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を，a) 従来手法で問題となった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，b) 利用者の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の出力を文重要度計算に導入する，という 2 点から精緻化した．また 3) についてはその解析の精度を向上させるために，動詞間の関係を解析するための辞書を自動的に生成する手法を検討した．

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は，評価の結果，文書数が比較的少ない時には有効であるものの，文書数が増加するに従って，文間の繋がりの良さ (結束性) が次第に悪くなる傾向にあった．また，これとは独立の課題として，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る．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を，(a) 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，(b) 利用者の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の出力を文重要度計算に導入する，という 2 点から精緻化した．また，上記 (3) についてはその解析の精度を向上させるために，動詞間の関係を解析するための辞書を自動的に生成する手法を検討した．

2　ハニング窓関数に基づく複数文書要約における文間結束性の改善

2.1　概要

本部分課題では，文書分類をされた後の複数文書を対象とし，要約を提示する一手法を提案する．複数文書から内容を網羅した要約を作成する為に，情報利得比 (IGR) を用いた語の重要度と MMR(Maximal Marginal Relevance) の統合による重要文抽出を行なう．また，抽出した重要文の間の結束性を高めるために，ハニング窓関数を用いる．そして，本手法を内容の網羅性と可読性の観点から検討する．

2.2　情報利得比に基づく語の重要度と MMR の統合による重要文抽出

図 1 に情報利得比に基づく語の重要度と MMR の統合による重要文抽出手法の概略を示す．詳細については前年度の報告書を参照されたい．これは，「情報利得比に基づく文の重要度の導出」と「MMR に基づく要約文中の冗長性の制御」の二点を次式により統合した重要文抽出手法である．

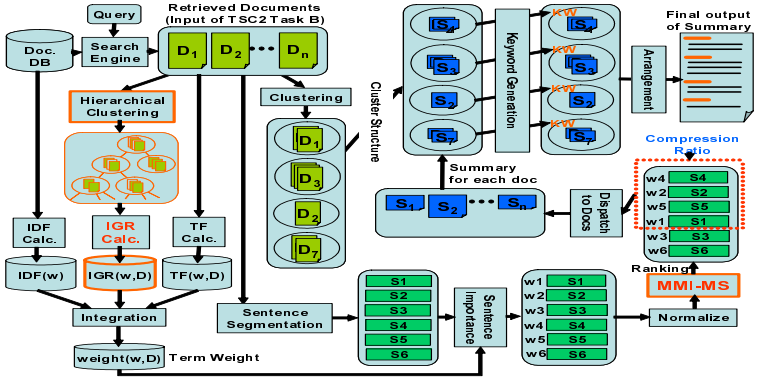

図 1: 情報利得比に基づく語の重みと MMR に基づく重要文抽出

$$MMI\text{-}MS(SS, A) \stackrel{\text{def}}{=} Arg \max_{S_i \in SS \setminus A} [\lambda \cdot Imp_{IGR}(S_i) - (1-\lambda) \max_{S_j \in A} Sim_s(S_i, S_j)] \quad (1)$$

式 (1) は初期文集合と既に選択されている文の集合を与えると，次に選択すべき文を返す関数である．ここで，SS は要約対象となるすべての文の集合，A は既に選択された上位の文の集合，$Imp_{IGR}(S_i)$ は文 S_i の IGR に基づく重要度，Sim_s は文間の類似度を表す尺度である．

2.3　ハニング窓関数による文重要度平滑化

前節で述べた手法について，NTCIR3 TSC2 課題 B[TSC 02] による評価実験を行なった結果，対象文書数が多い時には，多くの文書から少しずつ重要文を抽出し，文間の結束性が低下する傾向が見られた．そこで，重要な文がほぼ収まるような文の数を設定し，その範囲内で重要度が滑らかに変化するようにハニング窓関数を用いて重要度の平滑化を行なうことを提案する．

窓の幅 (重みを与える範囲) を W，窓の中心位置を l とすると，ハニング窓関数 $h_l(i)$ は式 (2) により与えられる．

$$h_l(i) = \frac{1}{2}(1 + cos 2\pi \frac{i-l}{W}) \qquad (|i-l| \leq \frac{W}{2}) \quad (2)$$

黒橋ら [黒橋 96] はハニング窓関数を用いて，語のテキスト中での出現密度分布を調べ，その高密度な出現位置を取り出す事によって，その語の重要説明箇所を特定する手法を提案している．栗山ら [栗山 02] は，文書中から複数の語を選定し，ハニング窓関数を用いて複数の語の出現密度分布を調べ，その高密度な出

