# 情報検索結果の知的提示のための自動要約ならびにインタフェースに関する研究

研究代表者　森 辰則　　横浜国立大学大学院・環境情報研究院
研究分担者　田村 直良　横浜国立大学大学院・環境情報研究院

**概 要**

本研究の目的は，情報検索の結果として得られた文書集合から，ユーザーの必要とする文書を効率よく選択するための情報ナビゲーション手法を提案することである．我々は検索文書間の関係から重要語を抽出するために，複数文書間の類似性構造というマクロな情報を，語の重要度というミクロな情報に写像する語の統計量に基づく新しい手法を提案している．昨年度までの研究で 1) 同手法に基づき文書分類による情報ナビゲーションと文書要約を同時に兼ね備えた情報ナビゲーションシステムを提案した，2) ナビゲーション過程や結果に現れる複数文書を対象とし，複数文書要約を生成する際の基本手法を検討し，特に，文間結束性の欠如をハニング窓関数に基づく文重要度計算手法の導入により改善する手法を提案した，3) より粒度の細かい複数文書要約において必要となる，特定の領域に依存しつつも精度良く文書からそのスキーマを抽出する手法を提案した．

本年度は，上記 2) に示す複数文書要約手法について，利用者の持つ複数の情報要求に同時に答える要約を生成するために質問応答エンジンを文重要度計算に利用する手法を検討するとともに，評価型ワークショップである NTCIR4 TSC3 において評価を行なった．また，3) については，活性度伝播に基づく事件記事の重要箇所抽出を検討し，実装システムに基づき定性的な評価を行なった．

## 1　はじめに

近年，検索エンジンのような情報検索システムが広く利用されるようになり，検索要求に関連のある文書を容易に得る事が出来るようになった．しかし，検索要求に関連性の低い文書を完全に排除できない，検索結果文書の構造化がなされていないなどの問題点がある．有効な方策としては，検索された各文書の要約の提示することによる元文書参照時間の削減，検索結果文書をクラスタリングすることによる関連文書と不要文書の分類などが従来提案されている．これらの方策に共通する必要不可欠な技術は検索結果文書集合を考慮した重要語の抽出である．その代表手法は検索要求中の語の重要度を高くする方法であり，これを自動要約に利用したものが Query-biased Summarization[TS98] である．この手法は直観的ではあるが，検索エンジンによる各種フィードバック等の工夫が反映されないという問題点がある．一方，我々は検索質問ではなく，検索文書間の関係から重要語を抽出することを検討している．これは，複数文書間の類似性構造というマクロな情報を，語の重要度というミクロな情報に写像する新しい手法である．この手法では，文書分類の過程で得られた文書間の類似性構造を適切に説明する度合を語の重要度とする．

昨年度までの研究では，(1) この重要語抽出手法が方法論として文書分類と自然に融合している点に着目し，文書分類に基づく情報ナビゲーションと文書要約を同時に兼ね備えた情報ナビゲーションシステムを提案・評価を行なった，(2) ナビゲーション過程や結果に現れる複数文書を対象とし，情報利得費に基づく語の重要度と MMR(Maximal Marginal Relevance) を統合した複数文書要約手法を検討し，文間結束性の欠如をハニング窓関数に基づく文重要度計算手法の導入により改善する方法について考察した，(3) より粒度の細かい複数文書要約においては，個々の文書の持つ情報構造を同一の枠組で捉える必要があるため，特定の領域に依存しつつも精度良く文書からそのスキーマを抽出する手法を提案・評価した．

上記 (2) で検討した複数文書要約手法における課題としては，利用者の持つ複数の情報要求に対して同時に答えられる要約文書の生成があった．特に複数文書要約においては内容が把握できるように，ある程度の要約文書量が必要となるため，利用者の知りたい事柄の各々について要約文書を生成すると最終的に利用者が読むべき文書量が増えてしまう．複数の要求の答を一度に概観ができることが望ましい．そこで，本年度は，上記 (2) に示す複数文書要約手法について，利用者の持つ複数の情報要求に同時に答える要約を

