自 然 言 語 処 理 154－3
（２００３．３．６）

メール型ヘルプデスク応対記録検索システムにおける ドメインモデルの利用

伊藤 元之*

現在我々は、社内情報システムに関する社内からの問い合わせに答える際に、社内ヘルプデスクオペレータが使用するための、過去の応対事例記録検索システムを作成している。これまで、全文検索システムをベースにシステムを開発してきたが、全文検索では、問い合わせの意味内容を捉えた検索が十分に実現できないという問題点がある。従来型の検索システムでも、(1)語に重要度を導入し、適合度評価の際に重み付けをする、(2)句や節といった構文構造単位の合致度に着目し、そのレベルでの合致性を適合度評価に加味する、といった補完的措置が試みられているが、根本の検索原理が、統計処理であるために、その種の、意味を考慮した補完措置との整合性のコントロールが難しく、安定に精度を上げていくことが難しい。本研究では、ドメインモデルの導入により、意味情報をより安定に利用できる検索システムを実現する方法について検討する。

Utilizing Domain Model for Semantic-based Document Search

Motoyuki Itoh

In the field of document search, many full-text search systems have developed by applying various statistical analysis methods. But to establish a truly flexible search system, we must introduce some semantic-based analysis methods to the system. We have now constructed the document search system which prepare for a domain model on which the system interprets input queries.

* （株）CSK 情報システム本部
Information System Division, CSK Corporation.

1.はじめに

現在我々は、社内情報システムに関する社内からの問い合わせに答える際に、社内ヘルプデスクオペレータが使用するための、過去の応対事例記録検索システムを作成している。これまで、全文検索システムをベースにシステムを開発してきたが、全文検索では、問い合わせの意味内容を捉えた検索が十分に実現できないという問題点があるため、本研究では、ドメインモデルの導入により、その事態を改善する手法について検討する。

2.基礎的考察

2.1.検索システムのタイプ

現在実用されている検索システムは、次の２種類に大別できる。

A.strict マッチ型（キーワード）検索システム

検索条件として与えられたキーワード及びその論理結合式で示されるキーワードの出現条件が、検索対象文書上で満たされているかどうかを、(表層文字列上で）厳密にチェックするタイプのシステム

B.あいまい検索・自然文検索システム

検索条件は、キーワード、その論理結合式、あるいは、自然文の形態で与えられる。入力された条件があいまいなものである可能性を積極的に認め、指定されたキーワードが一部出現しない検索解や、厳密な意味で検索条件を満たしているかどうか確認が取れない解であっても、一定の基準を満たせば、解として取り上げるタイプのシステム。

従来の検索タスクは、データベース検索のように、事前に十分に解析・整理・断片化された情報を検索対象とすることが多かったため、主に上記Ａ型のシステムで処理されることが多かった。しかしながら、近年は、加工されていない生の文書テキストを、あいまいな検索条件を手がかりとして検索する、というタイプの検索ニーズも高まっており、Ｂ型のアプローチのシステムが次々に登場してきている。

本論文で適用対象と考えているヘルプデスク関連の検索タスクでは、Ｂ型のシステムの適用が望まれるケースが多い。よって、以下では、Ｂ型のシステムの構築法を焦点にして、議論を進める。

2.2.あいまい検索・自然文検索に対する既存のアプローチ

あいまい検索・自然文検索の分野に限っても、様々な試みがなされているが、それらをタイプ分けする一つの切り口として、次のような分類を考えることができる。

(1) 単語を比較の基本単位とする手法
　統計的手法等により、入力条件と検索対象文書との単語出現パターン（どの単語がどれだけ使われているか）の類似度を算出することで、比較を行なう手法。
(2) 句を比較単位とする手法
　入力側と検索対象側とで、どれだけ同じ句（構造は比較的単純なもの）が出現したかを比較基準の一つとする手法。
(3) より大きな構文構造を比較単位とする手法
　節・文等、より大きな構造単位で比較を行い、その合致度を、比較基準の一つとする手法。

