様式 C-19

科学研究費補助金研究成果報告書

平成２１年５月２５日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：19500118
研究課題名（和文）　利用者との対話に基づく複数文書要約手法に関する研究

研究課題名（英文）　Study of multi-document summarization based on interaction with users

研究代表者
森　辰則（MORI TATSUNORI）
横浜国立大学・大学院環境情報研究院・教授
研究者番号：70212264

研究成果の概要：
本研究では、大量の文書に対する情報アクセス技術という観点から、文書群に対して生成された要約文章をディスプレイ上に電子的に提示したものを対話のインタフェースとすることを提案した。特に、システムが行う文書群の内容提示のみならず、利用者が行う情報要求の指示も、電子的に提示された要約文章の上で統合的に行う仕組みを検討した。また、動向情報を含む文章の要約を目的として、動向情報に関する情報抽出に関する検討を行った。動向情報は統計量名と値の組により表現されるので、統計量名の部品を構成する要素を定義し、これらを抽出する手法を提案した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	2,100,000	630,000	2,730,000
2008 年度	1,300,000	390,000	1,690,000
年度			
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：(1)自動要約, (2) Scatter/Gather, (3)関連性フィードバック, (4)ユーザインタフェース, (5)統計量抽出, (6)画像、文章、音声等認識

１．研究開始当初の背景

ICT 技術の普及に伴い、膨大な量の文書情報が入手可能となったが、一方で利用者が欲する情報に容易に到達可能であるとは言い難い。例えば、検索エンジンにより関連文書の効果的な絞込みが可能になりつつあるが、現在の検索エンジンは文書の順位付けを行うだけであるので、真に必要とする情報を得るためには文書の中身を精査する必要がある。そのため、情報を得るまでに利用者が読むべき文書を少なく抑え作業効率を向上させる研究が行われている。例えば、文書から不要な部分を削りより短い文書を生成する自動要約や、利用者が入力した質問文に対し知識源となる文書群からその答を見つけ出す質問応答などがある。一般に、利用者が詳細な情報要求を持っていない場合に適するのが自動要約である。

初期の研究では、少数の文書群が得られている状況を出発点とし、単文書要約技術により各文書を個別に要約し、その各々を提示す

ることにより利用者が行う情報取得の一助としていた。近年、要約技術が進展を見せており、複数の文書を統合して要約し、一つの文章にまとめる技術が提案されている。いずれの場合も、利用者の情報要求が反映されるのは、要約対象となる文書の選定過程にあり、文書検索等、利用者の情報要求を扱う情報アクセス技術との組合せが必要である。そのため、要約生成に際しても情報要求を考慮する手法が提案されている。

ここで注意すべきは、これらの研究において、要約文章は出力として位置づけられている点である。つまり、生成された要約文章に対し利用者が行う動作として、読むことだけが想定されている。一方で、利用者が真に行いたいことは、大量の文書集合が与えられたときに、情報要求に関連する文書部分を過不足無く読むことである。よって、望まれる技術は、文書を読むことに利用者が専念できるように支援するものである。提示した要約文章に、不足している情報や興味のある関連情報が存在するときには、要約文章に対する直接的かつ簡単な操作で、次々と適切な新しい(要約)文書が生成でき、よどみなく文書情報を読める仕組みが必要である。しかしながら、自動要約をこのような、情報アクセスインタフェースの観点から研究している事例はない。

２．研究の目的

本研究では、大量の文書に対する情報アクセス技術という観点から、自動要約技術に焦点を当てる。特に、文書群に対して生成された要約文章をディスプレイ上に電子的に提示したものを対話のインタフェースとすることを提案し、システムが行う文書群の内容提示のみならず、利用者が行う情報要求の指示も、電子的に提示された要約文章の上で統合的に行う仕組みを検討する。