生成するために質問応答エンジンを文重要度計算に利用する手法を検討するとともに，評価型ワークショップである NTCIR4 TSC3 において評価を行なった

また，(3) については，活性度伝播に基づく事件記事の重要箇所抽出を検討し，実装システムに基づき定性的な評価を行なった．

## 2　複数の質問に焦点を当てた複数文書要約手法

### 2.1　はじめに

大量の文書が溢れている昨今，その中から必要とされる情報を効率良く見つけたいという要求がある．情報検索や質問応答等の技術により情報要求に関連する文書群や答え自身を容易に得る事が出来るようになりつつあるが，最終的には原文書を調べる必要がある．これらの技術と相補的な関係にあるのが，検索文書群を対象とした複数文書要約技術である．特に，近年,「質問の答に焦点を当てた要約」(Answer Focused Summarization) が注目されている [HSI01, WRF02]．これは，情報検索過程においては利用者が情報要求を持っており，また，それらが質問文として記述できるという考え方に基づく．複数文書要約においては内容把握ができるように，ある程度の要約文書量が必要であるので，利用者の知りたい事柄の一つ一つについて別々の要約文書を生成すると，最終的に利用者が読むべき文書量が増えてしまう．複数の要求の答とその背景知識を一度に概観できるような要約が生成できることが望ましい．

以上を踏まえて，本章では，複数の質問文に対応可能な文重要度計算法として質問応答エンジンの解のスコアを利用する手法を提案する．

### 2.2　提案手法の概要

本稿では，要約対象文書群は，情報検索等の結果として得られており，また，利用者の情報要求は，複数の質問文として与えられているとする．質問文については，利用者がシステムとの対話の中で一つずつ与えていき，その都度，その答を含む文脈をそれ以前の要約文書との関連を考慮しつつ要約していくという設定が自然ではある．しかし，本稿では第一次近似として，複数の質問文が同時に与えられることを想定し，利用者との対話の中での要約生成は今後の課題としたい．

この状況下では，複数文書要約のために，1)「情報要求を考慮した重要箇所抽出」，2)「文書間の冗長箇所の削除」，3)「文書間の相違点の抽出」が必要であると考える．提案手法では，これらについて，a) 質問応答エンジンの出力スコアに基づく文の重要度計算 (上記 1 に対応)，b) 語の出現分布に関する情報利得比に基づく文の重要度計算 (上記 1，3 に対応)，c)MMR に基づく要約文書中の冗長性の制御 (上記 2 に対応) を用いる．更に抽出文間の結束性の担保のために，d) ハニング窓関数に基づく文重要度平滑化を採用する．

本システムへの入力は要約対象となる日本語文書 (の ID) の集合，情報要求に対応する質問文の集合，ならびに，求める抜粋の長さ (文字数もしくは文数) である．出力は文書集合の抜粋 (文の列) である．例えば，今回の評価で用いたテストコレクションである NTCIR4 TSC3 の Topic 0500 (クローン羊ドリーに関する記事群) の場合，図 1 に示す 9 つの文書 ID，図 2 に示す 10 の質問文，ならびに，要約文書長 491 文字が入力である．この時，本システムは，まず，各文に対して (a) 質問応答エンジンに基づく重要度計算，ならびに，(b) 語の確率分布に関する情報利得比に基づく重要度計算を行ない，その結果を統合して重要度を求める．次に，出力される要約における文間の結束性を維持するために，ハニング窓関数を用いて文の重要度の変化を平滑化する．そして，MMR を用いて，重要度を考慮しつつも冗長性が少なくなるように文を順位付けする．順位付けられた文集合から指定された要約長に相当する上位の n 文を選択する．最後に原文書群のクラスタ構造と記事の日付順を考慮して選択した文を配置する．図 3 に出力抜粋の例を示す．この中でゴシック体になっている部分が先ほどの質問の答の一つである．

### 2.3　実験と評価

本節では，評価型ワークショップである NTCIR4 TSC3 における Formal Run の課題により提案手法に基づくシステムを評価する．NTCIR TSC は国立情報学研究所主催の文書自動要約に関する一連の評価型ワークショップである [FO01]．NTCIR4 TSC3 の報告会は 2004 年 6 月に開催された [HOFN04]．本稿では，1)

JY-19990402J1TYEUG0400060, JY-19990527J1TYMAJ1400040, JY-19980424J1TYMAK1400070, JY-19980723J1TYMAJ1400050, JY-19980301J1TYMAP1400050, 980110135, 980723029, 980424152, 980215018

図 1: 入力例: 文書 ID の集合 (NTCIR4 TSC3 Topic 0500)