現在実用されているあいまい・自然文検索システムの多くは、基本的には(1)の手法を柱にしている。

(2)(3)は、自然文検索を主眼にすえたシステムにおいて、(1)と併用する形で用いられているケースが出てきている。

(1)の手法は、従来の strict マッチ型の検索システムにはない柔軟性を、検索システムにもたらした点で、非常に大きな意義があった。実用性の観点でも、一定の成果を出していると言える。しかしながら、(1)の手法だけでは十分な検索精度が達成できるわけではないことも明らかになりつつある。あいまい検索・自然文検索においては、ユーザは、「内容的」に入力と合致する検索結果を要求する度合いが高い。(1)の手法は「出現単語の分布が似ている→内容が近い」ということを大前提にした手法であるが、実際には、そのような前提が厳密に成立するわけではない（例えば、同じ単語群からでも、単語の並べ方によって、様々な異なる文脈を構築し得る)。そのギャップが検索精度が頭打ちになる一つの大きな要因になっていると見受けられる。

それに対し、(2)や(3)の手法は、単語のレベルでは捉えきれない「内容」的情報を、句や節・文といった単位で捕まえようとする、一種の内容分析手法であり、前述の(1)の手法の不足点を補完する目的で研究が進められてきている。(1)の手法だけの場合よりも、(2)(3)の手法を併用した場合の方が、検索精

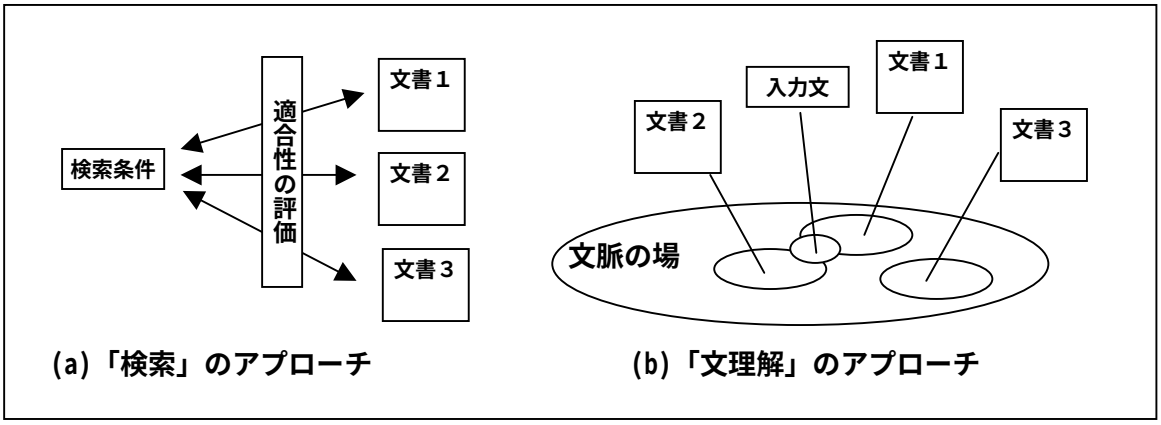

図１「検索」と「文理解」 アプローチの違い

度が改善する、という実験報告も実際に出てきている。

また、(1)(2)(3)、いずれのレベルの手法でも問題となるポイントとして、「重要度」の問題がある。単語にせよ、句にせよ、節・文にせよ、注視すべきものと、それほど注視すべきではないものがある。この問題に関しては、各単語、あるいは、句等の、処理上の着目単位要素毎に、「重要度」を付与し、検索時に、重要度の低い要素（単語、句等）は、比較に利用しない（あるいは、類似度への影響を小さくする）という形で、対処する方法が一般的に取られている。「重要度」の概念を導入した場合の方が、導入しない場合よりも、検索精度が良くなることも、実証的に認められている。

2.3.ドメインモデル

前述したような、句等のレベルの構文構造に着目する方向、重要なものとそうでないものの強弱を付ける方向、それぞれについては、我々も必要性を感じている。しかしながら、(1)のような統計的手法を主体にしておき、そこに足りない要素として、前述のような２つの「意味的」要素（構文構造、重要度）を付加して行く、という進め方には、やや違和感を覚える。

入力された文に対し、それに合致する情報を取り出す、というタスクは、確かに検索タスクであるが、それは、自然文理解の問題でもある。すなわち、自然文理解では、入力された情報が、自分の持つ知識・文脈情報中の、何に、どう対応するのかを解析することが、第 1 の課題となるが、検索のタスクは、まさに、その部分に相当するタスクであると見ることができる。