特に情報検索の一手法である Scatter/Gather 法を要約文章提示の観点から整理し直すことにより、利用者が行うべき作業、すなわち、概観する文書を読むことと情報を絞込むことの両者について、要約文章に対する操作に集約できるという方法を提案し、検討する。すなわち、Scatter 過程を、文書集合全体を概観する要約文章を生成し、利用者に提示することと捉え直し、Gather 過程を、提示された要約文章のうち、情報要求に適合する文章部分に利用者がマウス操作等により自由に印を付与することと考える。システムは印が付与された文章部分に関連する文書(あるいは文書部分)を集め、これらを改めて要約対象文書として、複数文書要約を行うことを繰り返す。Scatter/Gather 法と自動要約手法を組み合わせる従来手法においては、情報アクセスの途中に現れる情報(キーワード)と、最終的に利用者が取得すべき情報(文章)が乖離していた。これに対し、本研究で検討する上述の方法では、両者に境界が無いので、最終段階という概念が無く、要約文章を読み進めている最中に利用者の情報要求が満足されれば、そこで文章を読むことをやめるだけでよい。利用者は要約文章を読み進め、重要箇所に印をつける作業をするだけである。これを紙面の上で文章を読み進める状況に対応させて、比喩として説明をすると次のようになる。一枚の紙の上に記されている要約文章を利用者が読み進め、必要に応じて重要箇所に蛍光ペンで印をつけていくと、それに応じて焦点を変化させた新しい要約文章を記した紙が得られる。新しく得られた紙について同様の作業を繰り返すことにより、情報を読み進めることができる。この仕組みは、前節で述べた望まれる技術の多くの部分を実現するものである。

３．研究の方法

(1) 大量の文書に対する情報アクセス技術という観点から、文書群に対して生成された要約文章をディスプレイ上に電子的に提示したものを対話のインタフェースとすることを提案した。特に、システムが行う文書群の内容提示のみならず、利用者が行う情報要求の指示も、電子的に提示された要約文章の上で統合的に行う仕組みを検討した。
(2) また、動向情報を含む文章の要約を目的として、動向情報に関する情報抽出に関する検討を行った。動向情報は統計量名と値の組により表現されるので、統計量名の部品を構成する要素を定義し、これらを抽出する手法を提案した。

４．研究成果

(1) 対話型複数文書要約手法に関する検討

①概要

対話型複数文書要約システムは Scatter/Gather 法を要約提示の観点から捉え直す事により、提示した要約文章そのものに対し利用者が操作を行い、それによって利用者の興味を反映した新たな要約を提示するインタフェースである。図１に対話型複数文書要約システムの全体像を示す。

以下の手順に従って処理が行われる。
【手順 1】利用者は任意のキーワード列を入力し、文書データベース中の文書を検索する。
【手順 2】状況により、検索結果の中から要約したい文書群を利用者に選択してもらい、初期文書集合とする。
【手順 3】与えられた文書集合の要約(主要約文章)を生成し、表示する。要約処理過程で

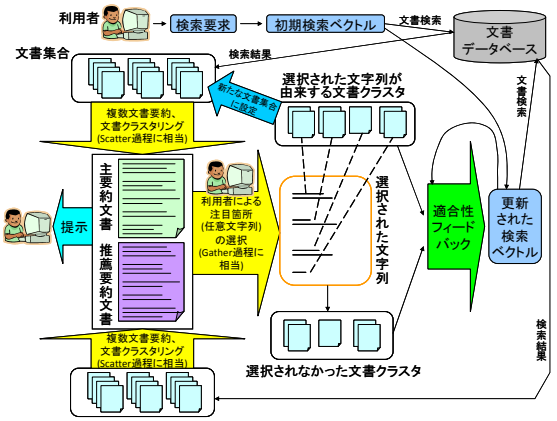


図 1 対話型複数文書システムの概要

は文書集合のクラスタリングが行われる。(Scatter 過程に相当)

【手順 4】表示された要約文章(主要約文章、ならびに、二巡目以降は推薦情報要約文章も含む)の中から、利用者は興味にしたがって自由に文字列を選択する。(Gather 過程の一部に相当)