「ドリー」は何の名前か？ ／クローン羊ドリーはどこで誕生したか？ ／クローン羊ドリーは、胎児細胞ではなく何の複製であることが実験で確認されたか？ ／クローン羊ドリーは何からつくり出されたか？ ／クローン羊ドリーの元になった雌羊の細胞とドリーの細胞とで、何が同一であると確かめられたか？ ／英国のロスリン研究所を率いているのは誰か？ ／クローン羊ドリーが妊娠中の羊から採った乳腺細胞をもとにつくりだされたことについて、どのような批判があったか？ ／クローン羊ドリーの細胞の寿命は普通の羊と比べてどうであることが分かったか？ ／クローン羊ドリーの出産はいつか？ ／クローン羊ドリーの出産により何が確認されたか？

図 2: 入力例: 質問文の集合 (NTCIR4 TSC3 Topic 0500)

８日付のイブニング・スタンダード紙によると、**英国のロスリン研究所**は、世界で初めて体細胞を卵子に組み込んで誕生させた**雌のクローン羊**「ドリー」（生後１年半）と雄の羊を交尾させたことを明らかにした。一昨年、世界初のクローン羊ドリーをつくることに成功した英エディンバラのロスリン研究所は**２３日**、ドリーがこのほど出産、**正常な生殖能力**があることが確認されたと発表した。詳細は二十三日発行の英科学雑誌「ネイチャー」に掲載される。ドリーは、**妊娠中の羊から採った乳腺（にゅうせん）細胞**をもとにつくり出された。この雌羊はすでに死んでいるが、組織の一部は、英国内で凍結保存されている。英国ロスリン研究所の**イアン・ウィルムット**博士のグループと、英国レスター大のグループは、クローン羊ドリーが**大人の雌羊**の体細胞核移植によるクローンであることをそれぞれ**ＤＮＡ**鑑定で裏付け、２３日発売の英科学誌「ネイチャー」に発表した。生まれたのは雄二頭と雌一頭で、いずれも元気だという。ドリーを誕生させた英国のロスリン研究所などのチームが二十七日発売の英科学誌「ネイチャー」で発表する。すでに二回妊娠し、元気な子供も産んだ。

図 3: 出力例: 要約文書 (抜粋)

モデル抜粋との比較による抜粋の性能，ならびに，2) モデル要約との比較による質問に対する解の被覆率に基づき評価を行なう．モデル抜粋とモデル要約はタスクオーガナイザにより準備がなされ，Formal Run の後に評価のために配布された．

同 Formal Run の課題は，30 トピックからなる．各トピックは，要約対象文書 ID のリスト (5〜19 文書)，トピックの表題 (検索要求を簡潔に表現したもの)，生成すべき要約文書の長さ (文字数，ならびに，文数．いずれも短いもの (Short, 要約率約 5%) と長いもの (Long, 要約率約 10%) の二種)，要約に含まれるべき事項を表した質問文のリスト (Short 用平均 7.6 文と Long 用平均 11.9 文の二種) から構成される．

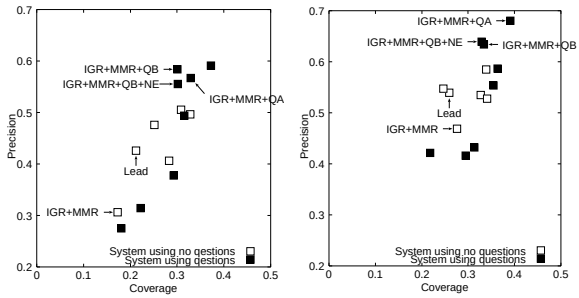
#### 2.3.1 重要文抽出の性能に関する評価

提案システムの出力抜粋の平均被覆率 (モデル抜粋中の文を再現できている割合) ならびに平均精度 (出力抜粋の各文がモデル抜粋に含まれる割合) を図 4(a)，(b) に示す．図中のラベル 'IGR+MMR+QA' は提案手法である．ラベル 'IGR+MMR' ならびに 'IGR+MMR+QB'，'IGR+MMR+QB+NE' は我々が用意したベースラインである．'IGR+MMR' は提案手法において質問応答エンジンによる文重要度を使わない場合に相当する．'IGR+MMR+QB' は Query-biased 手法に基づくベースラインであり，質問文中に含まれる語に追加の重みを与えるものである．'IGR+MMR+QB+NE' は 'IGR+MMR+QA' に加えて，NE の出現に重みを与えるものである．提案手法とこれらベースラインとの間の主な違いは質問応答エンジンの出力，すなわち，質問の答えに関する情報を使うか使わないかである．

