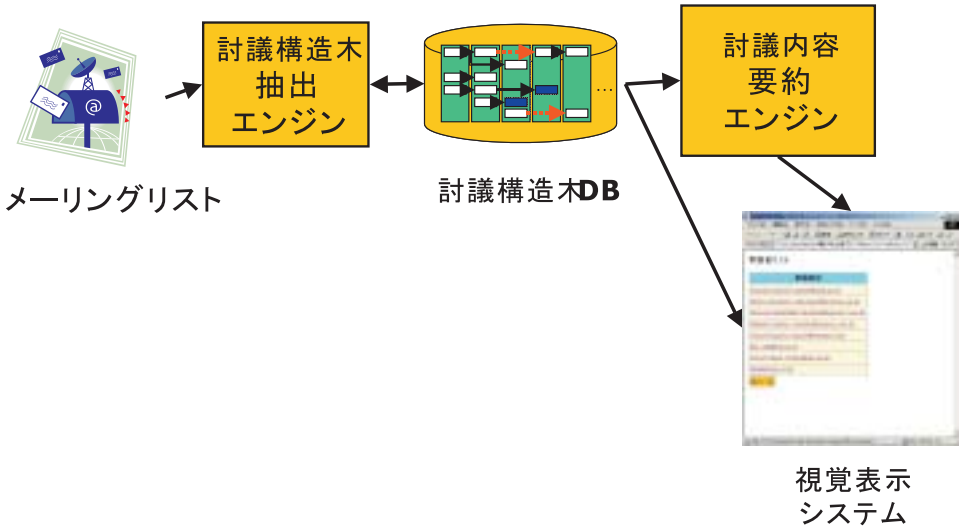

図 1.1: 討議内容の要約と呈示法へのアプローチ

2. 視覚表示システムを設計・実現する。このシステムは、討議の参加者を中心とし、指定した討議の参加者が関係している討議内容の要約とその討議内容を視覚表示する。また、全ての討議内容の要約とその討議内容も視覚表示する。

1.3 本論文の構成

本研究で実現した討議内容要約エンジンと視覚表示システムの概略を図 1.1 に示す。コミュニケーションを構成する電子メールは、まず討議構造抽出エンジンに送られる。送られた電子メールは、討議構造抽出エンジンによって発話に分割され、これまでに構築した討議構造に追加される。討議内容要約エンジンは、構築した討議構造を入力とし、各討議内容の要約を抽出する。討議の参加者は、視覚表示システムを通じて討議内容の要約とその討議内容を参照することができる。

本論文では、2 章で、討議の流れを構造化するための討議構造モデルと、討議構造を自動抽出する討議構造木抽出エンジンについて述べる。また、討議構造木抽出エンジンの精度向上の成果についても述べる。3 章では、討議構造参照機能を有するメールクライアント (ICEMail++) の概略と呈示法における問題点を述べる。4 章では、討議内容から重要な発話を抽出する討議内容の要約手法について述べる。5 章では、討議内容の自動要約である討議内容要約エンジンについて述べる。6 章では、討議内容と討議内容の要約を視覚表示するシステムについて述べる。最後に 7 章では、まとめと今後の課題を述べる。

第2章　討議構造木抽出エンジンの精度向上

本研究室において、討議の流れを構造化するための討議構造モデルを提案し、その構造を自動抽出するツールとして、討議構造木抽出エンジンの開発が行なわれた。本章では、討議構造モデルと討議構造木抽出エンジンについて紹介し、討議構造木抽出エンジンの精度向上を行なった結果について述べる。

2.1　討議構造木とは

討議構造木は、図 2.1 に示される構造を形式的に表現するために、対面による 2 者の会話モデルであるコントリビューションツリー [3] を拡張したモデルである [4][5]。

なお、図 2.1 では、発話と発話のつながりは、相手の発話と返答発話という関係しか表現しないが、討議構造木は、返答発話が、相手の発話意図を適切に理解したことを前提に行なわれたものであるか否かをも表現できる。

発話の定義

討議構造木を構成する発話の定義を以下に述べる。

- 相手からの返答を求めない、議論の流れに直接の影響を与えない文章 (以後、宣言的な発話と呼ぶ。)
- ある話題に関する質問や提案など、相手からの返答要求を示唆している文章 (以後、返答を要求する発話と呼ぶ。)
- 話題を終結するような同意や受諾などを示唆するような発話 (以後、話題を終結する発話と呼ぶ。)

フェーズの定義

コントリビューションツリーに基づく対面の会話モデルに、Novic によるコントリビューショングラフがある [6]。討議構造木では、コントリビューショングラフにおけるプライマリエビデンスとセカンダリエビデンスの概念を取り入れている。プライマリエビデンスとセカンダリエビデンスの定義を以下に示す。

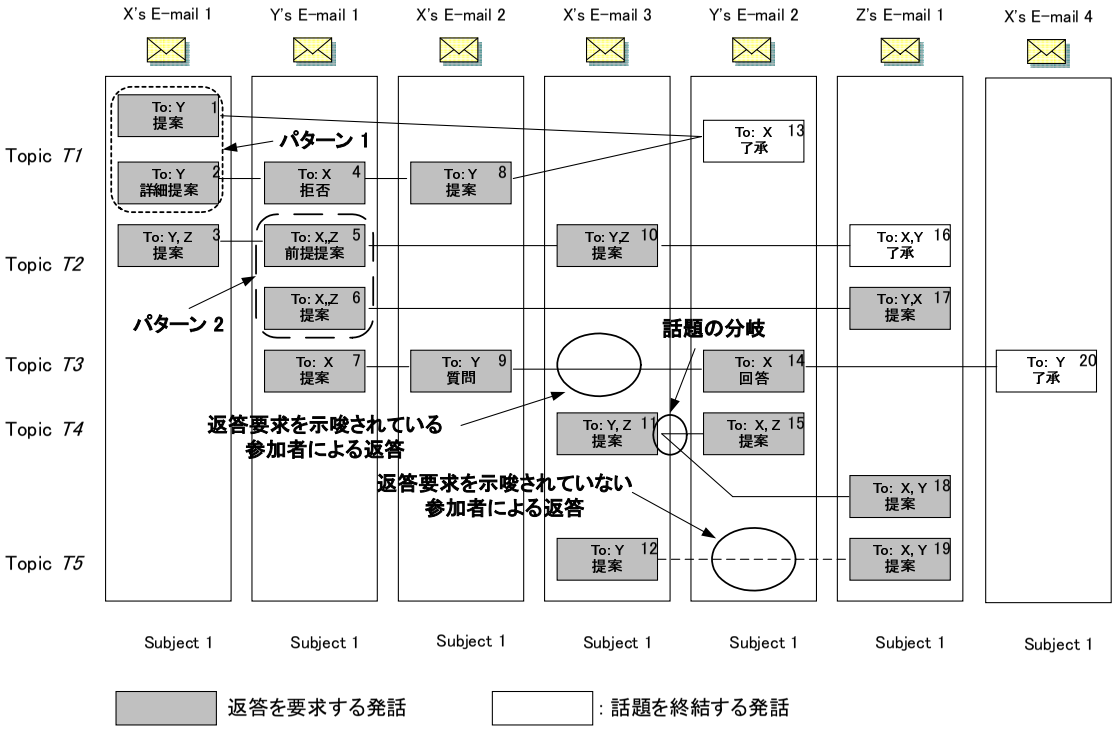

図 2.1: 電子メールを利用した討議構造

プライマリエビデンス: 聞き手の中の Y_i が、自分が話し手 X の意図する聞き手であると信じるときに示す行動 e のこと。つまり、Y_i が、自分との共通理解を示すことを作業者 X に要求されていると信じるときに示す行動のこと。

セカンダリエビデンス: プライマリエビデンスとは異なり、聞き手の中の Y_j が、X の意図する聞き手でない場合、もしくは、X がプライマリエビデンスを求めていないと Y_j が判断した場合の行動 e' のこと。

プライマリエビデンスとセカンダリエビデンスの概念を取り入れた。討議構造木における Pr、$Ac/InAc$ の定義を以下に示す。

Pr**:** 作業者 X が、複数の作業者の中のある作業者たちに、話題を継続/ 終結するといった意思を伝達するための発話 u を行なうフェーズのこと。ある作業者たちの中の Y_i がプライマリエビデンス e を示したならば、発話 u による X の意思を Y_i が適切に理解していると X が信じることができる。

Ac**:** ある作業者 Y_i が発話 u を受取り、発話 u で示される作業者 X の意思を適切に理解したことを意味するプライマリエビデンス e を示すフェーズのこと。このとき、作業

者 Y_i は、プライマリエビデンス e を示したならば、「X の意思を自分が理解していることを作業者 X が信じる」ということを信じることができる。

$InAc$: 作業者 X による返答を示唆されていない作業者 Y が、発話 u を理解したか否かを理解できると思われる行動、つまり、セカンダリエビデンス e' を作業者 X に与えるフェーズのこと。

属性の定義

C，Pr，Ac における属性を示す.

C(**トピック番号**) :

トピック番号は、サブジェクト番号毎に導入されたトピックに割り当てた番号のことで、導入された順番に、$T1, T2, T3...$ とナンバリングされる。

$Pr(WhoPr, WhomPr, PrS)$:

$WhoPr$ は、そのフェーズに属する発話を行なった参加者名を示す。$WhomPr$ は、発話者の意思を伝達する相手を示す。これは、以下の値をとる。

All : 参加者全員

参加者の名前の列 :例えば、A、B、C に対して意思の伝達を行なう場合、
$Whom = A/B/C$ となる。

PrS は、Pr に属している発話の状態を示す。

R : 返答を要求し、話題を継続している。

F : 返答を行ない、話題を終結している。

返答を要求する発話の場合、$PrS = R$ となり、話題を終結する発話の場合、$PrS = F$ となる。

$Ac(WhoAc)$,$InAc(WhoInAc)$:

$WhoAc$、$WhoInAc$ は、そのフェーズに属する発話を行なった参加者名を示す。

サフィックスの定義

C、Pr、Ac、$InAc$ にサフィックスを導入する[1]。$C_{x.y.z...}$ の $x.y.z...$ を、C のサフィックスとする。サフィックスの第 1 ラベル (x) をサブジェクト番号とする。サブジェクト番号とは、サブジェクト毎に割り当てた番号のことで、導入された順番 (1,2,3....) でナンバリングされる。y 以降のサフィックスの値は、コントリビューションが導入された順番を示す。Pr、Ac、$InAc$ のサフィックスは、それらによって構成されるコントリビューション

[1]文献 [3] では、コントリビューション、プレゼンテーションフェーズ、アクセプタンフェーズのサフィックスが定義されていない。

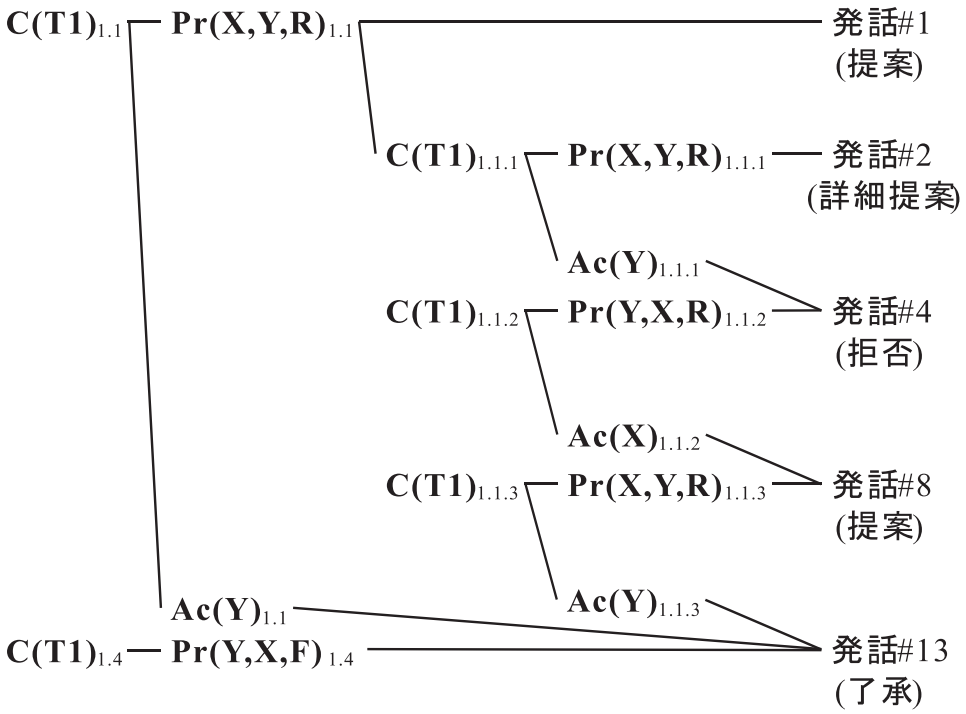

図 2.2: 討議構造木例

のサフィックスと同一である。

討議構造木例

図 2.2 に、討議構造木におけるパターン 1 の構造例を示す。この例では、参加者 X による発話#2 が、発話#1 の詳細を表しており、発話#2 は、参加者 Y に返答の要求を示唆している。続いて、参加者 Y の発話#4 が行なわれ、発話#1 に関する話題のサイドシーケンスが形成されている。参加者 Y による発話#13 により、サイドシーケンスと話題 $T1$ が終結している。発話#数字の# 数字は、対象とする討議において、発話された順番を表している。