【手順 5】利用者の選んだそれぞれの文字列の属するクラスタをまとめて次の要約処理の対象とする。なお、状況に応じて利用者は、それぞれの文字列の由来する文書のみを次の要約対象とすることができる。(Gather 過程の一部に相当)

【手順 6】手順 2 で選択した文書集合を、手順 4 で選択された関連文書集合と、それ以外の文書集合(非関連文書集合に対応する)に分割する。これらを用いて、適合性フィードバック手法により検索ベクトルを更新し、関連推薦文書の検索を行う。さらに、それらの要約文章(推薦情報要約文章)を生成する。

【手順 7】手順 2 に戻り、上記の推薦情報要約文章とともに、次の主要約文章を表示する。

上記の手順2から6 までを繰り返すことにより、利用者の興味に応じて文書集合が絞り込まれ、対応する要約文章が生成される。

②インタフェースの概要

図 2～4 にインタフェースの画面を示す。まず、図 2 に示すように、記事検索のキーワードを入力し、文書データベースから検索を行う。次に、中段の検索された文書のタイトルが表示され、この中から要約対象とする文書集合を選択する。(手順 1、2 に対応)

図 3 は、図 2 で選択した文書集合について要約を行った画面である。主要約が表示され、利用者はこの主要約の中からでもっと知りたい部分を選択する。(手順 3、4、5 に対応)

図 4 では、図 3 で選択された文字列をもとに、主要約および推薦情報要約が表示される。(手順(6)、(7))

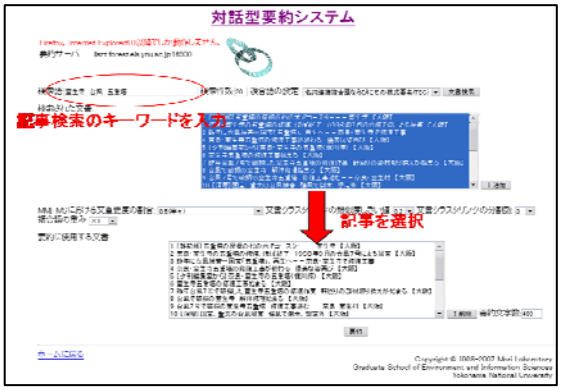


図 2 インタフェース画面 1

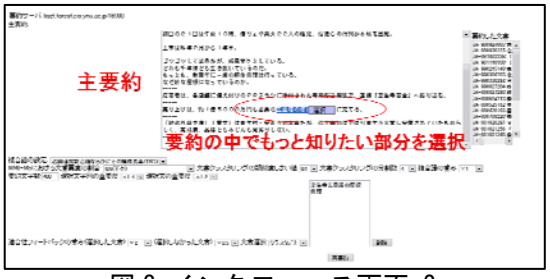


図 3 インタフェース画面 2

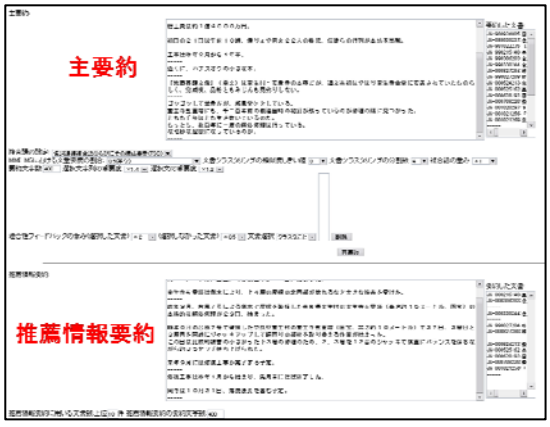


図 4 インタフェース画面 3

③複数文書要約における重要度計算

主要約および推薦情報要約を生成する際に、複数文書要約を行う。要約における文重要度を以下の語の重要度を用いて計算する。

TF, IDF…語の出現頻度、文書頻度による重要度

ICF, IGR…語のクラスタ頻度、クラスタ分割による情報利得比に基づく重要度

これらに加え、選択した文字列や複合語に対してバイアスをかける。

これらの尺度によって計算された重要度を用い、MMI-MS により重要文を決定する。MMI-MS とは文の重要度と要約の冗長性を同時に考慮する重要文抽出手法である。要約文章生成法としては、要約対象文書全体に対し MMI-MS を行う従来手法と本手法で得られたクラスタ構造を重視し、クラスタごとに