現位置を重要文として抽出し，抄録を作成する手法を提案している．これらの手法では，ハニング窓関数における横軸の単位を一語としているが，我々の手法では，横軸の単位を一文として考える． すなわち，指定した窓幅 W に含まれる全ての文 S_i について，文 S_i の重要度 $Imp(S_i)$ とハニング窓関数の値 $h_l(i)$ の畳み込み和 (次式) を求め，これを，結束性を考慮した文 S_l の重要度 $Imp_C(S_l)$ とする．

$$Imp_C(S_l) \quad = \quad \sum_{i=l-\frac{W}{2}}^{l+\frac{W}{2}} h_l(i) \cdot Imp(S_i) \tag{3}$$

ここで，複数の要約を対象に窓幅を変化させて予備実験を行なった結果，文間の結束性が感じられ，可読性 (要約の内容の理解のしやすさ) が最も高かった事から，窓幅を 8 文に決定した．

2.4 実験と評価

2.4.1 NTCIR TSC2 課題 B による順位付け評価

ハニング窓関数を用いないシステムにより NTCIR3 TSC2 課題 B[TSC 01] に参加し，評価を行なった．同 Formal Run は，30 のトピックから構成される． 要約の長さには，長い要約 (Long) と短い要約 (Short) の 2 種類がある．順位づけ評価では，各トピック，各要約文書長に対して，評価対象システムの要約以外に，baseline として次の 3 種類の要約が用意されている: (1) 人間による自由作成要約 (表中では '人手')，(2) lead 法による要約，(3) stein 法による要約 [SSW99]．要約評価者は，各要約文書と原文書集合とをトピック毎に読み，内容の網羅性 (C) と可読性 (R) の観点から要約文書の順位づけを行なう．つまり，評価対象のシステムの順位が上位 (小さい値) ほど良い．表 1 に対象文書数が少ない 15 トピック (7 文書以下) に関する結果を示す．文書数が少ない時には同表の示すように baseline と比較して評価が高かった．一方，文書数が多い残り 15 トピックでは評価が勝っているとは言えなかった．

表 1: baseline との優劣 (7 文書以下の 15 トピック)

	C Short			R Short			C Long			R Long		
	W	L	T	W	L	T	W	L	T	W	L	T
v.s. 人手	6	9	0	6	9	0	7	8	0	5	10	0
v.s. Lead 法	**11**	**4**	**0**	6	9	0	**10**	**5**	**0**	7	8	0
v.s. Stein 法	**13**	**1**	**1**	7	8	0	**10**	**4**	**1**	3	12	0

W:win L:lose T:tie

2.4.2 ハニング窓関数を組み込んだシステムに対する順位づけ評価

NTCIR3 TSC2 において baseline として用いられた Lead 法と Stein 法の出力結果とソースコードは現在のところ公開準備中であり入手できなかったが，人手による自由作成要約は公開されている為，次の 3 種類の要約についてトピック毎に順位づけ評価を行なった: (1) 人間による自由作成要約，(2) ハニング窓関数を導入した手法による要約，(3) NTCIR3 TSC2 課題 B 時の手法による要約 (表中では 'Formal Run')．要約評価は大学院生 4 人と大学生 2 人で行なった．なお，可読性に関する尺度として,「読みものとして，要約が述べる内容の理解のしやすさ」と定義し，被験者に周知した．

評価結果の内，Formal Run の時に評価が悪かった，対象文書数の多い 15 トピック (8 文書以上) に対する結果を表 2 に示す．

表 2: baseline との優劣 (8 文書以上の 15 トピック)

	C Short			R Short			C Long			R Long		
	W	L	T	W	L	T	W	L	T	W	L	T
v.s. 人手	0	15	0	0	15	0	0	15	0	1	14	0
v.s. Formal Run	3	6	6	**5**	**1**	**9**	**10**	**3**	**2**	**9**	**3**	**3**

W:win L:lose T:tie

2.5 考察

表 1 に示した NTCIR3 TSC2 課題 B の実験による評価結果より，我々のシステムは baseline である lead 法，stein 法と比較して，内容の網羅性においては優れていると考えられる．特に，対象文書数が少なく，要約率が小さい時においてその傾向が顕著に表れている．この結果は，IGR による語の重みづけと MMR の統合に基づく重要文抽出が効果的である事を示す．これに対して，上記と反対の状況では評価が格段に下がる．特に対象文書数が多い場合には，文間の結束性が低下する事が一因であると考えられる．