ところが、現在検索のタスクで取られている手法は、自然言語理解において取られている手法と、根本的な部分で大きな違いがある。検索タスクでは、図１(a)のように、入力と個々の検索対象との間で、個別に直接比較が試みられ、「対応するか否か」（適合性）の判断が行われる。自然言語理解の分野では、図１(b)のように、入力は入力として意味解析され、文脈情報の一部として文脈の場の上に配置され、一方、検索対象（文書等）は検索対象で別途意味解析され、同様に、文脈情報の一部として、文脈上に配置される。両者が、文脈という意味内容を記述した場の上に投影されることにより、はじめて比較が成立するようになる、というのが、自然言語理解の分野での考え方である。

確かに、純粋に(1)のような統計的手法に立脚する限りにおいては、図１(a)のような捉え方が妥当であるが、(2)(3)の手法のように、意味内容を意識した（句・節等の）構文構造に着目したり、あるいは、「重要度」のような、対象世界が定まっていないと議論できない属性を考え始めたりする段階に至っている状況では、むしろ、図１(b)のような考え方を導入するのが妥当ではないだろうか。すなわち、意味内容を問題にし始めるのであれば、図１(b)のように、入力に関しても、検索対象（文書）に関しても、それが、対象世界上で、意味的にどういうふうに位置付けられるのかを意識した上で、比較が行われるようになっていくべきではないかと考える。

そこで本研究では、自然言語理解における文脈の場に相当するものとして、検索対象となる世界を記述したドメインモデルを導入し、そのドメインモデルを用いて検索結果を求めるような検索方式を検討する。