MMI-MS を行う提案手法について検討する。

④評価実験および考察

提案システムの出力した要約を評価するために評価実験を行った。評価型ワークショップ NTCIR TSC3 Formal Run のタスクから無作為に選んだ 3 トピックについて、システムが出力した要約を正解要約に対する ROUGE-1 値により評価を行った。ROUGE-1 とは、1-gram の再現率による評価方法である。

【トピック 1】ニュートリノに質量があるとされることに関する記事群

【トピック 2】インディペンデンス艦載機の夜間離着陸訓練（ＮＬＰ）に関する記事群

【トピック 3】台風によって壊れた室生寺（五重塔）に関する記事群

これらのトピックについて、要約生成を行った。要約文字数は 400 字、クラスタ数は 6～8 とした。選択する文字列は、外的評価に用いられる質問で聞かれている内容とその答とし、要約が収束するまで再要約を行った。

複数文書要約手法について、以下の各手法を比較検討した。

【BASE】MMI-MS により複数文書要約を行う手法(ベースライン)

【TFIDF】各クラスタの文書に対して MMI-MS により複数文書要約を行う手法

【ICF】TFIDF に加え、重要度計算に ICF を用いる手法

【IGR】TFIDF に加え、重要度計算に IGR を用いる手法

【IGR＆ICF】TFIDF に加え、重要度計算に ICF、IGR を用いる手法

それぞれの手法において、最も ROUGE-1 の値が高かったクラスタ数の場合の結果を図 5～7 に示す。図中の ICF_8 とはクラスタ数が 8 のときに ICF の手法による結果を意味する。

いずれの場合も、クラスタ構造を考慮しないベースラインと比較して、クラスタ構造を考慮する提案手法のほうが優っている。トピック 1、2 においては、要約が収束したときの値はベースラインが最も低い結果となった。また、トピック 3 においては要約が収束したときの値は ICF&IGR が最も低かったが ICF&IGR の 7～10 回目の要約ではベースラインの最高値を上回っている。ベースラインよりも各提案手法が良い結果であることは、全体の文書集合に対して MMI-MS により冗長性排除を行うよりも、クラスタリングの結果を重要視し、各クラスタに対して MMI-MS により冗長性排除を行う方が、全体を概観できる要約が生成できることを示している。

⑤今後の課題

クラスタごとに複数文書要約を行う手法の優位性を示すことができたが、一方で、この場合、それぞれのクラスタから出力された要約間の冗長性については考慮されていない。評価実験の方法についても、改善の余地がある。また、対話の観点での評価を行う必要がある。