表 2 より，ハニング窓関数を組み込み，文間の結束性を向上させる事で，要約率が大きい時の内容の網羅性と，全トピックに亙っての可読性に対して，要約の品質が改善されていることがわかる．また，可読性 (R Short，R Long) に関しては対象文書数によらず，改善されるが，内容の網羅性に関しては，対象文書数が多く，要約率が大きい場合 (C Long) に大きく改善される傾向が見られる．また，対象文書数が少なく，要約率が小さい場合 (C Short) に関しては，多少評価が下がった．

2.6 本部分課題のまとめと今後の課題

本部分課題では，文書分類をされた後の複数文書を対象とし，情報利得比を用いた語の重要度と MMR の統合による重要文抽出を行なう事で内容の網羅性を考慮した要約を提示するシステムを提案した．NTCIR3 TSC2 による評価においては，我々のシステムは，内容の網羅性を考慮した複数文書要約を作成するにあたり，特に要約率が小さい時と，対象文書数が少ない時に効果的である事が示された．また，文間の結束性を高めるためにハニング窓関数による文重要度平滑化手法を導入し，人手による評価を行なったところ，組み込む前のシステムと比較して，内容の網羅性に関して，対象文書数が多い時に効果的である事が示された．

今後の課題として，ハニング窓関数の窓幅を対象文書数と要約率に対応するように自動的に調節し，より精度の良い重要文書部分の抽出を行なう事が考えられる．

3 質問応答エンジンを利用した答えに焦点を当てた重要文抽出

3.1 概要

本部分課題では，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を検討している．文書要約の目的は，原文書よりも少ない読書量で内容が「理解」できる短い文書を生成することにある．その際，原文書の概要全体を把握することが「理解」であることが一般的ではあるが，利用者が情報要求を持っている場合にはそれが満足されることが「理解」であると考えられる．後者の状況において，情報要求は「質問」として記述できるという考え方から，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．平尾ら [平尾 01] は質問型に一致する固有表現に高い重みを与え要約文書を生成する手法を提案している．Wu ら [WRF02] は質問型と要約生成の単位となる文書断片の長さの関係を検討し，これに基づいた要約文書生成手法を提案している．Gaizauskas らのプロジェクトでは，複数文書要約と質問応答を組み合わせたシステムを目指すとされている [Gai03] が，詳細は公開されていない．

ここでは，複数の情報要求が複数の質問文として与えられている状況において，対象となる複数文書を要約する手法について検討する．特に我々が開発している質問応答エンジンが保持する内部情報を利用する．このエンジンは，質問応答の過程において各語 (形態素) に解としての「良さ」を表すスコアを付与するので，このスコアを利用することにより，質問型の情報のみを利用する手法よりも精度の高い要約生成を行なえると期待される．

3.2 質問応答エンジンを利用した文の重要度計算

我々の提案している質問応答システムは，様々な文書や検索エンジンを利用できるようにするために，対象文書に対する前処理を必要としない仕組みとなっている．質問文が与えられた後に，形態素解析や構文解析，固有表現抽出などといった計算コストの大きい処理を行なうが，これを必要最小限に抑え実時間処理ができるようにするために，A^* に基づく解の探索制御を行なっている [MOFK03]．

このシステムは質問文が 1 つ与えられると対象文書中の各語に対して解としての「良さ」を表すスコアを付与する．我々はこのスコアを質問の解に注目した時の語の重要度と考え，重要文抽出を行なうことを提案する．ここでは，複数の質問文が与えられることを想定しているため，各形態素にはスコアの組が与えら

れることになる．各スコアの値はある質問文に対応する．質問文の複雑さや質問の型によりスコアの値が異なるため，本来，異なる質問文に対して求められたスコアは比較可能ではない．しかし，ある形態素について複数の質問文に対する単一の重要度を付与したいので，元のスコアを比較可能な値に正規化する．各質問文のスコアについて，その絶対値は問題ではなく，平均値からの隔たりが重要であると考え，偏差値(T-score) に変換した．これにより，複数の質問文に亙って平均値が同一になる．質問文 q に関する形態素 w の正規化スコアを $score_n(w,q)$ とする時，文 S_i の重要度 $Imp_{QA}(S_i)$ を式 (4) で求める．ただし，Q は与えられた質問文の集合，W_{S_i} は文 S_i に現れる形態素の集合である．いずれかの質問の答えが含まれているか否かと言う観点から文の重要度を決めるため，式 (4) ではある文の重要度をその文に含まれる形態素の最大スコアとしている．この重要度を式 (1) に組み込むと式 (5) となる．