2.4.本研究で提案するドメインモデルとその

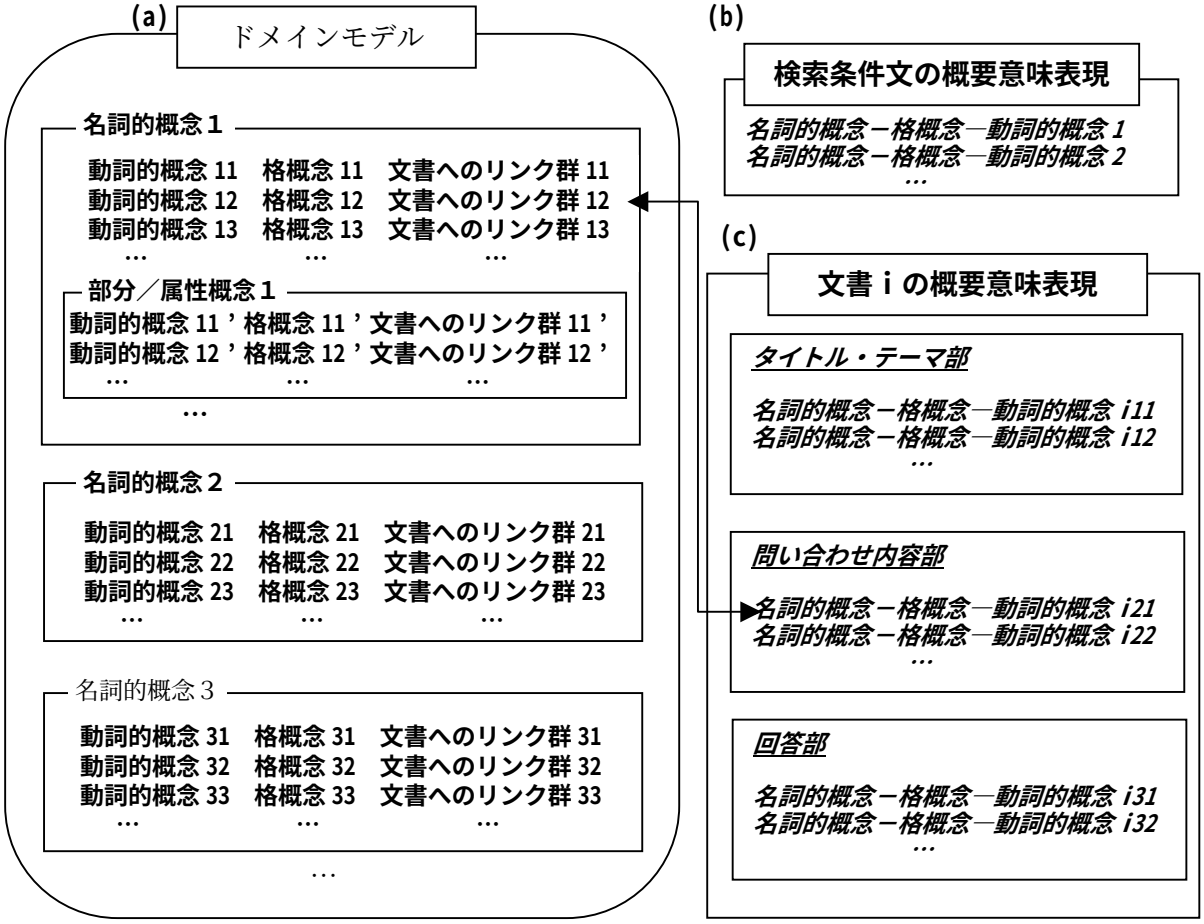

図２　ドメインモデルと入力及び文書の概要意味表現

上への入力・検索対象文書の位置付け

本研究では、文献[2]の成果を受けて、図２(a)のようなドメインモデルを考案した。

ドメインモデルは、基本的に、名詞的概念（「もの」）の集合として記述する。名詞的概念の中には、その名詞的概念が関わる各種の現象（の断面）が位置付けられるようにする。また、名詞的概念の属性や部分に関しても、その属性や部分の主体である名詞的概念の中に位置付けられるようにしておく。

入力・検索対象文書に対しては、形態素解析・構文解析を行い、その中に含まれる「名詞的概念－格概念－動詞的概念」「名詞的概念－連体助詞的概念－名詞的概念」を抽出し（図２(b),(c)　概要意味表現）、そのそれぞれが、ドメインモデル上のどの部分に対応するかをマッチングし、対応づいた要素との間でリンクを張るようにする。それにより、入力・検索対象文書が、ドメインモデル上に位置付けられることになる。なお、本研究で対象とするヘルプデスクの応対記録文書には、タイトル、問い合わせ内容、回答といった、文書上の構造があるため、文書の概要意味表現には、それに相当する構造を持たせる。

図３に具体的な例を示す。この図では、「パスワードがわからなくなった」という入力文、及び、パスワード再発行に関して、過去に行われた応対記録の文書（図中「文書１」）とが、ドメインモデル上に位置付けられている状態を示している。入力文の意味解析（概略意味表現生成）・位置付けは、入力文入力時に行うが、応対記録の文書に関しては、事前に意味解析し、ドメインモデル上に全て位置付けておく。

さて、以上が、本研究でのドメインモデルの表現であるが、フォーマットの問題とは別に、何をどこまで書いておくか、すなわち、具体的データ記述をどこまで詳細に用意しておくかという問題がある。

モデルを扱うシステムが、十分に、情報の適正な価値判断をする能力があるのであれば、いかなる情報をもこのモデル上に書き込んでおいて問題はないが、実際には、そこまで望めない。