2.2 XML による討議構造の表現

電子メール群に、討議構造木に関する情報を付与するための XML ボキャブラリ「UMML+Linkbase」を定義する。

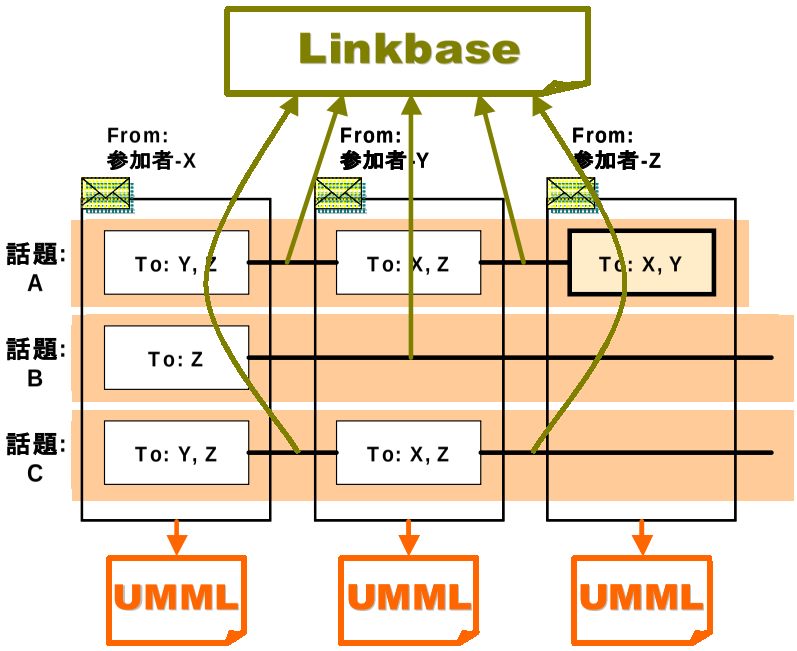

図 2.3: UMML+Linkbase

2.2.1 UMML

XML 仕様に準拠したマークアップ言語で、討議構造木中に現れる発話に関する情報 (Pr フェーズの属性など) を電子メールに付与するための言語 UMML(Utterances-in-Mail Markup Language) を定義した。図 2.3 に示すように、一つの UMML 文書に対しては、一つの電子メールが対応する。なお、別のメールに含まれる発話との接続関係は表現しない。

2.2.2 Linkbase

XLink とは、XML 文書のリンク機能に関する規約であり、リンク文書のリンク機能を規定するための名前空間と属性、制約を与えるものである [7]。図 2.3 に示すように、異なるメールに含まれる発話間の接続関係を、リンクベースに格納したサードパーティリンクとして表現することとし、UMML に対するサードパーティリンクを集めたリンクベースを記述するための XML ボキャブラリ Linkbase の定義を行なった。

2.2.3 UMML+Linkbase による討議構造木表現例

UMML+Linkbase による討議構造木表現の例を、図 2.4 に示す。

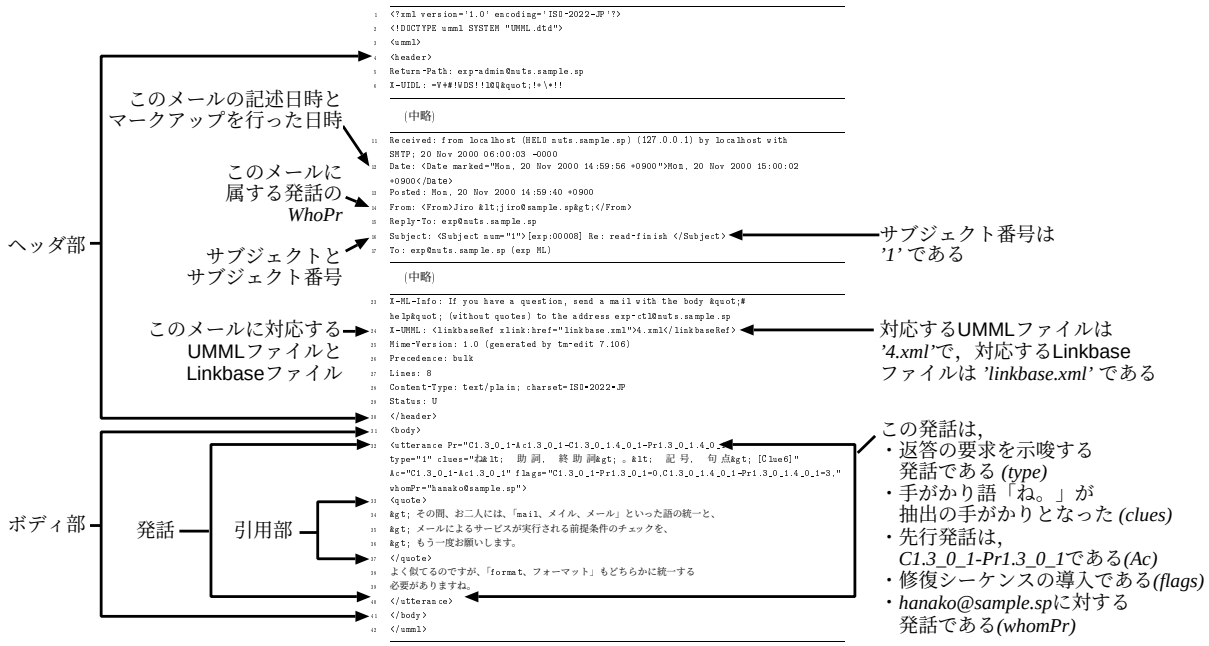

図 2.4: 討議構造木表現例

2.3 討議構造木抽出エンジンの概要

討議構造木抽出エンジンでは、電子メールを用いたコミュニケーションに現れる構造/言語的特徴を手がかりとして利用している。

2.3.1 電子メールボディ部の特徴

返答メールでは、引用文が利用されているという特徴がある。引用文とは、各行の先頭に “》”、“>” といった引用符を伴ってあらわれる文章であり、以前の議論であらわれた文章を引用している文章のことである。引用文は、引用文の直後にあらわれる発話の先行発話を示すことが多い。

また、メッセージの最初は、自分の名前や身分を示す宣言的な発話であることが多く、メッセージの最後は、一般に “署名” と呼ばれるものであることが多い。署名は、自分の名前や身分や連絡先などを示すもので、直前に “---”、“-=-=-=” といった記号だけからなる行 (セパレータ) を伴うことが多い。

電子メールを利用したコミュニケーションでは、対面での対話と異なり、会話の一時停止などを利用した話題転換の指示ができない。その代わりに、文章中に意図的に空行や改段落を挿入することによって話題転換の指示を行なっているものと考えられる。したがって、空行や改段落で区切られた文章が発話を構成する基本単位とする。ただし、発話の最低単位は一文とする。

示唆する特徴	略号	例
話題の並列展開	Clue#1	列挙します，挙げます
話題の転換	Clue#2	ところで，〜については
接続語による結束性 (詳細の関係)	Clue#3	具体的には，例えば
接続語による結束性 (前提の関係)	Clue#4	それで，だから
指示語による結束性	Clue#5	この場合，これは
返答の要求	Clue#6	か，しましょう，？
同意・了承	Clue#7	了解です，確に
対となる表現	Clue#8	〜と思いませんか—思います

表 2.1: 言語的手がかりの一覧

2.3.2 文章表現の特徴

先頭文の文頭や文末などの表現に、話題の転換/終結を示唆したり、文章末に、返答の要求を示唆する表現が含まれていることがある。このような特徴を抽出手がかりとした。表 2.1 に、その手がかり語をまとめる。

新たな話題の並列展開を示唆する表現

相手からの返答を求めない発話の中には、新たに議論される複数の話題を明示的に示しているものがある。例えば、「挙げておきます」のような表現は、続く空行や段落で分割される文章を、それぞれ異なった話題の最初の発話とするための手がかりになることが予想される。そこで、これらの表現を表 2.1 中の Clue#1 とする。

話題の転換を示唆する表現

会話における話題の転換では、それまでの話題を終始、あるいは停止し、新たな議論の開始を示唆する表現が用いられることがある [8][9]。同一メール中の隣接した発話間においても、発話の先頭文中に、話題転換を示唆する文副詞、接続詞があることがある。また、明示的に新たな話題の開始を示唆することによって、話題の転換を示唆することもある。これらの表現を、表 2.1 中の Clue#2 とする。

結束性を示唆する表現

文章の長さや読みやすさを考慮して、空行や改段落を用いて一つの話題の文章を分割することがある。その結果、メール中の空行や改段落によって分割された文章同士でも、同一の話題に言及している場合がある。そのような文章中には、後続の文章中に先行する文章との間の結束性を示唆する指示語や接続語が現れる。結束性とは、文や発話の間の「言語的なつながり」である [10]。結束性を示唆する接続語は、「具体的には」のように後続文章が先行文章の詳細に当たることを示唆する接続語群と、「だから」のように先行文章

が後続文章の前提に当たることを示唆する接続語群の２つのカテゴリに分類できる。これらの表現をそれぞれ、表 2.1 中の Clue#3、Clue#4、Clue#5 とする。

返答の要求を示唆する表現

返答を要求する発話には、相手への質問や提案などを示す語、つまり相手への返答の要求を示唆する表現が含まれていることが多い。これらの表現を表 2.1 中の Clue#6 とする。

同意・了承を示唆する表現

話題を終結する発話の場合、文章の先頭文の文頭あるいは文末に、相手の発話を理解し、同意を示す表現が含まれていたり、先頭文の文頭に感嘆を示す語が含まれる傾向にある。その上、先頭文以外の文末に、返答の要求を示唆する表現 (Clue#6) が含まれない傾向にある。これらの表現を表 2.1 中の Clue#7 とする。

対を構成する表現

発話ペアを構成する先行発話の最後の文末と後続発話の先頭文には、相手への問いかけと、それに対応した同意を示す表現の対が存在することがある。これらの対は、発話ペアを見つけるための手がかりになるだけでなく、後続発話が話題を終結する発話であるかを決定する手がかりにもなると思われる。これらの表現を表 2.1 中の Clue#8 とする。

2.4 討議構造木抽出エンジンの評価

シンポジウム開催のための事務局メーリングリストで行なわれた討議の一部を抽出エンジンを評価するためのデータとし、討議構造木抽出エンジンの出力結果の評価を行なった。抽出エンジンは、発話候補文章の切り出し、および発話候補文章からの発話抽出が基本機能である[2]。なお、このデータは 36 通のメールからなる。

2.4.1 発話抽出の精度

異なる発話を同一の発話としていないか、同一の発話を異なる発話として分割していないかを評価するために、以下に示す再現率の値を求めた。正解発話とは、人手によって、返答を要求する発話、あるいは話題を終結する発話として抽出された発話のこととする。

$$\text{再現率}\,(\%) = \frac{\text{抽出した正解発話数}}{\text{正解発話数}} \times 100 \tag{2.1}$$

抽出エンジンの出力結果を分析したところ、正解発話数が 57 で、抽出した正解発話の数が 37 となり、再現率は 64.9%となった。この低い評価結果となった主な原因は、抽出エンジンの基本機能である発話候補文章の抽出法にあった。そこで、2.5 節では、その抽出法の問題点を洗い出し、抽出エンジンの精度向上を行った結果について述べる。

2.5 精度向上の実現

本研究では、討議構造木抽出エンジンから得られる討議スレッドを要約の対象としているため、討議構造木抽出エンジンの精度を向上させる必要がある。そこで、すべての問題点を洗い出し、発話候補文章の抽出法の改善を行なった結果について述べる。

2.5.1 問題点の洗い出し

討議構造木抽出エンジンの出力結果と人手による抽出結果との比較を行い、以下に示す４つの例のような抽出間違いを見つけ出した。

例１

Internet 上での PR は、どなたがやっていただけるのでしょうか？ － PC Member への配信と論文集めの依頼： これは C 先生 & D さん？ － ML への配信：これは E がやります。 －関連 News Group への Post（１回かぎりではなく、時間をおいて何回か）： これはどなたが？