一方，'Lead' はタスクオーガナイザが提供した Lead 手法 (各文書の先頭部分を抽出する) によるベースライン，それ以外の点は他の参加システムである．ただし，トピック情報中の質問文群の利用については，参加グループの判断に委ねられている．そのため，Lead 法を含め質問文群を利用していないシステムが存在することに注意されたい．

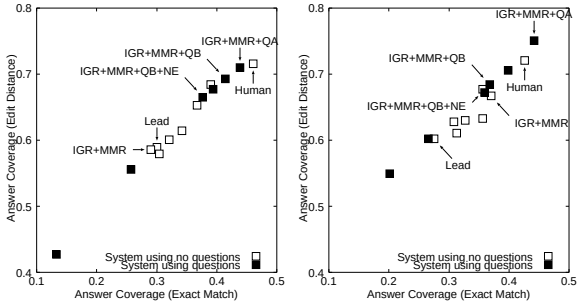
#### 2.3.2 質問に対する解の被覆率に基づく評価

各トピックについて，Short，Long の各要約文字数に対して，モデル要約に含まれる質問文の解が提案システムの出力抜粋に含有される度合 (解の平均被覆率) を調べた．図 5(a)，(b) に示す．尺度としては，正解文字列そのものが現れる割合の平均値 (Exact Match)，ならびに，正解文字列 $Ans_i$ と文 $S$ の間の編集距離に基づく尺度の平均値 (Edit Distance) の二種類がタスクオーガナイザにより提供されている．




(a) Short (b) Long

図 4: 抜粋の平均被覆率ならびに平均精度




(a) Short (b) Long

図 5: 質問に対する解の平均被覆率

### 2.4 考察

図 4(a) によると，要約長が短いとき (‘Short’) には，提案手法 (IGR＋MMR＋QA) は Lead 手法には勝っているが，ベースライン IGR＋MMR＋QB，IGR＋MMR＋QB＋NE とはほぼ同等である．つまり，質問文中の語だけでも抜粋生成について十分な情報があり，あえて解を求める必要はなさそうである．一方，要約長が長いとき (‘Long’) には，図 4(b) に示すとおり，すべてのベースラインならびに他参加システムに対して，その優位性が示されている．ただし，質問文の情報を利用しない参加システムもあることに注意されたい．QA エンジンを使わない IGR＋MMR と比較すると性能の改善は著しく，QA エンジンによる重み付けが非常に有効に機能していることがわかる．

ところで，Long については提案手法の抜粋精度が 0.680 と高いのに対して，抜粋被覆率は 0.391 と低い．これは，別の文書に由来する同一もしくは非常に似通った文を抽出する例が見受けられるためである．出力文書の冗長制御を行なっている MMI-MS では，名詞の重要度を成分とする文ベクトルの類似度を用いているが，各語の重要度は文書によって異なるために，全く同一の文であっても類似度が 1 にならない．文間類似度計算の精緻化が今後必要である．

次に質問の解の被覆率について考察する．図 5 によると，提案手法は各種ベースラインと比較して，Short，Long の要約長のいずれにおいても，改善されていることがわかる．

### 2.5 本章のまとめと今後の課題

本章では，複数の情報要求に対して一度に答えることができる複数文書要約を目標として，質問応答エンジンを用いた文重要度計算を汎用の文重要度計算に融合する手法を提案した．さらに，NTCIR4 TSC3 Formal Run に基づく評価により，その有効性を示した．

今後の課題としては，先に述べた MMI-MS における文類似度の精緻化の他に，質問の解解析の高速化が挙げられる．現在は，与えられた文書中の全ての形態素についてスコアを求めているため，平均的な PC を利用した時に 1 質問当り数十秒の処理時間がかかっている．

また，質問文の取り扱いについて，最初に多くの質問文が同時に与えられるという想定は第一次近似であることを本稿の最初で述べた．本来は，利用者がシステムとの対話の中で少しずつ質問を与えていき，その都度，その答を含む文脈をそれ以前の要約文書との関連を考慮しつつ要約していくという設定が自然である．そこで，質問応答における関連性のある一連の質問に答えるタスク [KFM04] との関連性を考察しつつ，対話型の要約システムの実現を検討したい．

## 3 活性度伝播に基づく事件記事の重要箇所抽出

### 3.1 はじめに

文章からの情報抽出について検討している．このとき，どのような観点から重要事項を判断するかが中心的なメカニズムとなる．また，文章の構成要素は種々の関係が影響しあい，抽出にはこの影響も考慮する必要がある．ここでは，枠組みとして「活性度伝播」を用い，特に，新聞の事件記事を対象に，記事の意味的