そこで本研究では、本研究で対象とするドメインにおいて、着目するに値すると人間（モデル作成者）が明瞭に判断できた概念だけを、モデル上に記述す

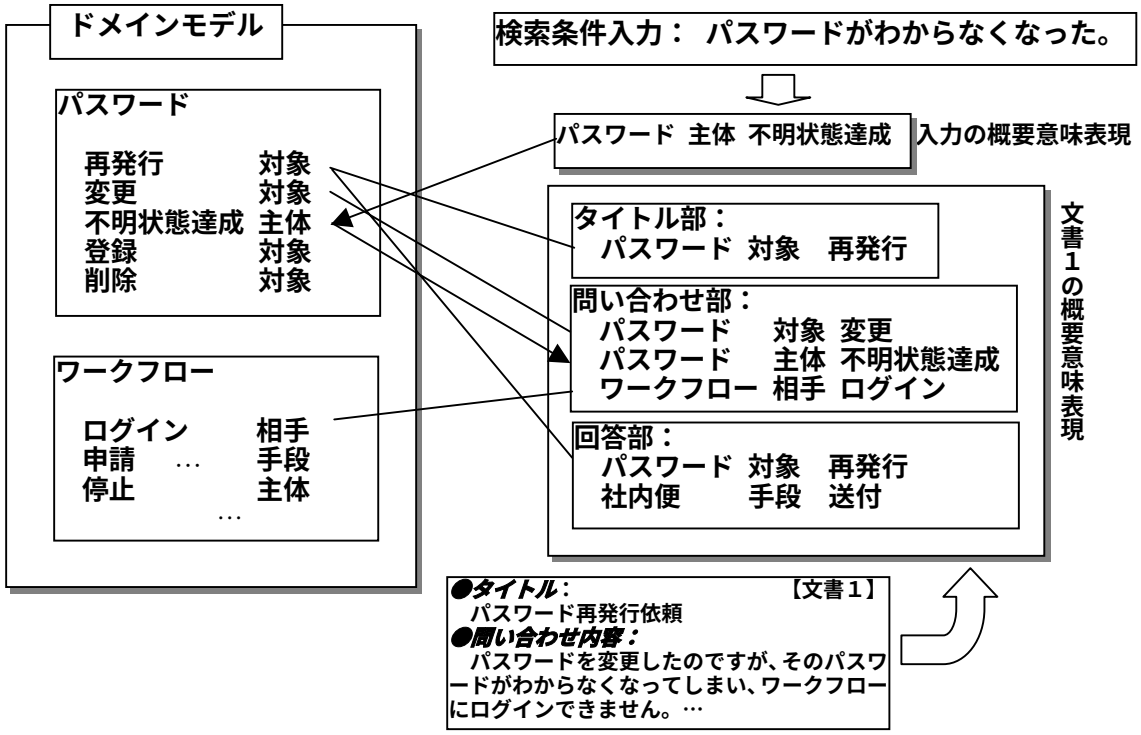

図３ ドメインモデル上への位置付け例

ることにする。

そうすることにより、逆に、このモデル上に投影された概念は、このドメインにおいて着目すべき概念であることがわかることになる。すなわち、結果的に、このドメインモデルが、事物の重要性の判断の一定の尺度を与えてくれるようなものになる。

2.5.検索結果の同定

以上のように、入力文や検索対象とが、同一のモデル上に、意味内容によって位置付けされていれば、そこから、検索結果に相当する文書を同定する方針を立てることは、それほど困難ではない。

まず、ドメインモデル上で、入力文がリンク付けられた場所と、同じ場所にリンク付けられた文書群が、検索結果の候補となる。特に、本研究で検索対象としている文書のように、対象文書の構造上に、重要性の違うセクションが存在する場合には、もっとも重要なセクション（タイトル部等）を介してモデルにリンク付けられている文書群を最有力候補にするのがよいと思われる。また、入力文が位置付けられているリンク先の名詞概念モデルに対して、たくさんのリンクを持つ文書を有力解とする、という考え方もある。さらに、モデルへの位置付けの状況が同一である文書群は、ほぼ同一の内容を持つものと考えることができる。その数があまりに多い場合には、適宜検索解から省略するといったことも、実用的見地からは重要である。

最終的に、どのように検索解を順序付け、絞込むかに関しては、いろいろな方策が考えられると思われる。どれを選ぶのがよいかは、現時点では一概に判断できないが、いずれにせよ、

- モデル上で、入力文と同じ場所に位置付けられた検索対象文書を解候補とできる。
- その位置を中心として、（モデル上でその）周辺に位置付けられている情報（文書）を手繰ることで、関連解の候補が順次得られる。
- リンクを多く持つということは、このドメインにおいて、着目すべき特徴を多く持つ情報であると判断してよい。

と判断してよいと思われる。

このように、対象世界のモデルを価値基準あるいは意味内容の座標系として、一定の手順で、検索解を内容に基づいて取捨選択していくことができることが、従来の統計的手法での１対１比較ベースの検索手法では、実現し得ないメリットである。