[2]電子メールを利用したソフトウェア仕様のレビュー作業におけるコミュニケーションから、発話抽出に利用するための言語的手がかりをあらいだしている [4]。

例１のように、箇条書きにより文章がインデントされる場合、抽出エンジンは、「－ PC Member への配信と論文集めの依頼：」の行と、「これは C 先生 & D さん？」の行を２つの発話候補文章として抽出してしまう。

これは、発話候補文章の切り出しの前に、メッセージ部を整形処理をすることにより、回避できる見込みがある。

例２

```
＞部屋の大きさ、分割の可不可を、投稿していただけますか？

    現在予約しているのは、
    大集会室（机と椅子を入れて、160 名収容可） 会議用
    第１会議室（12 人） 事務局用
    です。

    あと借りれそうなのは、
    第５, ６会議室 (計 60 人） 会議用
    第３会議室（14 人） ツール展示用
    です。

    事務局用とツール展示用を交換しても良いかもしれません。
```

例２のような文章を読み易くするために、文章が均等にインデントされる場合、抽出エンジンは、空行を認識せずに「現在予約しているのは、」の行から「事務局用とツール展示用を交換しても良いかもしれません。」の行までを１つの発話候補文章として抽出してしまう。

これも、例１同様に発話候補文章の切り出しの前に、メッセージ部を整形処理をすることにより、回避できる見込みがある。

例３

```
部屋の大きさ、分割の可不可を、投稿していただけますか？
＞ G さん
```

例３の場合、“＞ G さん” は、話題をなげかけている討議参加者を示しているが、抽出エンジンは、引用文として識別してしまう。

発話抽出の際に、同一話題の文章を捉える手がかりとして、結束性を示す指示語を利用している。しかし、以下に示す例では、文中の指示語が同一文中にある語を指示しているため、話題の分割を誤っている。

例４

```
＞第５, ６会議室を借りますか？

借りて下さい。

期間をどうするか（１日目からやるか、２日目だけにするか)、
ツール展示募集をどうするか（CFP に載せるならその文面)、
展示者からの料金はどうするか、
など、原案をお願いします。＞ F さん
```

例４の場合、抽出エンジンは、「CFP に載せるならその文面」の「その」が、「借りて

下さい。」の文中の語を指示すると判断し、同一発話（話題）として抽出してしまう。

例３、４の抽出間違いは、いづれも頻度が少ない。例３では、引用文の言語的手がかりの充実、例４では、文脈処理などが必要になるものと考えられる。

2.5.2 発話候補文章の抽出法の改善と評価

2.5.1 節の問題点の洗い出しによって、討議構造木抽出エンジンの改善ポイントは、発話候補文章の抽出法の見直しであると考えられる。

改善前の討議構造木抽出エンジンでは、以下の４つを手がかりに発話候補を切り分けており、

- 引用符
- (署名と本文との間の) セパレータ
- 空行
- 行頭の空白

この４つの手がかりだけでは、箇条書きの文章や均等にインデントが含まれた文章に対する処理がなされていない。そこで、発話候補の切り出し手がかりを追加し、発話候補文章の抽出法の改善を行なった。以下に追加した発話候補の切り出し手がかりを示す。

- 箇条書きの文章 (2.5.1 節 例１) に対する改善方法

  行頭に「・」や「●」、「ー」などの記号を伴って現れる箇条書きの文章に対して、各箇条ごとに１つの発話候補文章として切り分けるように修正を行なった。

- 均等にインデントが含まれた文章 (2.5.1 節 例２) に対する改善方法

  行頭が空白であり、直前の行も行頭が空白であれば、同一の発話候補文章とみなし、空行までの行を１つの発話候補文章として切り分けるように修正を行なった。

改善の結果、正解発話数が 57 で、抽出した正解発話の数が 40 となり、再現率 (2.4.1 節の式 (2.1)) は、70.2%となった。改善前の再現率が 64.9%であったので、5.3%増加したことになる。図 2.5 では、実験データ中の各 subject における改善前と改善後の抽出した正解発話数を示す。

このような低い増加率となった原因は、今回行なった改善において、以下に示す例では改善されていないためであると考えられる。

例１ (2.5.1 節 例１と同様)

Internet 上での PR は、どなたがやっていただけるのでしょうか？
　－ PC Member への配信と論文集めの依頼：
　　これは C 先生 & D さん？
　－ ML への配信：これは E がやります。
　－関連 News Group への Post（１回かぎりではなく、時間をおいて何回か)：
　　これはどなたが？

	正解発話数	抽出した正解発話数
subject#1	16	15
subject#2	12	5
subject#3	6	5
subject#4	11	3
subject#5	12	9
合計	57	37

改善前の結果

	正解発話数	抽出した正解発話数
subject#1	16	15
subject#2	12	6
subject#3	6	5
subject#4	11	5
subject#5	12	9
合計	57	40

改善後の結果

図 2.5: 討議構造木抽出エンジンの精度向上前後の比較

例１では、発話候補文章の抽出において、各箇条毎に１つの発話候補として抽出するが、発話抽出の際に、「— ML への配信：これは E がやります。」の行の「これ」が、「— PC Member への配信と論文集めの依頼：これは C 先生 & D さん？」の文中の語を指示すると判断し、同一発話として抽出してしまう。これは、発話抽出の際の結束性を示す指示語による抽出誤りである。

例２

```
＞もしツール展示を行なうとすると、どの部屋を会場にするか等の
＞問題が残っています。

    返事を書こうとしたら、I さんから三点のツール展示の
    意向が示されました。
      この調子で考えれば、20 セット位になりそうです。
    従って、第 5,6 会議室 (計 60 人) を借りればいいと思いますが、
    実行委員長をはじめとした皆さんのご意見はいかがでしょうか。
```

例２では、均等インデントが含まれた文章中に、発話候補の切り出し手がかりである行頭の空白 (「この調子で考えれば、20 セット位になりそうです。」の行) が存在していても、それを認識せずに「返事を書こうとしたら、I さんから三点のツール展示の」の行から「実行委員長をはじめとした皆さんのご意見はいかがでしょうか。」の行までを１つの発話候補文章として抽出してしまう。

今後も、多種多様のデータを解析することにより、討議構造木抽出エンジンの精度向上を行なう必要がある。

第3章　討議構造参照機能を有するメールクライアントの問題点の洗い出し

本章では、本研究室において実現された討議構造を呈示するためのメールクライアントについて紹介する。また、メールクライアントの呈示法における問題点を洗い出し、それについて述べる。

3.1　システム構成

討議構造木が有する情報をユーザーに視覚呈示することにより、円滑なコミュニケーションの支援を目指す。構造木抽出エンジンが出力するUMML+Linkbaseファイル群を入力とし、討議の道筋を呈示するソフトウェアのプロトタイプとして、ICEMail++を開発した。図3.1にその画面例を示す。これは、Javaで記述されたメールクライアントICEMail[11]を拡張し、討議構造呈示部を追加したものである。討議構造呈示部は、討議構造の表形式およびツリー形式呈示機能を有する。この2つの呈示機能は、ユーザーによる選択可能となっている。

3.2　討議構造呈示機能

表形式

表形式呈示部では、ユーザーがメール一覧から選択したメール中に含まれる発話を、表形式で呈示する。抽出された発話は、導入順でセルに格納し、発話の種類をセルの背景色で区別している。また、引用部分から披引用部分をリンクづけることにより、対象とした発話の先行発話が、披引用部分から引用部分をリンクづけることにより、対象とした発話の返答発話をたどることができる。図3.1に示す討議構造呈示部は、表形式呈示を行なっている。

ツリー形式

ツリー形式呈示部では、ユーザーがメール一覧からメールを選択することにより、選択したメール中の発話が含まれる討議のスレッドを自動配置し、討議構造木に基づく討議構造をツリー状に呈示する。図3.2に示すように、1つのフレームは1つの発話に対応し

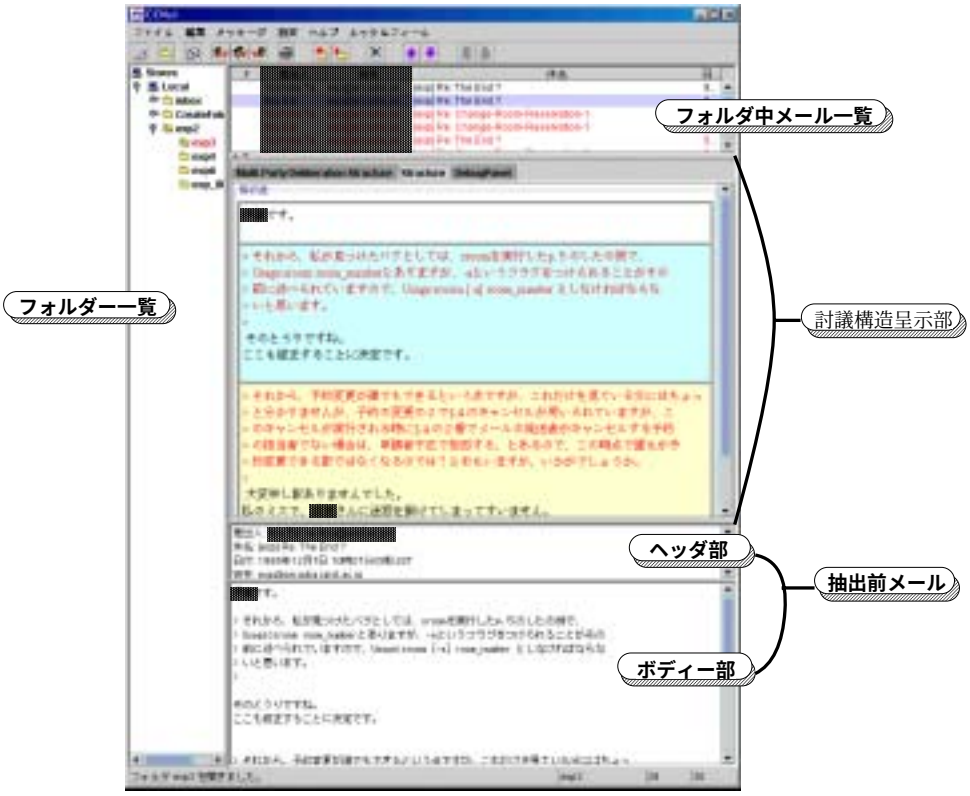

図 3.1: ICEMail++

ており、フレーム間を矢印で結び、討議のスレッドを表現している。フレーム内の背景色は、発話の種類を示し、フレームを選択することにより、その発話が行なわれたメールを参照することが可能である。宣言的な発話は白色に、返答を要求する発話は黄色、話題を終結する発話は青色で表現している。この視覚呈示により、例えば、ある話題の議論が進んでいない場合に、自分が返答を要求されているにもかかわらず、未返答であることが原因であると認識できるという効果が考えられる。

討議構造木では、返答を要求されている参加者による返答と、そうでない参加者による返答を区別して表現することができる。図 3.3 に示すように、返答を要求されている参加者 (聞き手) による返答には、青実線の矢印で結び、返答を要求されていない参加者 (聞き手) による返答には、赤実線の矢印で結ぶことにより表現する。

討議構造木では、同一話題中に、提案 (質問) とその提案 (質問) に対する詳細の提案 (質問) が含まれている場合 (パターン 1) や、提案 (質問) とその提案 (質問) の前提となる提案が含まれている場合 (パターン 2) の討議の構造を表現することができる。図 3.4 に示すように、パターン 1 の場合は、緑破線の矢印で発話間を結び、パターン 2 の場合は、赤破線の矢印で発話間を結ぶことにより表現する。

なお、各々発話は、返答を要求されている参加者や、発話抽出の手がかりとなった語彙情報を保持しており、図 3.4 に示すように呈示可能である。

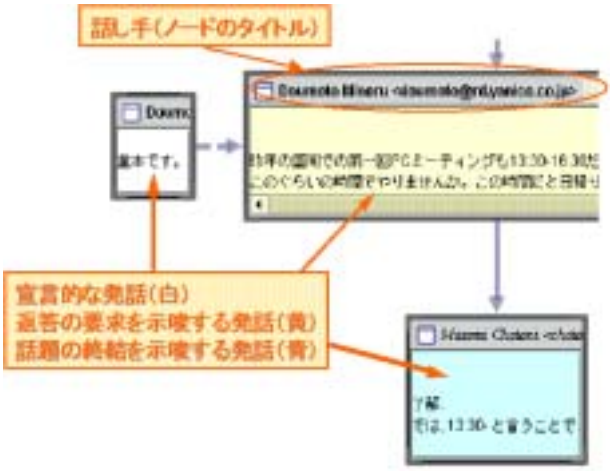

図 3.2: ツリー形式呈示部 (発話の種類)