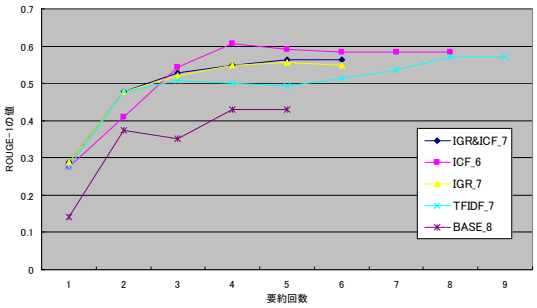


図 5 トピック 1 に対する評価

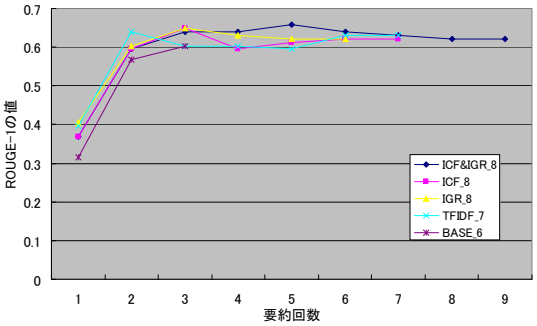


図 6 トピック 2 に対する評価

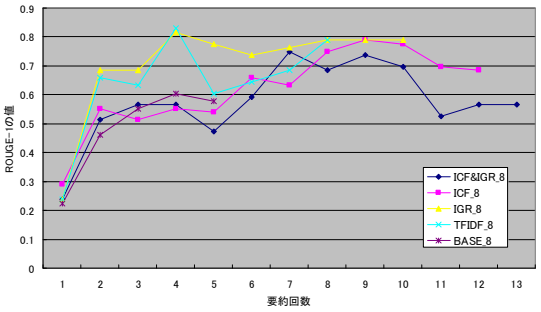


図 7 トピック 3 に対する評価

(2) 動向情報に関する情報抽出に関する検討

①概要

各種文書に現れる動向情報を集約して、その要約と可視化を行う場合には、文書から統計量に関する情報を抽出する必要がある。例えば、「大手自動車メーカーが２４日に発表した１０月の国内生産台数によると、トヨタ自動車は１４万台と前年実績を上回った。」という文においては、「１０月の国内生産台数」、「トヨタ自動車」という表現から推定される「トヨタ自動車の１０月の自動車の国内生産台数」という統計の調査方法と、それに対応する値の表現である「１４万台」の組が、統計量の抽出結果となる。そこで、まず、前者の文書中での表出を統計量名と定義し、そ

の自動抽出を検討した。特に、動向情報の集約を念頭に置き、統計量名を成す構成要素を分類された部品として抽出する。次に、文書中に散在している統計量名を成す要素を組み合わせて、一つの統計量名に同定することを検討した。

②文章中の表現と統計量との間の関係

次の二つの例文を考えよう。

例文１「Ａビールが発表した３月のビール出荷量は、２００万ケースだった。」

例文２「４月のＡのビール出荷数量は、２２０万ケース。」

統計量については、どのような統計であるかを表す表現(例えば、「４月のＡのビール出荷数量」)と対応する値を表す表現(例えば、「２２０万ケース」) の組で現れる。本研究では特に複雑な構造を持つ前者に注目をする。

まず、以下の概念を導入し、統計量の整理を行った。図８にその概念を図示する。

【統計の調査方法】ある統計量の値がどのように統計を取って得られたのかを示す概念. 文章中に現れるものではない. (「３月のＡビール社のビール出荷量」に対応する概念)

【統計量名】統計の調査方法を指し示すために文章中に表出する表現を分類し組み合わせたもの。例えば、後述の分類に従うと例文１の統計量名は{agent:Ａビール, time:３月,obj:ビール, foot:出荷, head:量}となる。

【統計量】ある「統計の調査方法」と，それに対応する値の組。

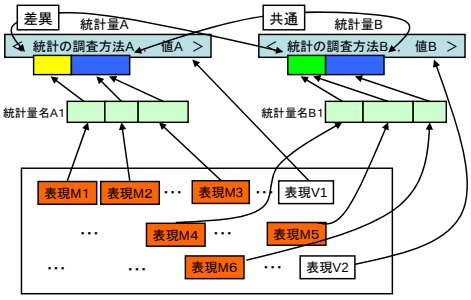


図８ 文章中に現れる統計量表現の構造

③統計量名の種類

統計量名は少なくとも３種に分類できる。

【動作型】ある動作によって生じた物の統計量を扱うもの。例えば、「１９９８年度のパソコンの国内出荷台数は７３５万台と前年度比１０％増で、前年実績を上回った。」

【属性型】ある物の状態や性質が統計量となっているもの。例えば、「１７日に中東のドバイ原油価格は１バレル当たり９・９８ドルであった。」

【定義型】外部で定義された何らかの式等に従って計算されるもの。例えば、「１月の景気動向指数は６２・５％となり、景気判断となる分かれ目である５０％を越えた。」

④統計量名の抽出タスクの構造

統計量名を構成する部品は文章中に単語の連続として出現するとは限らず、離れて出現する場合が多い。例えば、「国内のビール大手５社は１３日、１月の課税出荷数量を発表した。全体の数量は３０５万４０００ケースで、前年同月比１２５％と好調な滑り出し。」という文章では、「１月」、「課税出荷数量」、「全体の数量」が組み合わさって統計量名を構成している。本研究では、これら１つ１つの表現を統計量名の要素と呼ぶ。このとき次のタスクより抽出タスクが構成されると考えられる。