3.実験システムの作成

3.1.システムの概要

本研究では、図４のような実験システムを作成した。入力文を入力すると、その内容に相当する過去の応

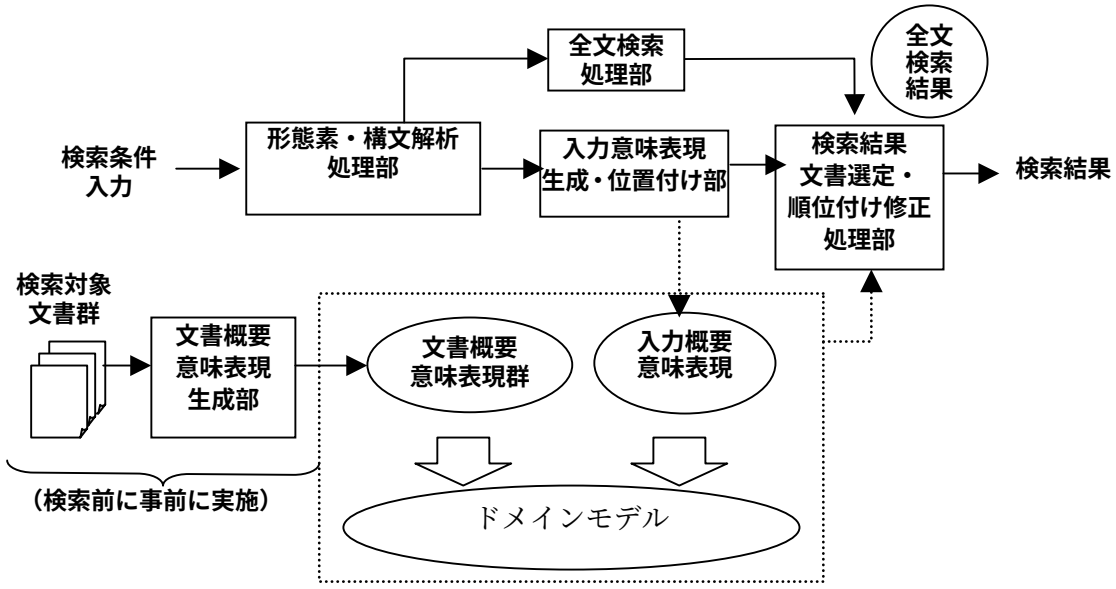

図４ 試作システム概要

対記録文書群が検索されるシステムである。検索対象の応対記録としては、8000 程の文書が用意されている。

今回のシステムでは、別途用意した、全文検索システムの検索結果と、今回のドメインモデルを用いて選択した検索結果とを総合したものを検索結果として提示するようになっている。将来的には、全文検索システム部分の必要性は薄れると思われるが、現段階の実用システムでは、十分な検索語句（文）を入力しないユーザがいるなどするため、安全のために、残してある。

システムは、基本的には Windows NT/2000 上で、ASP(VBScript)で記述している。なお、形態素解析システムについては、茶筅を用い、構文解析システムについては、簡便なものを自作している。

3.2.ドメインモデルの作成

今回ドメインモデルは、上記応対記録文書のタイトル文の解析結果を元に作成した。すなわち、タイトル文の解析結果の中から、重要な名詞概念や、重要な句表現を拾い、それをベースにドメインモデルを記述した。タイトル文は、対象ドメインに関する重要な語句・概念を含んでおり、モデルの骨格となる概念を抽出する元としては、非常に優れている。

3.3.実験・評価

実験・評価については、一般検索者を対象にした本格的なものは、これから行う段階である。

５０文程度の試験的検索入力文で実験をしてみた結果では、検索結果の上位３位までに正解が来る割合は 80%を超えており、この結果は、併用している全文検索システム単独で実験した場合より、10%程よい結果となっている。

文献

[1] 伊藤元之，久保寺正晃，"質問文の句構成に着目したＱＡ事例集検索手法について",信学技法,TL2002-6,pp.31-36，May．2002.

[2] 松原隆男，伊藤元之，高木朗,“ナビゲーション対話システムにおける意味解析手法の検討”，情処研究報告,NL-108，1995

[3] “質問応答（QA）技術最前線 － QA の現状と今後の可能性” 講習会資料,電子情報通信学会 NLC 研究会，Jan．2003

[4] 清田陽司，黒橋禎夫，木戸冬子，"大規模テキスト知識ベースに基づく自動質問応答 －ダイアログナビ－"，言語処理学会第８回年次大会発表論文集，pp.271-274，Mar．2002.

[5] Microsoft 話し言葉によるサポート技術情報検索 ,http://www.microsoft.com/japan/enable/nlsearch/

[6] NEC121ware レスキュー，http://121ware.com/rescue/

[7] SONY VAIO カスタマーリンク，http://vcl.vaio.sony.co.jp/index.html

[8] 長野徹，武田浩一，那須川哲哉，"テキストマイニングのための情報抽出"，情処研究会報告,FI-60-5,pp.31-38,Sep.2000.

[9] 那須川哲哉，"コールセンターにおけるテキストマイニング"，人工知能学会誌，Vol.16,No.2,pp.219-225，2001

[10]松村真宏，大澤幸生，谷内田正彦，"AAS:過去の回答の自動組み合わせに基づく質問応答システム"，信学技報,AI98-64,pp.17-24,Jan．1999

[11] Salton，G．and McGill，M.J．”The SMART and SIRE experimental retrieval systems”，in Introduction To Information Retrieval，McGraw-Hill，pp 118-156，1983.