$$Imp_{QA}(S_i) = \max_{w \in W_{S_i}, q \in Q} score_n(w,q) \tag{4}$$

$$MMI\text{-}MS(SS,A) \stackrel{\text{def}}{=} Arg \max_{S_i \in SS \setminus A} [\lambda(\alpha \cdot Imp_{IGR}(S_i) + (1-\alpha) Imp_{QA}(S_i)) - (1-\lambda) \max_{S_j \in A} Sim_s(S_i, S_j)] \tag{5}$$

3.3 本部分課題のまとめと今後の課題

本部分課題では，利用者の持つ情報要求を複数の質問文という形で受け取り，複数文書要約に反映させる手法として，質問応答エンジンのスコアを利用する方法を提案した．同手法の評価は，報告書執筆時点でタスク評価が行なわれている NTCIR4 TSC3 において行なう予定である．

4 新聞記事中の行動抽出のための動詞間関係推定辞書

4.1 概要

本部分課題では，新聞の事件記事中の人物の行動を記述する動詞間[1] の関係解析のための辞書生成を目的とし，抽出パターンによりコーパスから有用な動詞間関係を抽出する．

我々は，十分重要で，かつ多量にある事件記事を扱っている．事件記事についての本来の意味とは，容疑者はじめ事件に関わる人々についての情報，人々の行動，犯罪の場合の動機，供述などで，犯罪スキーマとしてモデル化して昨年度報告した．その中で容疑者の名前や被害者の年齢などは一つの値として二次利用可能な形で比較的容易に抽出可能であるが，事件に関わる人々の行動については被害者，加害者，ある場合には警察などが相互にリンクし合う動作の記述列として，さらに構造化されるべきである．

そうした行動情報の抽出において，我々は，時刻が明示的に示されている部分で文章を区切り，これを時間セグメントと呼んで時間セグメント間の整列化を行い，時間セグメント内の行動については行動を表す文を手がかり語により構造化 (整列化) およびデフォルトとして手がかり語のない接続関係について順接の関係として，行動情報の意味抽出としてきた．

しかし，手がかり語やデフォルトの扱いによる文間の関係解析には限界がある．一方，たとえば,「刺される」,「死亡する」に対しては，事象発生の時間の順序関係，事象間の因果関係が自然に推論できるように，ある種の動詞については，二つの動作を記述する動詞間の関係がある程度明白なものが存在し，これにより動作間の関係解析を進めることが可能である．さらに，この種の関係は，辞書として知識ベース化することが必要である．

4.2 動詞間関係辞書生成

犯罪行為動詞対抽出システムは，(1)JUMAN による形態素解析，(2) 接続表現のよる重文の分割，(3) 品詞情報によるタグづけ，(4) パターン・マッチングによる動詞対抽出，(5) 犯罪行為動詞の確認の順で解析を進める．実際のパターンは，Perl の正規表現により記述されている．

動詞対抽出に用いたパターンは,「XXX は，YYY ため ZZZ」，および「XXX は YYY ので WWW を ZZZ」なる形式を表しており，品詞に関して制約が設けられ，動詞対 (YYY,ZZZ) ならびにそれらの間の関係を抽出する．

[1] コーパス (日経新聞 1993 年版) 一年分の記事 (約 18 万件) から「容疑者」なる文字列を含む記事約 1700 件を抽出し，さらにその中から格フレームを構成している動詞を抽出し，特に，被害者，警察，容疑者の行動で使われそうな動詞を 4 名の多数決により分類した．

動詞数	(サ変名詞数)	被害者		容疑者		警察		その他	
3165	(1439)	20	(11)	275	(118)	78	(44)	2792	(1266)

表 3: 抽出された犯罪行為動詞対の一部

前事象	後事象	前事象	後事象	前事象	後事象
切り付ける	刺す	ごまかす	供述する	盗み出す	偽造する
持ち出す	切り付ける	密売する	供述する	自供する	供述する
密入国する	供述する	供述する	供述する	供述する	奪う

4.3 本部分課題のまとめと今後の課題

本研究では，新聞の事件記事において容疑者など登場人物の行動解析のための動詞間関係辞書構築を目指し，コーパスからの (犯罪行為) 動詞対抽出について述べた．

犯罪行為動詞対抽出のためのパターンの数は，現在 16 個であるが，さらに追加，精密化する必要がある．その際，格にとなる名詞についての制約の導入が有効と思われる．

参考文献

[Gai03] Robert Gaizauskas. Cubreporter web page. http://nlp.shef.ac.uk/cubreporter/index.html, 2003.