な情報としてもっとも重要なものの一つと考えられる「容疑者の行動」の抽出を検討する．

### 3.2 解析処理のモデル化

#### 3.2.1 事件記事のモデル化

文章を，いくつかの素性を持ち，言及される対象や概念に対応するノード (具体的には句に相当する) と，対象 (ノード) 間の関係 (リンク) ととらえ，グラフ構造としてモデル化する．この構造では，それぞれの関係ごとにグラフ構造が存在し，全体としては，それらを重ね合わせたものとなる．用いた素性を表 1 に示す．

表 1: 実験システムでの素性，関係，重み

| 素性 | | 重み | 素性 | | 重み |
|---|---|---|---|---|---|
| 動作 | 動作を表す概念 | 5 | 人 | 人を表す概念 | 0 |
| 物 | 物の相当する概念 | 0 | 主題 | 文中の主題を表す | 3 |
| 題述 | 文中の題述を表す | 0 | 動作主 | 述語に対して動作主格をあらわす | 5 |
| 対象 | 述語に対して対象格をあらわす | 0 | 位置 | 位置を表す概念 | 0 |
| 時刻 | 時刻を表す概念 | 0 | 日付 | 日付を表す概念 | 0 |
| 受け身 | 受け身の述語である | 0 | 文末 | 文末である | 2 |
| 関係 | | 重み | 関係 | | 重み |
| 格関係 | 格関係 | 5 | 語彙連鎖 | 語彙連鎖の関係 | 5 |
| 行動連鎖 | 連続する行動としての前後関係 | 5 | 修飾関係 | 後続する文の要素を修飾している | 0 |

#### 3.2.2 活性度伝播

前述のグラフ構造で表される意味構造においては，例えば，ひとつのノードについての注目の度合いは，それにリンクしている他のノードにも伝搬しうる．この注目の度合いを「活性度」と定義する．

活性度は関係に応じてリンクを辿り，次々と他のノードへ伝播していく．他のノードからの活性度の伝播もあり，以下の方程式を満たすようにバランスするものと考える．

$$l_{ij} = \sum_k c_{ijk} w_k \qquad v_i = \sum_k w_k a_{ik} + \sum_j v_j l_{ji}$$

ノード $i$ からノード $j$ へのリンクの「伝搬率」$l_{ij}$ は，ノード $i$ からノード $j$ へ関係 $k$ が存在するかどうか ($c_{ijk}$，1 または 0) と，その関係の重み $w_k$ から計算される．ノード $i$ の活性度 $v_i$ は，ノード $i$ の素性 $k$ の重み ($w_k$) 付き合計と，他のノードからの活性度の伝搬のから計算される．

結局のところ，ノードの活性度は，素性の有無の判定とノード間の関係判定の後，素性重み，関係の重みが決まれば連立方程式により，決定される．

#### 3.2.3 実現と事件記事での行動抽出，検討

各素性，関係の解析には,「南瓜」,「日本語語彙大系」,「文末，様相の解析」,「修辞構造理論」などの要素技術を用いている．システムは，VC++にて実現され，重みを調整できるスライドバーを持つ．スライドバーが操作されるごとに，活性度が計算され表示される文章の色の濃さの違いとして表示される．

図 6 は，新聞記事 (日本経済新聞 1993 年 2 月 8 日版から) を解析した結果である．活性度の高い要素ほど濃く印刷されている．

定性的な評価として，次がある．(1) 重要と思われる個所が活性度に応じて濃く表示されている，(2) 活性度計算は十分に早い，(3) 重みを決める戦略が不確定である，(4) 素性，関係の解析の精度が不十分であり，それが結果に影響している．

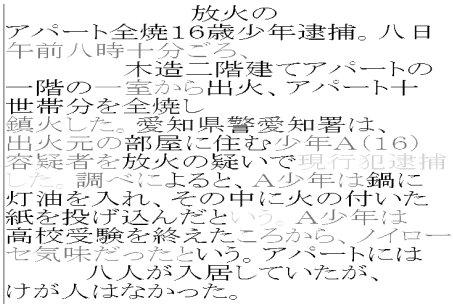
### 3.3 本章のまとめと今後の課題

ミクロ的には，各基礎技術の応用，それを統括するモデルとして活性度伝搬モデルを導入した．句をノードの単位とし，句間の関係により文章を構造化した．動的に解析できるが，重みの決定法に関して課題がある．