【統計量名の要素を抽出するタスク】文章中から統計量名の要素となるものすべてを取り出すタスク。

【統計量名の要素を組み合わせるタスク】取り出された要素を組み合わせて１つの統計量名を作るタスク。

⑤統計量名の各要素を注釈付けするためのタグセット

各種表現を分類し例文に注釈付けするために以下のタグを含むXMLタグセットを用意した。

【動作型に関するタグ】

obj ：対象となる部分。「ビール」など。

foot：対象が受けた動作の部分。「出荷」「生産」など。

head：統計量の数え方。「数」「量」など。

prop:統計量の数え方が割合で表されている部分。「シェア」など。

【属性型に関するタグ】

obj ：対象となる部分。「原油価格」における「原油」など。

attr： 対象の属性を表す部分。「原油価格」における「価格」など。

【定義型に関するタグ】

def：定義された式にしたがって計算された統計量。「景気動向指数」など。

【「条件」に関するタグ】(各型に共通)

time:統計量の値を集計した期間を表す部分。

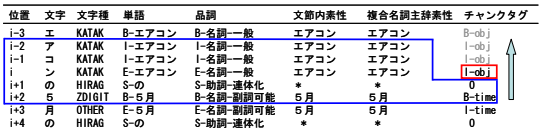
locat：統計量の値を集計した地域。

agent：会社名や機関名など。

⑥文字のチャンキングに基づく統計量名の要素の自動抽出

定義した各要素が、比較的標準的な抽出方法によってどれくらいの精度で抽出できるかを調べるために、文字を構成単位としたチャンキング問題として、統計量名の要素の抽出を捉えることを考える。チャンキングとは,任意の解析単位(トークン)をある視点からまとめ上げていき、まとめ上げた固まり(チャンク)をそれらが果たす機能ごとに分類することであり、固有表現抽出などで用いられる。そこで、統計量名の要素の抽出には中野らの固有表現抽出手法に対して、我々独自の複合名詞主辞素性を導入したものを用いる。図９

に同手法によるチャンクタグ推定の例を示す。また、各要素の自動抽出結果の評価を表 1 に示す。動作型の統計量名の主要素であり動作に対応する foot は適合率、再現率ともにほぼ 80%であり、数え方に対応する head に関しては適合率、再現率の両者が 85%以上であったため、動作型の主要素をある程度の精度で抽出できたと考えられる。



位置	文字	文字種	単語	品詞	文節内素性	複合名詞主辞素性	チャンクタグ
i-3	エ	KATAK	B-エアコン	B-名詞-一般	エアコン	エアコン	B-obj
i-2	ア	KATAK	I-エアコン	I-名詞-一般	エアコン	エアコン	I-obj
i-1	コ	KATAK	I-エアコン	I-名詞-一般	エアコン	エアコン	I-obj
i	ン	KATAK	E-エアコン	E-名詞-一般	エアコン	エアコン	I-obj
i+1	の	HIRAG	S-の	S-助詞-連体化	*	*	0
i+2	5	ZDIGIT	B-5月	B-名詞-副詞可能	5月	5月	B-time
i+3	月	OTHER	E-5月	E-名詞-副詞可能	5月	5月	I-time
i+4	の	HIRAG	S-の	S-助詞-連体化	*	*	0