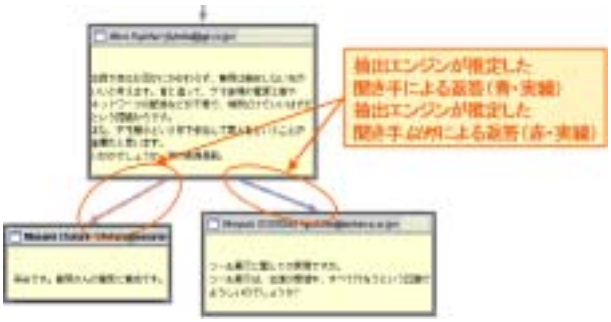

図 3.3: ツリー形式呈示部 (返答発話の種類)

3.3 呈示法における問題点

討議構造呈示部では、フォルダ中メール一覧にある一つのメールに含まれる発話同士の関係や、その発話が含まれる討議のスレッドを呈示する。しかし、選択した以外のメール中に含まれる発話に関する情報を知るためには、フォルダ中メール一覧に戻ってメールを再選択するという手間が必要になるため、思考の流れの妨げになる。図 3.5 では、討議の構造の模式図を用いて、討議構造呈示部で呈示される討議構造の領域の例を示している。この例では、選択されたメール中に二つの発話が含まれており、それぞれの発話が含まれる二つの討議のスレッドのみを討議構造呈示部で呈示することになる。

以上のことから、この呈示法では、討議の概観を一目で見ることができないといった問題がある。そこで、自然言語処理の分野で研究されているテキスト自動要約技術を用いて、すべての討議内容の重要箇所を呈示し、また、利用者が求める討議構造も呈示できれば、円滑なコミュニケーションの促進や効率の良い共同作業の実現が期待される。

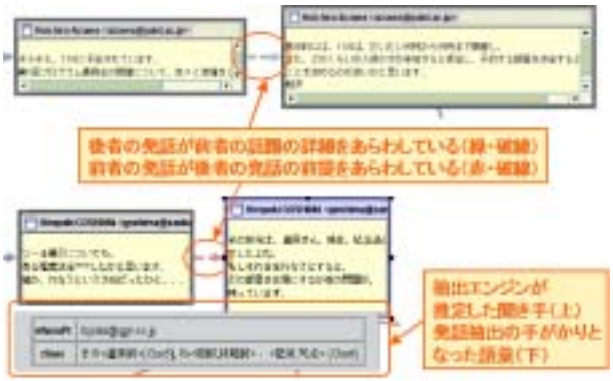

図 3.4: ツリー形式呈示部 (同一話題中の複数発話の表現)

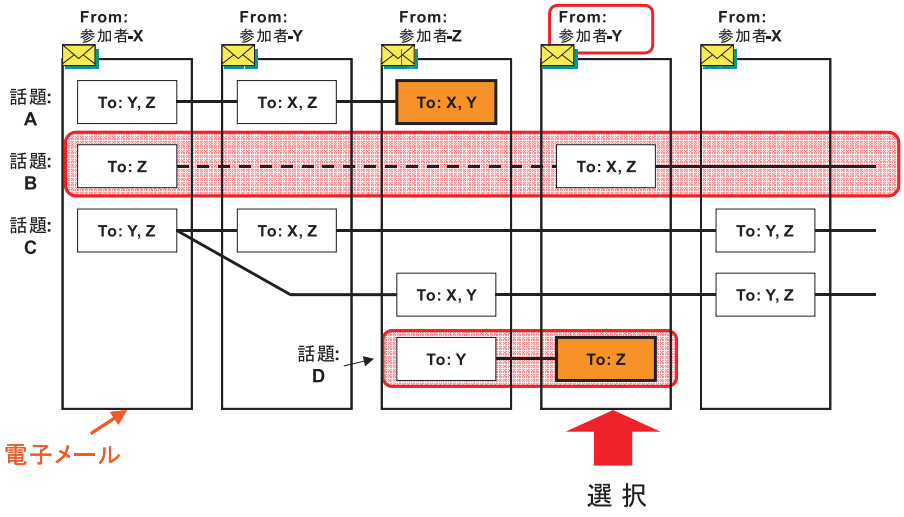

図 3.5: 討議構造呈示部の呈示領域の例

第4章　討議内容の要約手法の考案

本研究では、討議構造木抽出エンジンから得られる UMML+Linkbase ファイルを入力とする討議内容の自動要約システム（討議内容要約エンジン）を実現した。討議内容要約エンジンの設計と実現を行なうにあたり、討議内容の要約手法を考案する必要がある。そこで、自然言語処理の分野におけるテキスト自動要約技術に関する調査を行ない、討議内容の要約手法として効果的な手法を考案した。

4.1　テキスト自動要約技術

自然言語処理の分野において、読み手にテキストの内容を容易に把握させるという目的で、テキスト自動要約技術に関する研究が行なわれている。重要文抽出による要約手法や、近年では、抽象化、言い換えによる要約手法が考案されている [2]。テキスト要約は、元の文の大意を保持したまま、テキストの冗長さ、複雑さを減らす処理といえる。その過程は、大きく３つの項目に分けられる。テキストの解釈（文の解析とテキストの解析結果の生成)、テキスト解析結果の要約の内部表現への変形（解析結果中の重要部分の抽出)、要約の内部表現の要約文としての生成である。これまでのテキスト自動要約研究の多くのものは、テキスト中の文を一つの単位とし、それらに何らかの情報を基に重要度を付与し、その重要度で順序付け、重要な文を選択するという手法がとられている。この重要度評価の際に用いられるテキスト中の情報は、Paice により次の７つに分類されている [12]。

(1) テキスト中のキーワードの出現頻度

(2) テキスト中あるいは段落中での位置情報

(3) テキストのタイトル等の情報

(4) テキスト中の文間関係を解析したテキスト構造

(5) テキスト中の手がかり表現

(6) テキスト中の文あるいは単語間のつながりの情報

(7) テキスト中の文間の類似性の情報

分類された７つの情報それぞれについて従来研究を紹介する。

(1) テキスト中のキーワードの出現頻度

テキスト中によく出現する内容語はテキストの主題を示す傾向があるとの仮定が情報検索分野などではしばしば用いられる。この仮定に基づき、テキスト中で出現頻度の高い名詞をキーワードと考えたり (tf 法)、また、これに合わせて、出現するテキスト数も考慮することで、そのテキスト固有の出現の度合いを計算したり (tf*idf 法) など、情報検索分野では、様々な単語の重み付け技法が用いられている。単語の重要度を基に、文自体に重要度を付与するという重要文抽出手法が、1950 年代から提案されている。

(2) テキスト中あるいは段落中での位置情報

テキストはジャンルに依存して、ある程度構造に規則性を持っている。学術論文では、序論、本論、結論のような構造を持つ。新聞は、見出し、小見出しの後に、本文が来ることが多い。テキスト中での文の位置情報をその文の重要度計算に利用する手法がいくつか考えられている。新聞記事からの重要文抽出では、本文の先頭数文を抽出する lead 手法と呼ばれる方法が良いとされている。

(3) テキストのタイトル等の情報

ジャンルにより決まったテキストの構造から得られる、もう一つの情報として、本文以外に、テキスト中に付与されたタイトル、見出しの情報がある。学術論文の場合は、テキスト自体がタイトルを持つ場合もあり、また、各章、節にもタイトルが付与されることが多い。また、新聞には、見出し、小見出しが本文とは別に付与されることもある。タイトル、見出しに現れる内容語を含む文が重要であると考え、タイトル、見出し中の単語を重要文抽出に利用する手法が提案されている。

(4) テキスト中の文間関係を解析したテキスト構造

自然言語処理の分野では、テキスト中の接続詞等の手がかり語情報などを基に、文間の構造を解析し、テキスト構造を得る研究がいくつか見られる。このようにして得られたテキスト構造を利用して重要文を抽出する研究が行なわれている。

(5) テキスト中の手がかり表現

テキスト、文の主題を表す内容語ではないが、テキスト中の重要箇所を指示すると考えられる手がかり表現がいくつか存在する。学術論文などでは、'this report'、'in conclusion'、'our work' などの表現は、論文の主題を表す文中に出現すると考えられる。このような手がかり表現を利用して、テキスト中の重要文を抽出する研究も存在する。これとは逆に、重要文と負の相関関係にあると考えられる手がかり語を考慮することもできる。「たとえば」などの例示を示す接続語で始まる文は重要度が低いと考えられるのはその一例である。

(6) テキスト中の文あるいは単語間のつながりの情報

テキスト中の文間のつながりの情報を重要文抽出に利用する手法を紹介する。

Skorokhod'ko は、文をノード、文間の関係をリンクとするグラフでテキストを表現し、多くの文と関係のある文が重要であるという考えに基づき、重要文を抽出する手法を示している [13]。Halliday と Hassan は、表層的な文間のつながりを表す指標として、５種類の結束性 (cohesion)、すなわち、指示 (reference)、代入 (substitution)、省略 (ellipsis)、接続 (conjunction)、語彙的結束性 (lexical cohesion) を挙げている [14]。語彙的結束性とは、語の意味的なつながりによって文と文をつなぐことである。Hoey は、この語彙的結束性の情報を利用し、文間で単語によるつながりが多いほど、文間のつながりが強いと考え、他の文とのつながりの強さに基づき、要約を作成する手法を示している [15]。

(7) テキスト中の文間の類似性の情報

テキストを、その中に出現する単語の重みのベクトルとして表現することが多い。テキスト間の類似度は、テキストを表現するベクトル間の内積などで計算することができる。テキスト中の文 (段落) を１単位として、それらの間の類似度を計算し、この類似度を文 (段落) 間のつながりの度合と考え、重要と考えられる文 (段落) を抽出する手法が提案されている。亀田は、２文間にどれほど共通な単語 (キーワード) が現れるかに基づいて計算した文間関連度 (の平均) と、ある文が他の文とどの程度広く関係があるかに基づいて文の重要度を計算する手法がある [16]。

4.2　本研究における要約手法の方針

本研究では、討議構造木抽出エンジンから得られる討議スレッドから、討議内容を要約するための手法を考案する。本研究において討議内容の要約とは、話題が終結した討議内容の要約として、**議題**と**それに対する結論となる情報**を抽出し、話題が未終結な討議内容の要約として、**議題**と**最終提案となる情報**を抽出することである。議題となる情報とは、討議内容において「何が、どうなったか」という情報の「何が」に相当するもので、例えば、ある会議の開始時間を決める討議で開始時間が 13:00 からとなった場合、「ある会議の開始時間」が議題となる情報である。それに対する結論となる情報とは、「どうなったか」という情報であり、先の例では、「13:00 から」が結論となる情報である。また、話題が未終結な討議内容における最終提案となる情報とは、最後に提案した情報のことで、先の例で未終結な討議とした場合、「13:00 からでいいですか？」が最終提案となる情報である。

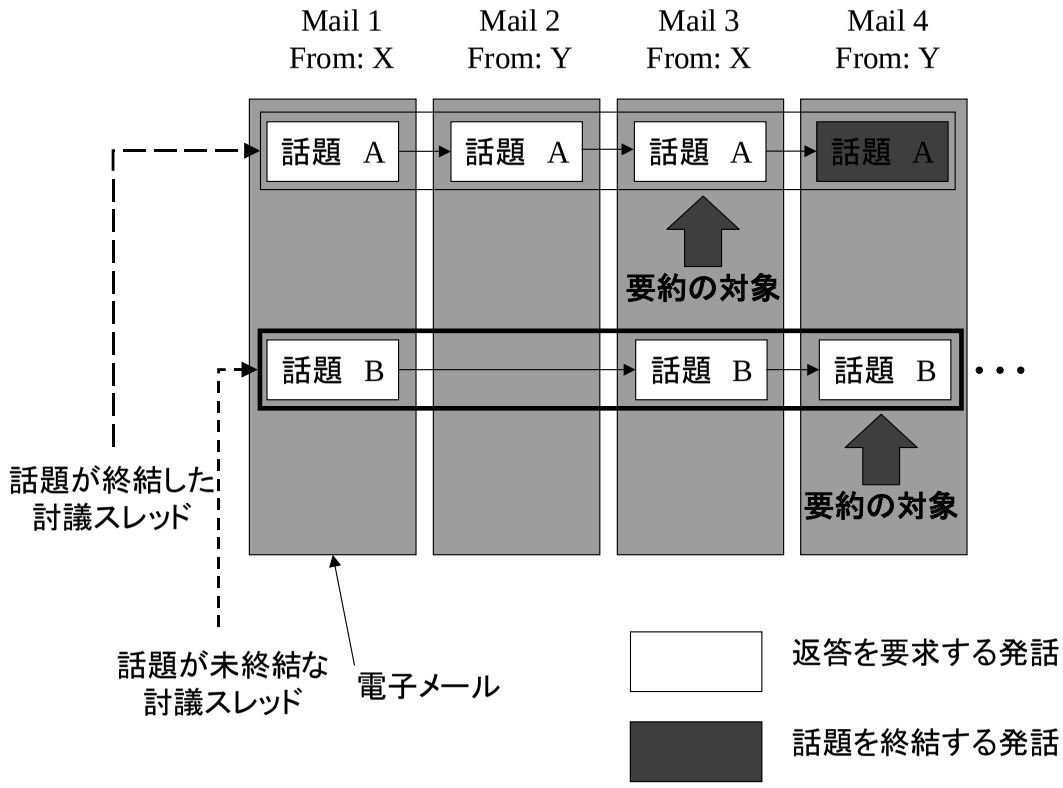

図 4.1: 討議内容の要約対象

4.3 手がかり表現を利用した要約手法の考案

討議内容の要約においては、「テキスト中のキーワードを用いた出現頻度」や「テキスト中あるいは段落中での位置情報」を利用した重要文の抽出は、効果的な手法とは言えない。なぜなら、討議内容中のキーワードと考えられる subject は、話題の異なった討議内容に対して、同じキーワードとなるからである (例えば、図 4.1 における Mail1 の話題 A と話題 B)。また、討議内容は、発話の連接ペアによる会話で構成されているため、テキストにおける起承転結などの位置情報を利用することも効果的ではない。