[MOFK03] Tatsunori Mori, Tomoharu Ohta, Katsuyuki Fujihata, and Ryutaro Kumon. An A* search in sentential matching for question answering. *IEICE Transactions on Information and Systems*, Vol. E86-D, No. 9, pp. 1658–1668, September 2003. Special Issue on Text Processing for Information Access.

[SSW99] Gees C. Stein, Tomek Strazalkowski, and G. Bowden Wise. Summarizing Multiple Documents using Text Extraction and Interactive Clustering. In *Proceedings of the sixth Pacific Association for Computational Linguistics (PACLING 99)*, pp. 200–208, 1999.

[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.

[TSC 01] TSC 実行委員会. NTCIR 3 テキスト自動要約タスク (automatic text summarization task)/TSC-2(text summarization challenge 2). http://lr-www.pi.titech.ac.jp/tsc/tsc2.html, 2001.

[TSC 02] TSC 実行委員会. NTCIR 3 テキスト自動要約タスク (automatic text summarization task)/TSC-2(text summarization challenge 2). http://lr-www.pi.titech.ac.jp/tsc/tsc2.html, 2002.

[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.

[栗山 02] 栗山義明, 絹川博之. ターム群の出現密度分布を用いた重要文抽出方式—ターム種別重みの評価実験—. 第一回情報科学技術フォーラム講演論文集 (FIT2002), pp. LE–9, 2002.

[黒橋 96] 黒橋禎夫, 白木伸征, 長尾真. 出現密度分布を用いた語の重要説明箇所の特定. 自然言語処理研究会報告 96-NL-115-7, 情報処理学会, 1996.

[平尾 01] 平尾努, 佐々木裕, 磯崎秀樹. 質問に適応した文書要約手法とその評価. 情報処理学会論文誌, Vol. 42, No. 9, pp. 2259–2269, 2001.

研究成果

研究発表

- Tatsunori Mori, Tomoharu Ohta, Katsuyuki Fujihata and Ryutaro Kumon: “An A* Search in Sentential Matching for Question Answering”, IEICE Transactions on Information and Systems, Vol. E86-D, No. 9, 2003.
- 中川 裕志，湯本 紘彰，森 辰則: “出現頻度と連接頻度に基づく専門用語抽出”, 自然言語処理, Vol. 10, No. 1, pp.27–45, 2003.
- Hiroshi Nakagawa and Tatsunori Mori: “Automatic Term Recognition based on Statistics of Compound Nouns and their Components”, Terminology, Vol. 9, No. 2, 掲載予定.
- 金山淳一, 田村直良: “犯罪スキーマに基づく新聞記事からの事件情報抽出”, 言語処理学会第 9 回年次大会,pp.73–76, 2003.
- 北條孝, 田村直良: “語の分布と主題の展開に着目した文章の構造化”, 言語処理学会第 9 回年次大会,pp.605–608, 2003.
- 太田 知宏, 藤畑 勝之, 公文 隆太郎, 森 辰則: “質問応答における探索制御付き命題照合の精度向上”, 言語処理学会第 9 回年次大会,pp.621–624, 2003.
- 佐々木 拓郎, 野澤 正憲, 森 辰則: “情報利得比に基づく語の重要度と MMR の統合による複数文書要約”, 言語処理学会第 9 回年次大会,pp.198–201, 2003.
- 高野 祐介, 村松 哲, 森 辰則: “空間分割型 PLSI を用いた言語横断情報検索”, 言語処理学会第 9 回年次大会,pp.389–392, 2003.
- 春日隆緒, 田村直良: “文章のセグメント間関係解析に基づく文章構造解析”, 情報処理学会 自然言語処理研究会 2003-NL-155, pp.59–64, 2003.
- Tatsunori Mori: “Recent Research at Mori laboratory, Yokohama National University”, Exhibiton Brochure of the 41st Annual Meeting of the Association for Computational Linguistics,p.23, 2003.
- 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”, 情報処理学会自然言語処理研究会 2003-NL-159, 発表予定.