図 9 素性集合に対する分類に基づくチャンクタグの推定

表 1 要素の自動抽出に関する適合率と再現率

	obj	foot	head	prop	attr	def
頻度	978	672	417	275	500	168
適合率	76.5	80.1	86.0	74.0	80.7	84.7
再現率	64.4	79.3	85.4	76.4	74.6	79.3

	time	locat	agent	age	add	range
頻度	2067	486	484	44	217	2362
適合率	73.3	73.0	74.9	83.3	72.5	76.2
再現率	69.8	59.0	68.8	83.3	72.9	67.1

⑦統計量名の要素の組同定

次の手続きに従って、統計量名の要素の組同定を試みる。ここで、一つの統計量名を構成する要素の組において、最後に出現する、基準点となる要素を「最終要素」と呼ぶ。

1) 文章を先頭から走査し、出現する各要素に対し、「最終要素」であるかを判定する。
2) 文章を再び先頭から走査し、以下の(3)から(5)までを新しい「最終要素」が出現しなくなるまで繰り返す。
3) 「最終要素」が出現した時、空のフレームを用意し、その「最終要素」を対応する種類のスロットに入れる。他のスロットに対応する要素の各種類に対して、「最終要素」以前に出現した要素の中のどの要素と結びつくか、または、どの要素とも結びつかないかの判定を行う。結びつく要素があれば、それを抽出し当該スロットに保存する。
4) (3)で保存された要素群を、1つの統計量名を構成する要素群として同定、出力をする。
5) フレーム上に保存されている要素を全て破棄し、走査を再開する。

上記手続きの中で二段階の判定を行っている。(1)での判定を第一段階、(3)での判定を第二段階とする。本研究では、いずれの段階も Support Vector Machine (SVM) を用いた機械学習手法により、その判断を行う分類器を教師情報から自動的に獲得することを試みた。

第一段階の学習過程においては、各要素が「最終要素」であるか否かを判定する分類学習を行い、分類器を生成する。

また第二段階では、判断された最終要素の各々について、フレーム構造を用意し、スロットに対応する各要素の種類毎に適切な要素を一つずつ決定する。要素のある一つの種類に注目すると、この過程は、i) テキスト中に現れる同じ種類の要素の中から、ある基準に従って候補群を集め、ii) その候補群の中から、現在の最終要素と適切に結びつく候補を一つ選択する、という手順からなる。複数の候補の中から適切な一つを選択する手法としては飯田らのトーナメントモデルを使用した。評価結果を表 2～4 に示す。

第一段階目の最終要素の決定は F 値にして、0.9 程度であり十分に高い精度で同定できることが分かった。また、二段階を続けた場合には、適合率、再現率がいずれも 0.7 程度であることが分かった。

表 2 第一段階の精度

トピック	適合率	再現率	F 値
AirConditioner	0.944	0.934	0.939
BeerIndustry	0.871	0.856	0.863

表 3 第二段階の精度

トピック	1 対 1	トーナメント
AirConditioner	0.885	0.789
BeerIndustry	0.819	0.803

表 4 二段階の過程を続けた場合の精度

トピック	適合率	再現率
AirConditioner	0.695	0.720
BeerIndustry	0.674	0.697

5．主な発表論文等

〔雑誌論文〕（計 1 件）

① 森辰則, 藤岡篤史, 村田一郎. 動向情報編纂のためのテキストからの統計量表現の自動抽出. 人工知能学会論文誌 Vol. 23, pp. 310-318 (2008) 査読有

〔学会発表〕（計 2 件）

①森辰則, 上野史紀. 動向情報編纂のためのテキストからの統計量表現の自動抽出: 統計量名の構成要素に関する組の同定. 第 22 回人工知能学会全国大会論文集, 3K3-6, 2008 年 6 月 13 日, 旭川

②森辰則, 藤岡篤史, 村田一郎. 動向情報編纂のためのテキストからの統計量の自動抽出. 第 21 回人工知能学会全国大会論文集, 3H9-4, 2007 年 6 月 22 日, 宮崎

6．研究組織

(1)研究代表者

森 辰則 (MORI TATSUNORI)
横浜国立大学・大学院環境情報研究院・教授
研究者番号: 70212264

(2)研究分担者

(3)連携研究者