そこで、「テキスト中の手がかり表現」を用いて、討議内容の要約手法を考案した。その手法とは、手がかり表現を利用して、重要度が低いと考えられる文や不要な接続詞の削除である。重要度が低いと考えられる文として、「例えば」などの例示を示す接続語で始まる文を用いた。また、不要な接続詞として、「従って」や「ということで」などの文頭に現れる接続詞を用いた。

以下に話題が終結した討議内容、及び未終結な討議内容の要約手法の手順を示す。

1. 話題が終結した討議内容に対して、了承の意図を表す返答が行われているメールの引用文を要約の対象とし、話題が未終結な討議内容に対しては、最終提案を要約の対象とする (図 4.1　要約の対象部)。
2. 手がかり表現を利用して、重要度が低い文や不要な接続詞を削除することにより要約文を作成する。

4.3.1　評価

シンポジウム開催のための事務局メーリングリストで行なわれた討議の一部を、討議内容の要約手法を評価するためのデータとし、評価を行なった。評価値として、以下に示す正解率を用いた。

$$\text{正解率 (\%)} = \frac{\text{要約手法に適した討議スレッド数}}{\text{討議スレッド数}} \times 100 \tag{4.1}$$

要約手法に適した討議スレッド数とは、自然な要約文 (文間のつながりが悪く要約の内容が理解できない要約文や主語が含まれていない要約文などは、自然な要約文ではない) が得られた討議スレッド数とし、討議スレッド数とは、討議構造木抽出エンジンの出力結果で、正解発話によって連鎖した討議スレッドの数とする。図 4.2 に自然な要約文の例を示す。

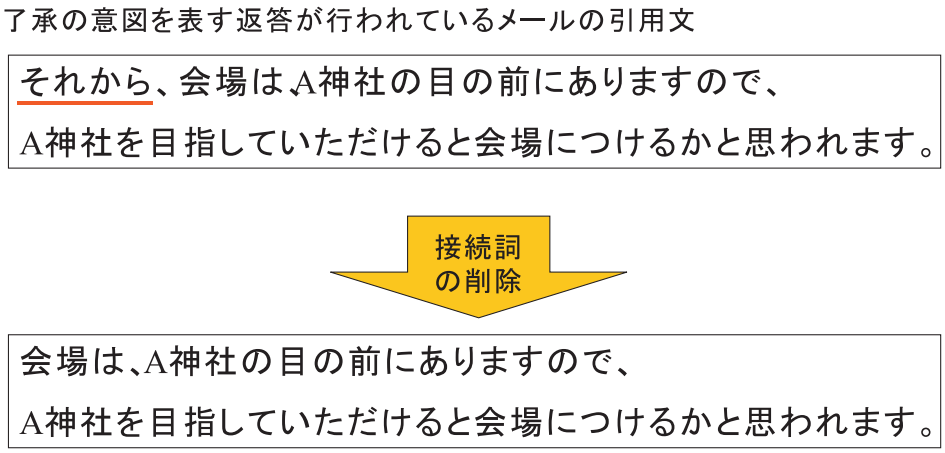

図 4.2: 自然な要約文の例

この要約手法では、討議スレッド数 29 で、要約手法に適した討議スレッド数が 11 となり、正解率が 37.9%であった。

4.3.2 問題点

このような低い評価結果となった主な原因は、要約に必要な情報が要約の対象外の発話に含まれていることである。以下に例を示す。

- 了承の意図を表す返答が行われている発話中に、議題の結論となる情報が含まれている。

  例

了解しました。 では、13:30 から会議を行います。

- 引用文を利用する際に、議題となる情報が含まれていない。

  例

＞ 午前中仕事があるので、 ＞ 15:00- にはできないでしょうか？

- 討議スレッド中で、話題が転換した場合に議題となる情報の抽出が困難である。（例えば、初期の提案の一部分だけが拒否された場合、初期の提案と新たな提案を組み合わせて議題となる情報を抽出しなければいけない。）

この問題点は、要約の対象が討議スレッドの終盤だけに限定しているために、それ以外の箇所にある必要な情報が欠けてしまっているためであると考えられる。

4.4 要約手法の改良案の方針

4.3.2 節の問題点より、要約の対象を討議スレッド全体とする要約手法を考案する必要がある。また、議題とそれに対する結論となる情報のどちらか一方または、議題と最終提案となる情報のどちらか一方が含まれていないという問題があり、これについては、要約の対象を１つの発話に限定した手法を改良する必要がある。

そこで、本研究では、要約の対象を討議スレッド全体とし、重要な発話の抽出手法を考案した。4.2 節で述べたように、要約手法の方針は、話題が終結した討議内容の要約として議題とそれに対する結論となる情報を抽出し、話題が未終結な討議内容の要約として議題と最終提案となる情報を抽出することである。この方針から、話題が終結した討議内容と話題が未終結な討議内容の重要発話の定義を以下に示す。

話題が終結した討議内容

- 議題となる情報を含んだ発話
- 結論となる情報を含んだ発話

話題が未終結な討議内容

- 議題となる情報を含んだ発話
- 最終提案となる情報を含んだ発話

この定義をもとに、4.2 節で述べた「ある会議の開始時間を決める討議」の例を当てはめた場合、議題となる情報を含んだ発話は「会議の開始時間は何時にしますか？」のような発話となり、結論となる情報を含んだ発話は「13:00 から始めましょう」のような発話となる。また、最終提案となる情報を含んだ発話は「13:00 からでいいですか？」のような発話となる。このように、各討議スレッドから 2 つの発話を抽出することで、討議内容の概要を理解することが可能である。よって、討議内容の要約として、各討議スレッドから 2 つの発話を重要発話として抽出する。

4.5 tf*idf 法を利用した要約手法の考案

討議スレッド中の重要発話とは、討議内容の特徴を表わしている発話であると考えられる。その特徴を表わしている発話を選び出す手がかりとして、その討議の内容を表わしている単語を用いることが考えられる。そこで、tf*idf 法を用いた単語の重み付け技法による討議内容の要約手法を考案した。この手法では、討議スレッド中の各発話毎に、その討議の内容をよく表していると考えられる単語 (索引語) を選択し、その索引語に tf*idf 重み付けを行ない重要発話を抽出する。

4.5.1 tf*idf とは

tf*idf とは、情報検索分野などにおいて、単語の重要性を計算する一手法である [17]。まず、単語の重要性を計算する上で、文章中の全ての単語を対象とするのではなく、文章の内容を表わしている単語 (索引語) を選び出す。次に、ある文章 d における索引語 t の出現頻度を求める。この値を tf(term frequency) という。また、索引語 t が全文章中のどれぐらいの文章に出現するかを求める。これを idf(inverse document frequency) という。この tf と idf をかけあわせた値をその文章の索引語の重みとする。この重みを tf*idf 重みという。以下に索引語の tf*idf 重み付けについて述べる。

索引語の tf*idf 重み付け

索引語の重み付けとは、抽出した索引語にその索引語の重要度を表わす尺度を与えることである。索引語の重みを考える場合に、索引語の網羅性と特定性という 2 つの性質を考慮する必要がある。すなわち、一般的な指針としては、網羅性と特定性を兼ね備えた索引語の重要度が上がるように重みを与える。

tf (term frequency)

網羅性は、索引語の頻度をもとに重みを計算することによって考慮することができる。ある文章 d 中に出現する索引語 t の頻度を索引語頻度 (term frequency) と呼び、*tf(t,d)* で表わす。この *tf(t,d)* を文章 d における索引語 t の重み w_t^d と考えることができる。索引語頻度に基づく重み付けの背景には、「何度も繰り返し言及される概念は重要な概念である」という仮定がある。

索引語の出現頻度に関して注意すべき点は、文章の長さと頻度の関係である。文章が長くなると平均的に語の出現頻度も多くなる傾向にある。従って、索引語の出現頻度を文章中の全ての索引語の出現数で割った相対頻度を重みとして採用することにする。

tf(t,d) は次式のように定義される。

$$tf(t,d) = \frac{t}{T} \tag{4.2}$$

t : ある文章 d 中に出現する索引語 t の出現数
T : 文章中の全ての索引語の出現数

idf (inverse document frequency)

索引語の頻度は、索引語の網羅性を高める上で貢献するものの、索引語の特定性については必ずしも役に立たない。*tf(t,d)* は、各文章内の頻度のみを考慮するにとどまり、ほかの文章の索引語の分布について考慮していない。ある索引語が、どの程度その文章に特徴的に現れるのかという特定性を考慮するためには、ほかの文章中の索引語の分布も考慮する必要がある。特定性を表わす尺度として idf が知られている。

idf は、ある索引語が全文章中のどれくらいの文章に出現するかを表わす尺度であり、次式のように定義される。

$$idf(t) = log(\frac{N}{df(t)}) \tag{4.3}$$

N : 検索対象となる文章集合中の全文章数
$df(t)$: 索引語 t が出現する文章数

tf*idf 重み付け

索引語の網羅性と特定性の両方を併せ持つように、２つの尺度を組み合わせて索引語の重みを計算する。すなわち、索引語 t の重み w_t^d は、次式のように定義される。

$$w_t^d = tf(t,d) * idf(t) \tag{4.4}$$

これを tf*idf 重み付けと呼ぶ。

索引語	出現数	索引語	出現数
11/5	2	予定	2
する	12	いる	2
プログラム	1	委員	1
開催	3	準備	1
具体	2	何時	2
いう	1	決める	1
良い	2	時間	3
思う	3	借りる	4
・	・	・	・
・	・	・	・

表 4.1: ある討議スレッドにおける索引語の出現数

名詞	プログラム, 人数, 部屋, 時間, ss2000, 会場, 会議, パーティション, PC, ツール, ミーティング, 担当, ISO, オープンソース, Java, Smalltalk, 企業 •••
動詞	決める, 思う, 空く, 外す, 出来る, 借りる, 分かる, 区切る, 行なう, 残る, 違う, 集める, 申し込む, 関わる, 考える, 寄せる •••
形容詞	良い, つらい, 狭い, 早い, よろしい •••

表 4.2: 索引語リスト

4.5.2 索引語の選択

索引語の選択は人手で行なうことも考えられるが、その場合、討議の参加者に対して索引語の情報提供を求めなければならない。本研究では、討議の参加者に負担を与えない討議内容の自動要約手法を考案することを目的とするため、自動による索引語の選択方法を考案する必要がある。

語には、特定の概念を表わす内容語と、語と語の間の関係を表わす機能語とがある。一般に、機能語は文章を特徴付ける上で役には立たない。内容語は、その全てが文章を特徴付ける上で役に立つかというと必ずしもそうではない。「ある」や「なる」などの一般的によく使われる語は文章を特徴付ける上で、あまり役には立たない。そこで、本研究では実験データをもとに索引語の選択を行なった。実験データは、「シンポジウム開催のための事務局メーリングリストで行なわれた討議の一部」を用いた。索引語の対象として、名詞、動詞、形容詞とする。機能上内容を表わしていないと考えられる代名詞、接尾語などは不要語とする。また、実験データ中に高頻度で出現した語「する」も不要語とする (表 4.1 参照)。索引語リストを表 4.2 に、不要語リストを表 4.3 に示す。

代名詞	私, それ, これ, あれ, 皆, どなた, なに, こちら, どちら …
接尾語	回, 会, 的, ぐらい, 市, 用, 場, 室, 様, 料, 群 …
高頻度語	する