応用としては，各種開発のプラットホーム，視覚化による文章構造検討ツール，文章の読み上げ (音声合成の抑揚チューニング) が考えられる．

愛知・日進町，放火の疑い，アパート全焼１６歳少年逮捕．八日午前八時十分ごろ，愛知県愛知郡日進町の木造二階建てアパートの一階の一室から出火，(中略) 調べによると，A少年は鍋に灯油を入れ，その中に火の付いた紙を投げ込んだという．A少年は高校受験を終えたころから，ノイローセ気味だったという．アパートには三世帯，八人が入居していたが，けが人はなかった．

(a) 原文




(b) 出力

図 6: 活性度伝播に基づく事件記事の重要箇所抽出:出力例

## 参考文献


[FO01] Takahiro Fukusima and Manabu Okumura. Text summarization challenge: Text summarization evaluation in Japan. In *Proceedings of the NAACL 2001 workshop on Automatic Summarization*, pp. 40–48, 2001.

[HOFN04] Tsutomu Hirao, Manabu Okumura, Takahiro Fukushima, and Hidetsugu Nanba. Text Summarization Challenge 3 — Text summarization evaluation at NTCIR Workshop 4 —. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp. 407–411, June 2004.

[HSI01] Tsutomu Hirao, Yutaka Sasaki, and Hideki Isozaki. An extrinsic evaluation for question-biased text summarization on qa tasks. In *Proceedings of the NAACL 2001 workshop on Automatic Summarization*, pp. 61–68, 2001.

[KFM04] Tsuneaki Kato, Jun'ichi Fukumoto, and Fumito Masui. Question Answering Challenge for Information Access Dialogue — Overview of NTCIR4 QAC2 Subtask 3 —. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp. 291–296, June 2004.

[OY03] Paul Over and James Yen. An introduction to DUC 2003: Intrinsic evaluation of generic news text summarization systems. In *Proceedings of Document Understanding Conference 2003*, 2003.

[TS98] A. Tombros and M. Sanderson. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 2–10, 1998.

[WRF02] Harris Wu, Dragomir R. Radev, and Weiguo Fan. Towards answer focused summarization. In *Proceedings of the 1st International Conference on Information Technology and Applications*, 2002.


## 研究成果

### 研究発表

- 鈴木朋子, 田村直良: “音声物理量からの抑うつ傾向判定”, 電子情報通信学会論文誌, Vol. J87-D-II, No. 6, pp.1349-1358, 2004.
- Tatsunori Mori, Masanori Nozawa and Yoshiaki Asada: “Multi-Answer-Focused Multi-Document Summarization Using a Question-Answering Engine”, ACM Transactions on Asian Language Information Processing (TALIP), to appear.
- Tatsunori Mori, Masanori Nozawa and Yoshiaki Asada: “Multi-Answer-Focused Multi-Document Summarization Using a Question-Answering Engine”, The 20th International Conference on Computational Linguistics (COLING 04), pp.439–445, 2004.
- 上野友司, 森辰則, 木戸冬子, 中川裕志: “係り受けの 2 部グラフと共起関係を利用した同義表現抽出”, 言語処理学会第 10 回年次大会, 2004.
- 伊藤直之, Nikolay Elenkov, 森辰則: “情報検索ナビゲーションにおけるユーザへの提案機構と可視化インタフェース”, 言語処理学会第 10 回年次大会, 2004.
- 森辰則, 野澤正憲, 浅田義昭: “質問応答エンジンを利用した複数文書要約手法”, 言語処理学会第 10 回年次大会, 2004.
- Tatsunori Mori, Masanori Nozawa, and Yoshiaki Asada: “Multi-Document Summarization Using a Question-Answering Engine: Yokohama National University at TSC3”, Working Notes of the Fourth NTCIR Workshop Meeting, 2004.
- Tetsu Muramatsu and Tatsunori Mori: “Integration of PLSA into Probabilistic CLIR Model: Yokohama National University at NTCIR4 CLIR”, Working Notes of the Fourth NTCIR Workshop Meeting, 2004.
- 森辰則, 野澤正憲, 浅田義昭: “複数の質問に焦点を当てた複数文書要約手法”, 自然言語処理研究会報告 2004-NL-162, 情報処理学会, 2004.
- Tatsunori Mori: “Japanese Q/A System using A* Search and Its Improvement: Yokohama National University at QAC2”, Working Notes of the Fourth NTCIR Workshop Meeting, 2004.