表 4.3: 不要語リスト

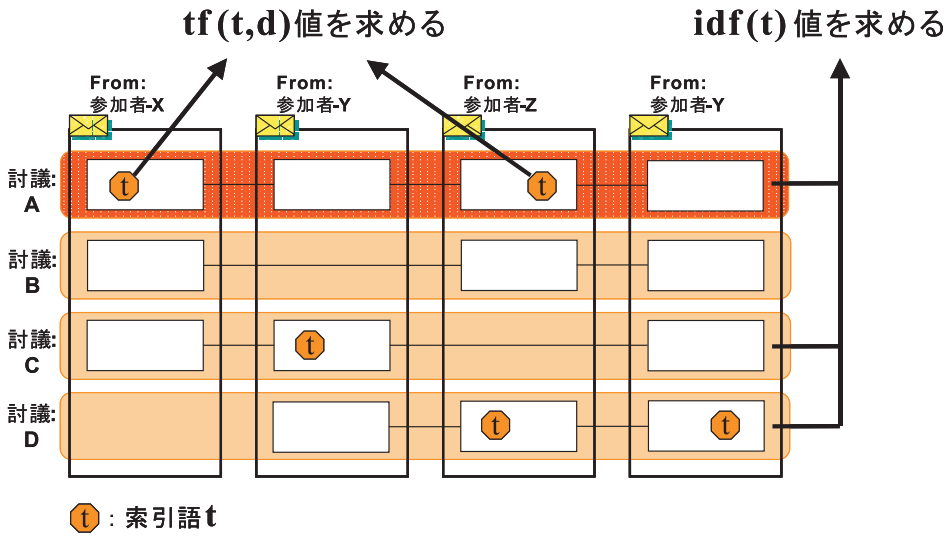

図 4.3: *tf(t,d),idf(t)* 値の求め方の概要

4.5.3 重要発話の抽出

本研究では、重要発話の抽出手がかりとして、各発話中の索引語に tf*idf 重み付けを行う。本研究における *tf(t,d)* の値の求め方は、4.5.1 節の式 (4.2) において、t はある討議スレッド d 中に出現する索引語 t の出現数、T は討議スレッド d 中の全ての索引語の出現数である (図 4.3 の場合、t の値は 2 となる)。また、*idf(t)* の値の求め方は、4.5.1 節の式 (4.3) において、N は検索対象となる討議スレッドの集合中の全討議スレッド数、*df(t)* は索引語 t が出現する討議スレッド数である。つまり、idf は、ある索引語が少数の討議スレッドにしか出現しない場合に大きくなり、どの討議スレッドにも出現すると最小の値になる (図 4.3 の場合、*df(t)* の値は 3 となる)。この *tf(t,d)* と *idf(t)* の値を組み合わせたものが、索引語の tf*idf 重み付けとなる (4.5.1 節の式 (4.4) 参照)。

各発話中の索引語の tf*idf 重み付けをもとに討議内容から重要発話として 2 つの発話を抽出する。重要発話とは、4.4 節で述べた情報を含んだ発話であり、また、簡潔にまとまった発話のことである。そこで、索引語の tf*idf 重み付けにより得られる *tf(t,d)*idf(t)* 値を用いた重要発話の抽出法として、以下の 3 つの方法を検討した。

方法 1 最も高い *tf(t,d)*idf(t)* 値の索引語が多く含まれている発話を抽出する。

方法 2 発話毎に高い *tf(t,d)*idf(t)* 値の索引語をある基準によって取り出し、その取り出した複数の索引語の *tf(t,d)*idf(t)* 値をたしあわせた値が高い発話を抽出する。

方法 3 各発話中の *tf(t,d)*idf(t)* 値の合計値を各発話中の索引語数で割った値が高い発話を抽出する。

高い *tf(t,d)*idf(t)* 値の索引語は、討議スレッド中の出現頻度が高いという特徴を持っている。この特徴から、方法１では、ある話題の詳細を記述した発話のように発話内の文章が長い発話を抽出すると考えられる。実験データをもとに方法１を適用してみた。以下に示す例は、実験データ中のある討議内容である。

発話 1
A:そろそろ，11/5 に予定されています,
　第 1 回プログラム委員会の開催について，色々と準備をしてはいかがでしょうか?

発話 2
A:具体的には，11/5 は，だいたい何時から何時まで開催し,
　また，どのくらいの人数の方が参加すると仮定し，予約する部屋を決定するといったことを決めるのが良いかと思います.

発話 3
D:時間は、13:30-18:00 ぐらいではないでしょうか。
　人数は、最多で 40 名と思います。

発話 4
G:今のところ，ss2000 の開催予定会場の方の会議室は空いてるそうです.
　30 名収容の会議室が 2 つあって，2 つの会議室はパーティションを外せば,
　1 つの会議室に出来そうです.
　もし 40 名になりますと，2 つの会議室を借りなければならないです.

発話 5
D:意味が良くわからないのですが...
　もう少し具体的に説明してもらえないでしょうか?

発話 6
G:わかりづらい表現ですいません.
　2 つの会議室があって，その会議室はパーティションで区切られているので,
　それを外すと，1 つの会議室として機能します.
　利用する側からすると，1 つの会議室として問題ないです.

借りる際に，2つの会議室として借りなければならないという意味です.

発話7
D:了解しました．それでは，その2つの会議室を借りてください.
それでよろしいでしょうか? ＞ C先生

発話8
C:会議室の予約については，それでOKです.

実験データ中のある討議内容の例

この例では、最も高い *tf(t,d)*idf(t)* 値の索引語は「会議」であり、方法1による重要発話の抽出を行なうと、発話4と発話6が抽出される。このことにより、方法1では、ある話題の詳細を記述した発話のように発話内の文章が長い発話を抽出する傾向にある。よって、方法1は、議題に対する結論となる情報を含んだ簡潔な発話を抽出することができず、本研究における重要発話の抽出法に適切ではない。

方法2では、最も高い *tf(t,d)*idf(t)* 値の索引語が複数存在する発話が抽出されると考えられ、方法1と同様にある話題の詳細を記述した発話のように発話内の文章が長い発話を抽出すると考えられる。実験データをもとに方法2を適用してみた。上記の例において、2番目に高い *tf(t,d)*idf(t)* 値の索引語は「借りる」であり、続いて「2」である。方法2による重要発話の抽出を行なうと、「会議」という索引語が多く含んだ発話4と発話6が抽出されることになる。この結果は、方法1と同様である。このことにより、方法2でも方法1と同様に、ある話題の詳細を記述した発話のように発話内の文章が長い発話を抽出する傾向にある。よって、方法2も、議題に対する結論となる情報を含んだ簡潔な発話を抽出することができず、本研究における重要発話の抽出法に適切ではない。

方法3については、各発話中の索引語1つ当りの *tf(t,d)*idf(t)* 値を求めていることになり、発話毎の *tf(t,d)*idf(t)* 値の密度を手がかりに発話を抽出することになる。この方法では、*tf(t,d)*idf(t)* 値の密度の高い発話が抽出され、発話内の文章の長さに依存しない結果が得られると考えられる。実験データをもとに方法3を適用してみた。方法3による重要発話の抽出を行なうと、発話6と発話8が抽出される。発話6は、議題となる情報を含んだ発話として考えることができ、発話8は、発話6の結論となる情報を含んだ発話として考えることができる。このことにより、方法3では、発話の長さに依存しない結果が得られ、討議スレッド中から重要とされる発話を2つ抽出することができると考えられる。よって、本研究における重要発話の抽出法に適切であると考えられる。

以上のことより、本研究における重要発話の抽出法として、方法3による各発話中の索引語1つ当りの *tf(t,d)*idf(t)* 値を手がかりとする手法を用いる。

4.6 討議内容の要約手法

本研究における要約手法の手順を以下に示す。

1. 発話毎に索引語を抽出する。
   本研究における索引語とは、名詞、動詞、形容詞から代名詞、接尾語、高頻度語 (「する」) を除いたものである。
2. 抽出した索引語に対して、*tf(t,d)* の値を求める
   本研究では、4.5.1 節の式 (4.2) において、t はある討議スレッド d 中に出現する索引語 t の出現数、T は討議スレッド d 中の全ての索引語の出現数である (図 4.3 の場合、t の値は 2 となる)。
3. 抽出した索引語に対して、*idf(t)* の値を求める
   本研究では、4.5.1 節の式 (4.3) において、N は検索対象となる討議スレッドの集合中の全討議スレッド数、*df(t)* は索引語 t が出現する討議スレッド数である。つまり、idf は、ある索引語が少数の討議スレッドにしか出現しない場合に大きくなり、どの討議スレッドにも出現すると最小の値になる (図 4.3 の場合、*df(t)* の値は 3 となる)。
4. 発話毎に索引語の *tf(t,d)*idf(t)* 値の合計を求める (4.5.1 節の式 (4.4) 参照)。
5. 各発話中の索引語１つ当りの *tf(t,d)*idf(t)* 値を求める。
6. 討議スレッド毎に各発話の索引語１つ当りの *tf(t,d)*idf(t)* 値を降順にソートし、値の高い２つの発話を抽出する。

5 章では、討議内容要約エンジンのアルゴリズムについて述べ、討議内容要約エンジンを用いた要約手法の評価について述べる。

第5章　討議内容要約エンジンの実現

本研究では、討議内容の自動要約である討議内容要約エンジンを実現した。討議内容要約エンジンは、討議構造木抽出エンジンから得られる UMML+Linkbase ファイルを入力とし、tf*idf 法を利用した要約手法を用いて重要発話を抽出する。

5.1　概要

討議内容要約エンジンのアルゴリズムの概要を以下に示す。

1. UMML ファイルから全討議内容の発話とその発話の情報 (WhoPr、WhomPr など) を抽出する。
2. Linkbase ファイルから発話間の接続関係の情報を抽出する。
3. 1 で抽出した発話に対して、茶筌による形態素解析を行ない、その結果から索引語の抽出を行なう。
4. 発話毎に索引語 1 つ当りの $tf(t,d)*idf(t)$ 値を求める。
5. 討議内容毎に各発話の索引語 1 つ当りの $tf(t,d)*idf(t)$ 値を降順にソートし、値の高い 2 つの発話を抽出する。
6. 5 で抽出した 2 つの発話を XML 文書へ出力する。

概要の 6 において、得られた討議内容の要約の全てを XML 文書へ出力するのではなく、以前に討議内容要約エンジンを起動した時点から、追加されたメールに含まれる発話に関係する討議内容のみの要約を XML 文書へ出力する。なぜなら、tf*idf 法における idf が、全ての討議内容を対象に索引語の重み付けを行なっているため、追加されたメールの索引語に依存して $idf(t)$ 値が変化し、過去に終結した討議内容の要約などが変更する可能性があるからである。本研究では、UMML+Linkbase ファイルに追加されたメールに対する情報がないため、手動で UMML ファイルに追加されたメールの情報を付加した。

討議内容要約エンジンでは、討議構造木抽出エンジンから得られる UMML+Linkbase ファイル中の以下の情報を用いて、討議内容毎に重要な発話を抽出する。

入力
- 発話内容 (UMML ファイル)
- WhoPr, WhomPr, 追加メールなどの発話に関する情報 (UMML ファイル)
- 発話間の接続関係 (Linkbase ファイル)

出力 • 各討議内容の重要発話とその発話情報 (WhoPr, WhomPr など)

抽出手順は、大きく次の 3 つからなる。

手順 1 UMML+Linkbase ファイルの読み込み

手順 2 討議内容毎に重要発話を抽出

手順 3 XML 文書への出力

各手順の詳細について、順に説明する。

5.2 手順 1. UMML+Linkbase ファイルの読み込み

手順 1.1. UMML ファイルの読み込み

UMML ファイルから発話内容や WhoPr, WhomPr, 追加メールなどの情報を読み込む。発話内容は utterance 要素から読み込み、WhoPr は From 要素、WhomPr は utterance 要素の whomPr 属性から読み込む。また、本研究では、utterance 要素の add_flag 属性として、追加されたメールの発話であるかどうかの情報を追加した (1 が追加を表わす)。以下に UMML ファイルの例を挙げる。

```
 1  <?xml version='1.0' encoding='ISO-2022-JP'?>
 2  <!DOCTYPE umml SYSTEM "UMML.dtd">
 3  <umml>
 4  <header>
 5  Return-Path: exp-admin@nuts.sample.sp
 6  X-UIDL: /]D&quot;!_8]&quot;!?&amp;]!!mWj!!
 7  Return-Path: &lt;exp-admin@nuts.sample.sp&gt;
      (中略)
14  Date: <Date>Mon, 20 Nov 2000 14:52:10 +0900</Date>
15  Posted: Mon, 20 Nov 2000 14:51:34 +0900 (JST)
16  From: <From>taro@sample.sp (Taro)</From>
17  Reply-To: exp@nuts.sample.sp
18  Subject: <Subject>[exp:00007] Re: read-finish</Subject>
19  To: <To>exp@nuts.sample.sp (exp ML)</To>
      (中略)
26  X-ML-Info: If you have a question, send a mail with the body &quot;#
    help&quot; (without quotes) to the address exp-ctl@nuts.sample.sp
27  Precedence: bulk
28  Lines: 9
29  Content-Type: text
```

```
30  Status: U
31  </header>
32  <body>
33  <utterance type="0" whomPr="">
34  太郎です。
35  </utterance>
36  <utterance type="1" Pr="C1.2-Pr1.2" whomPr=""
    flags="C1.3_0_1-Pr1.3_0_1=0," add_flag="1">
37  今、読み終りました。
38  足りない内容の把握のため 15:00 までかけてもう一読みします。
39  その間、お二人には、「mail、メイル、メール」といった語の統一と、
40  メールによるサービスが実行される前提条件のチェックを、
41  もう一度お願いします。
42  </utterance>
43  </body>
44  </umml>
```

手順 1.2. Linkbase **ファイルの読み込み**

Linkbase ファイルから発話間の接続関係を読み込む。接続関係は、arc 要素の xlink:from 属性と xlink:to 属性から読み込む。以下に Linkbase ファイルの例を挙げる。

```
 1  <?xml version='1.0'?>
 2  <!DOCTYPE linkbase SYSTEM "Linkbase.dtd">
 3  <linkbase>
      (中略)
 9  <link xlink:type="extended">
10  <locator xlink:type="locator" xlink:label="C1.3_0_3-Pr1.3_0_3"
    xlink:href="4.xml#C1.3_0_3-Pr1.3_0_3"
    xlink:title="4.xml#C1.3_0_3-Pr1.3_0_3"/>
11  <locator xlink:type="locator" xlink:label="C1.3_0_1-Pr1.3_0_1"
    xlink:href="3.xml#C1.3_0_1-Pr1.3_0_1"
    xlink:title="3.xml#C1.3_0_1-Pr1.3_0_1"/>
12  <arc xlink:type="arc" xlink:from="C1.3_0_1-Pr1.3_0_1"
    xlink:to="C1.3_0_3-Pr1.3_0_3" xlink:show="new"
    xlink:actuate="onRequest" xlink:title="Ac_Pr"/>
13  </link>
      (中略)
83  </linkbase>
```

5.3 手順2. 討議内容毎に重要発話を抽出

手順 2.1. 索引語の抽出

まず、UMML ファイルから読み込んだ発話内容に対して、茶筌による形態素解析を行なう。以下に解析結果の一部を示す。

部屋	ヘヤ	部屋	名詞-一般
を	ヲ	を	助詞-格助詞-一般
区切っ	クギッ	区切る	動詞-自立
て	テ	て	助詞-接続助詞
狭く	セマク	狭い	形容詞-自立
する	スル	する	動詞-自立
か	カ	か	助詞-副助詞／並立助詞／終助詞

形態素解析を行なった結果から、以下に示す索引語の対象となるものを抽出し、その索引語の対象となるものから不要語を除いたものを抽出する。これにより、抽出されたものが索引語である。

- **索引語の対象**

  名詞・動詞・形容詞

- **不要語**

  代名詞・接尾語・「する」(高頻度語)

手順 2.2. 各発話に対して索引語１つ当りの *tf(t,d)*idf(t)* 値を求める

手順 2.1. で抽出した索引語に対して、*tf(t,d)*idf(t)* 値 (4.5.1 節の式 (4.4) 参照) を求め、発話毎に *tf(t,d)*idf(t)* 値の合計を求める。その合計値を各発話中の索引語数で割り、発話毎に索引語１つ当りの *tf(t,d)*idf(t)* 値を求める。

手順 2.3. 重要発話の抽出

手順 1.2. から得られる全ての討議について、各発話の索引語１つ当りの *tf(t,d)*idf(t)* 値を降順に並べ、値の高い発話を２つ抽出する。これにより、抽出されたものが重要発話である。

5.4 手順3. XML 文書への出力

本研究では、各討議内容とその要約を用いて、討議の概観を視覚表示するシステムを実現する。そのシステムを実現するために、以下に示す４つのパターンの XML 文書を作成した。

- 全討議内容の参加者リスト (main.xml)
- 各参加者が関係している討議内容の要約リスト (participant*.xml)
- 各討議内容の全ての発話内容 (deliberation*.xml)
- 全討議内容の要約リスト (summary.xml)

各パターンの詳細について、順に説明する。

全討議内容の参加者リスト (main.xml)

この XML 文書は、全ての討議内容の参加者を要素とする。以下に例を示す。

```
 1  <?xml version="1.0" encoding="Shift_JIS" standalone="yes"?>
 2  <?xml-stylesheet type="text/xsl" href="main.xsl" ?>
 3  <main>
 4    <participant>
 5      <name>
 6        Hanako &lt;hanako@sample.sp&gt;
 7      </name>
 8      <link>
 9        participant1.xml
10      </link>
11    </participant>
12    <participant>
      (中略)
35    </participant>
36    <all>
37      <link>
38        <xml>
39          summary.xml
40        </xml>
41        <title>
42          要約一覧
43        </title>
44      </link>
45    </all>
46  </main>
```

この XML 文書は、ルート要素である <main> 要素の中に、複数の <participant> 要素で構成されている。<participant> 要素の内容は、<name> 要素と <link> 要素である。

<name> 要素の内容は参加者名で、<link> 要素の内容はリンク先の XML 文書である。最後の段落は、<all> 要素となっており、<link> 要素の中に、<xml> 要素と <title> 要素で構成されている。<title> 要素の内容は「要約一覧」で、<xml> 要素の内容は「要約一覧」のリンク先の XML 文書である。

各参加者が関係している討議内容の要約リスト (participant*.xml)

この XML 文書は、各参加者が関係している討議内容の要約を要素とする。以下に例を示す。

```
 1  <?xml version="1.0" encoding="Shift_JIS" standalone="yes"?>
 2  <?xml-stylesheet type="text/xsl" href="participant.xsl" ?>
 3  <summaries>
 4    <pageTitle>
 5      Ichiro &lt;ichiro@sample.sp&gt;
 6    </pageTitle>
 7    <thread>
 8      <summary contribution="C1.1-Pr1.1" type="R" whoPr="Ichiro
         &lt;ichiro@sample.sp&gt;" whomPr="ALL">
 9        読み終わりましたが、
10        皆さんいかがでしょうか？
11      </summary>
12      <summary contribution="C1.1-Pr1.1-C1.1.2-Pr1.1.2" type="F"
         whoPr="Taro &lt;taro@sample.sp&gt;" whomPr="ALL">
13        僕も読み終わりました。
14        開始時刻はいつでも OK です。
15      </summary>
16      <link>
17        <xml>
18          deliberation1.xml
19        </xml>
20        <title>
21          D-1
22        </title>
23      </link>
24    </thread>
25    <thread>
      (中略)
91    </thread>
```

```
92  </summaries>
```

この XML 文書は、ルート要素である <summaries> 要素の中に、<pageTitle> 要素と複数の <thread> 要素で構成されている。<pageTitle> 要素の内容は参加者名であり、<thread> 要素の内容は、複数の <summary> 要素と <link> 要素である。<summary> 要素の内容は、<pageTitle> 要素の参加者が関係している討議の要約であり、属性として、”contribution”, ”type”, ”whoPr”, ”whomPr”をもつ。<link> 要素の内容は、<xml> 要素と <title> 要素である。<title> 要素の内容は「D-*」で、<xml> 要素の内容はリンク先の XML 文書である。

各討議内容の全ての発話内容 (deliberation*.xml)

この XML 文書は、各討議内容の全ての発話内容を要素とする。以下に例を示す。

```
 1  <?xml version="1.0" encoding="Shift_JIS" standalone="yes"?>
 2  <?xml-stylesheet type="text/xsl" href="deliberation.xsl" ?>
 3  <deliberation>
 4    <pageTitle>
 5      Deliberation 1
 6    </pageTitle>
 7    <utterance contribution="C1.1-Pr1.1" type="R" whoPr="Ichiro
       &lt;ichiro@sample.sp&gt;" whomPr="ALL">
 8      読み終わりましたが、
 9      皆さんいかがでしょうか？
10    </utterance>
      (中略)
30    </utterance>
31  </deliberation>
```

この XML 文書は、ルート要素である <deliberation> 要素の中に、<pageTitle> 要素と複数の <utterance> 要素で構成されている。<pageTitle> 要素の内容は討議名 (Deliberation *) である。<utterance> 要素の内容は、討議内容中の発話であり、属性として、”contribution”, ”type”, ”whoPr”, ”whomPr”をもつ。

全討議内容の要約リスト (summary.xml)

この XML 文書は、全ての討議内容の要約を要素とし、各参加者が関係している討議内容の要約を要素とする XML 文書 (participant*.xml) と構造がほぼ同じである。違いは、<pageTitle> 要素の有無である。以下に例を示す。

```
 1  <?xml version="1.0" encoding="Shift_JIS" standalone="yes"?>
 2  <?xml-stylesheet type="text/xsl" href="summary.xsl" ?>
 3  <summaries>
 4    <thread>
      (中略)
48    </thread>
49    <thread>
50      <summary contribution="C2.3-Pr2.3" type="R" whoPr="Ichiro
         &lt;ichiro@sample.sp&gt;" whomPr="hanako@sample.sp">
51        電子メールとメールと分ける必要はあるのですか？
52      </summary>
53      <summary contribution="C2.6-Pr2.6" type="F" whoPr="Hanako
         &lt;hanako@sample.sp&gt;" whomPr="ALL">
54        なかった。「メール」で統一。
55      </summary>
56      <link>
57        <xml>
58          deliberation3.xml
59        </xml>
60        <title>
61          D-3
62        </title>
63      </link>
64    </thread>
65    <thread>
      (中略)
91    </thread>
92  </summaries>
```

5.5 要約手法の評価

討議内容要約エンジンを用いて、要約手法の評価を行なった。要約手法の評価として、以下の尺度を用いて評価を行なった。

- 正解率：抽出された 2 つの発話が、本研究における要約手法の方針にあった発話を抽出できているか

5.5.1 正解率による評価結果

評価を行なうためのデータとして、3 人の被験者によるメーリングリストを利用した輪講室予約システムの仕様書レビューでやりとりされた討議内容を用いる。話題が終結した討議には、結論となる情報を含んだ発話が含まれており、その発話は討議中の重要発話と考えることができるため、その発話を評価の対象として用いる。評価は、話題が終結した討議と話題が未終結な討議に分けて行なった。

話題が終結した討議についての評価

評価値として、以下に示す正解率を用いた。

$$\text{正解率}\,(\%) = \frac{\text{議題と結論となる情報を含んだ発話を抽出した討議スレッド数}}{\text{討議スレッド数}} \times 100 \tag{5.1}$$

議題と結論となる情報を含んだ発話を抽出した討議スレッド数とは、討議内容の要約として抽出された 2 つの重要発話が、議題となる情報を含んだ発話と結論となる情報を含んだ発話である討議スレッドの数である。結論となる情報を含んだ発話とは、了承の意図を表わす発話であることが多い。以下に例を示す。

例 1:
　僕も読み終わりました。
　開始時刻はいつでも OK です。

例 2:
　そうですね。
　オープンウィンドウで行なっている方の事を考え、10 件としましょう。

但し、「了解。」、「OK です。」などの了承の意図を表わす言葉のみの発話に関しては、その了承の意図を表わす返答が行なわれているメールの引用文となる発話が結論となる情報を含んだ発話であることが多い。また、議題となる情報を含んだ発話とは、結論となる情報の本質と掛け離れていない内容が示された発話である。以下に議題とそれに対する結論となる情報を含んだ発話を抽出した例と抽出できていない例を示す。

例 A. 議題とそれに対する結論となる情報を含んだ発話の例

例 A-1:
　発話 1
　　先にキャンセル処理を行うのではなく、予約申請を行うべき。
　　でないと、キャンセルだけされて、他の部屋の予約がとれない状況に。
　発話 2
　　確かに、先に予約申請を行って場所がとれるか確保してから、
　　キャンセル処理を行うほうがいいですね。

例 A-2:
　発話 1
　　その節割りで、その人を中心に初めの節から順に全員で進めていくってのは
　　どうですか？
　発話 2
　　了解。

例 B. 議題とそれに対する結論となる情報を含んでいない発話の例

例 B-1:
　発話 1
　　日常的に使い慣れている語は、極力、カタカナで統一しましょう。
　　例：メール、アドレス、フォーマット
　　「unix」と「E-mail」はそれぞれ、
　　「システム」と「電子メール」と書くことを望みます。
　発話 2
　　電子メールとメールと分ける必要はあるのですか？

例 B-2:
　発話 1
　　上記の内容に賛成です。
　発話 2
　　遅くなりましたが、賛成します。

例 A-1 では、発話 1 が議題となる情報を含んだ発話であり、発話 2 が結論となる情報を含んだ発話である。例 A-2 では、発話 2 が了承の意図を表わす言葉のみの発話であるが、発話 1 の結論となる情報を含んだ発話である。発話 1 は、議題となる情報を含んだ発話である。例 B-1 では、発話 1、発話 2 ともに返答を要求する発話であり、結論となる情報を含んだ発話が抽出されていない。例 B-2 では、発話 1、発話 2 ともに結論となる情報を含

んだ発話である。これは、複数人での討議であるために起こりうることである。

3 人の被験者によるメーリングリストを利用した輪講室予約システムの仕様書レビューでやりとりされた討議内容を用いて、話題が終結した討議内容についての評価を行なった結果、討議スレッド数が**24** で、議題と結論となる情報を含んだ発話を抽出した討議スレッド数が**17** となり、正解率は**70.8**% となった。討議スレッド数 24 のうち、結論となる情報を含んだ発話が抽出されていた討議スレッド数は 22 であったため、抽出間違いとなった主な原因は、議題となる情報を含んだ発話が抽出できなかったことにあった。

話題が未終結な討議についての評価

評価値として、以下に示す正解率を用いた。

$$\text{正解率}\ (\%) = \frac{\text{議題と最終提案となる情報を含んだ発話を抽出した討議スレッド数}}{\text{討議スレッド数}} \times 100 \tag{5.2}$$

議題と最終提案となる情報を含んだ発話を抽出した討議スレッド数とは、討議内容の要約として抽出された 2 つの重要発話が、議題となる情報を含んだ発話と最終提案となる情報を含んだ発話である討議スレッドの数である。以下に議題と最終提案となる情報を含んだ発話を抽出した例と抽出できていない例を示す。

例 A. 議題と最終提案となる情報を含んだ発話の例

例 A-1:
発話 1
P3 の上のほうの希望は一つもかかなくてもいいですし、人数も
省略可能ですという文章は、P2 の予約申請のアイテマイズの
3 番目の本文に人数もしくは、第 n 希望の項目ではじまる行が
あるというルールと矛盾していると思います。
発話 2
この文章自体なくてもいいのでは？

例 B. 議題と最終提案となる情報を含んでいない発話の例

例 B-1:
発話 1
年月日は私の所で出て来るけど、曜日は出てこないぞ。
日時は 3 で説明することにして、曜日はそちらに任せる。
発話 2
よく読むと、曜日を「指定する」場面は無い気がする。

例 A-1 では、発話 1 が議題となる情報を含んだ発話であり、発話 2 が最終提案となる情報を含んだ発話である。例 B-1 では、発話 1、発話 2 の両方が討議内容の序盤に現れており、最終提案となる情報を含んでいない。3 人の被験者によるメーリングリストを利用した輪講室予約システムの仕様書レビューでやりとりされた討議内容を用いて、話題が未終結な討議内容についての評価を行なった結果、討議スレッド数が**11** で、議題と最終提案となる情報を含んだ発話を抽出した討議スレッド数が**8** となり、正解率は**72.7**%となった。抽出間違いとなった原因は、討議内容の終盤に現れる最終提案となる情報を含んだ発話を抽出することができていないことである。

以上の結果から、討議内容の要約手法の正解率は 70%ぐらいであるといえる。さらに精度を上げるとするなら、各索引語の *tf(t,d)*idf(t)* 値も重要発話の抽出手がかりとして追加することだと考えられる。6 章では、討議内容抽出エンジンから得られる討議内容の要約を用いて、討議構造の呈示法について述べる。

5.5.2 抽出した発話に関する考察

本研究における要約手法の方針の１つに、討議内容の要約として、各討議スレッドから２つの発話を重要発話とした。この方針に対して、重要発話を討議スレッド中の２つとして適切であったかを考察する。

5.5.1 節に挙げた話題が終結した討議内容と話題が未終結な討議内容の正解例 A-1,A-2 において、抽出された２つの発話から討議の概要を理解できると判断した。また、不正解となった討議スレッドに対して、３つの目の発話を抽出してみたところ、不正解となった 10 個の討議スレッドのうち３つの討議スレッドしか正解となる発話を抽出できなかったため、２つの発話を重要発話とすることは適切であると判断した。４つ目以上の発話を抽出することに関しての評価は、要約という観点から離れると考え、それ以上は考察する必要がないと判断した。

第6章　討議構造の呈示システムの実現

6.1　システムの概要

討議内容要約エンジンが生成した４つのパターンの XML 文書を用いて、討議内容の要約とその討議構造をブラウザに呈示するシステムを実現した。このシステムは、討議の参加者を中心とし、指定した討議の参加者が関係している討議内容の要約とその討議内容を視覚表示する。また、全ての討議内容の要約とその討議内容も視覚表示する。

6.2　XML を用いたブラウザ呈示

４つのパターンの XML 文書をブラウザに呈示するため、それぞれに対して XSL(Extensible Stylesheet Language) ファイルと CSS(Cascading Style Sheets) ファイルを作成した。この作成した XSL ファイルと CSS ファイルより XML 文書を HTML 化し、ブラウザに呈示する。

6.3　討議構造の呈示法

メイン画面として、参加者リストを呈示する。また、最下部には、要約一覧を呈示する欄を追加した (図 6.1)。参加者名と要約一覧にはリンクが張られており、ある参加者を指定するとその参加者が関係している討議の要約一覧を呈示する (図 6.2)。また、要約一覧を指定すると全ての討議の要約一覧を呈示する (図 6.3)。要約一覧は、討議の番号を示す D-No. に対して２つの重要発話とそれぞれの発話の WhoPr、WhomPr を呈示する。各討議の終結・未終結の情報は、D-No. の背景色で区別できるようにした。また、D-No. にはリンクが張られており、ある D-No. を指定するとその討議の全ての発話を呈示する (図 6.4)。呈示する各発話には、WhoPr と WhomPr の情報を追加した。発話の種類 (返答を要求する発話か話題を終結する発話) は、発話の背景色で区別できるようにした。

図 6.1: 参加者リスト

図 6.2: ある参加者が参加した討議の要約一覧の例

図 6.3: 全討議の要約一覧

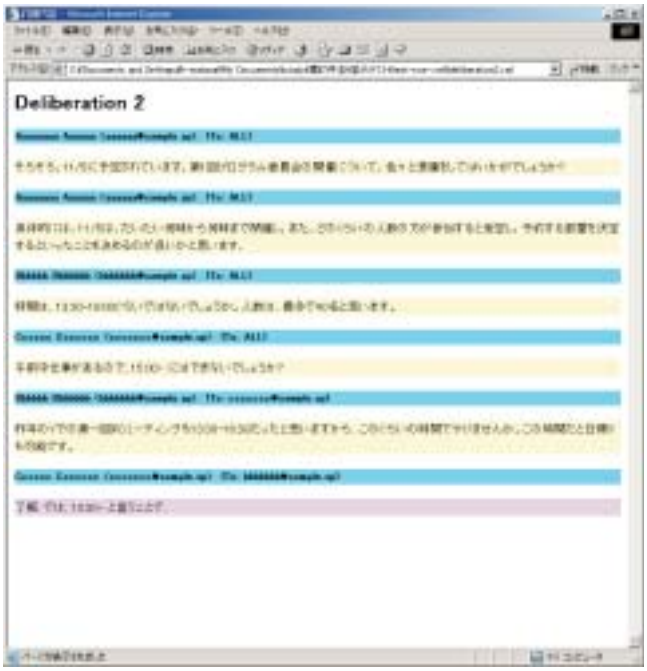

図 6.4: ある討議の全発話

第7章　おわりに

7.1　まとめ

本研究では、自然言語処理の分野で研究が行なわれているテキスト自動要約技術を利用し、討議内容の自動要約を実現した。また、討議内容の自動要約から生成される要約を用いた討議構造の呈示システムの開発を行なった。

討議内容の要約手法として、手がかり表現を利用した要約手法を考案したが、要約の対象が討議内容の終盤だけに限定しているため、それ以外の箇所にある必要な情報を抽出することができないという結果となった。そこで、要約の対象を討議全体とし、tf*idf 法を用いて単語の重み付け技法による討議内容の要約手法を考案した。その要約手法は、討議スレッド毎に各発話中の索引語１つ当りの $tf(t,d)*idf(t)$ 値を求め、その値が高い２つの発話を抽出することである。この手法では、70％の正解率が得られ、討議内容の要約に適切であると判断した。また、２つの発話を重要発話とすることも適切であると判断した。この討議内容の要約手法をもとに、討議内容の自動要約である討議内容要約エンジンを実現した。討議内容要約エンジンは、討議構造木抽出エンジンから得られる討議スレッドを入力とし、討議内容の要約と討議内容を XML 文書へ出力する。その出力結果を用いて、討議の参加者を中心とし、指定した討議の参加者が関係している討議内容の要約と討議内容を視覚表示するシステムを実現した。

7.2　今後の課題

今後の課題を以下に示す。

- 要約手法の再考

  本研究では、要約手法の１つの行程として、索引語の選択を行なった。索引語の選択において、「名詞」、「動詞」、「形容詞」の３つを主として用いたため、十分な索引語の絞り込みができていないと考えられる。よって、要約手法の改良案として、この３つの品詞の中から更に絞り込み、索引語の選択を強める必要があると考える。

- 運用実験による評価

  本研究の目的は、討議の概観を視覚表示するシステムの実現であり、討議の参加がそのシステムによって討議の概要を理解でき、円滑なコミュニケーション支援の実

現ができることである。そこで、実際にこのシステムを利用することで、電子メールを用いるコミュニケーションにおける新たな知見を得る必要があると考える。

謝辞

本研究を行なうにあたり、終始変わらぬ御指導を賜わりました落水浩一郎教授に心から深く感謝申し上げます。

本研究を進めるにあたり、随所で貴重なご意見を賜わりました村越広享助手に深く感謝致します。

研究の節目節目において、適切な助言を下さいました島津明教授に深く感謝致します。

論文審査にあたり、片山卓也教授、篠田陽一教授には御助言、御意見をいただき深く感謝致します。

藤枝和宏助手、服部哲助手には多大な御助言をいただき深く感謝致します。

本研究室卒業生の山見太郎氏には、本研究を始めるにあたり、多大な御助言をいただき深く感謝致します。

最後に、落水研究室の皆様の日頃の討論と助言、励ましに深く感謝致します。

参考文献

[1] 村越広享, 山見太郎, 島津明, 落水浩一郎. 電子メールを利用した学習者間のコミュニケーション支援技術の開発. 教育システム情報学会誌, Vol.18, No3・4, pp.308-318, 2001.

[2] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向. 自然言語処理, Vol.6, No.6, pp1-26, 1999.

[3] Herbert H. Clark and Edward F. Schaefer. Contributing to Discourse. *Cognitive Science*, Vol.13, No.2, pp.259-294, 1989.

[4] Hiroyuki Murakoshi, Akira Shimazu, and Koichiro Ochimizu. Construction of deliberation structure in e-mail communication. *International Journal of Computational Intelligence*, Vol.16,No.4, pp.570-577, 2000.

[5] 村越広享, 島津明, 落水浩一郎. メーリングリストを利用した共同作業における討議構造の自動構築法. コンピュータソフトウェア, Vol.18, No.3, pp.19-23, 2001.

[6] D. G. Novick, L. Walton, and K. Ward. Contri-bution graphs in multiparty discourse. In *Pro-ceedings of International Symposium on Spoken Dialogue*, pp.53-56, 1996.

[7] XML Linking Language (Xlink) Version 1.0, Jul. 2000. http://www.w3.org/TR/2000/CR-xlink-20000703/.

[8] Rachel Reichman. Conversational Coherency. *Cognitive Science*, Vol.2, No.4, pp.283-327, 1978.

[9] 泉子・K・メイナード. 会話分析. くろしお出版, 1993.

[10] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, 1976.

[11] Tim Enders, Jeff Gay, and Y. Miyadate. ICE-Mail. http://www.icemail.org/.

[12] Paice, C. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing and Management*, 26 (1), pp.171-186, 1990.

[13] Skorokhod'ko, E. Adaptive method of automatic abstracting and indexing. In *Information Processing 71*, pp.1179-1182. North Holland, 1972.

[14] Halliday, H. A. K. and Hassan, R. *Cohesion in English*. Longman, 1976.

[15] Hoey, M. *Patterns of lexis in text*. Oxford University Press, 1991.

[16] 亀田雅之. 疑似キーワード相関法による重要キーワードと重要文の抽出. 言語処理学会第 2 回年次大会発表論文集, pp.97-100, 1996.

[17] Salton, G. *Automatic Text Processing*. Addison-Wesley, 1989.

本研究に関する発表

[**1**] 村越広享, 渡邊大貴, 島津明, 落水浩一郎. 討議構造参照機能を有するメールクライアント. 教育システム情報学会第 26 回全国大会, デモセッション, August 2001.

[**2**] 渡邊 大貴, 村越 広享, 島津 明, 落水 浩一郎: 電子メールコミュニケーションにおける討議スレッドの要約手法, 平成 13 年度 電気関係学会北陸支部連合大会, pp.321, October 2001.