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．昨年度は，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を行なった．

本年度は，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．特に，上記重要語抽出手法を MMR(Maximal Merginal Relevance) と呼ばれる冗長性制御機構と組み合わせることにより重要文抽出に基づく複数文書要約が行なえることを示した．また，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．そこで，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．昨年度は，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．

しかしながら，同要約手法は基本的には単文書要約であるので，利用者が把握できる文書数まで，利用者との対話を続けなければならず，途中の文書クラスタを概観する要約は生成できなかった．また，最終的に絞り込まれた文書集合においても関連事項の有無に関わらず個別に要約を提示することしかできなかった．そこで，本年度は，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．特に，上記重要語抽出手法を MMR(Maximal Merginal Relevance) と呼ばれる冗長性制御機構と組み合わせることにより重要文抽出に基づく複数文書要約を行なう．また，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．そこで，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・評価を行なう．

2　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

2.1　概要

本課題では，利用者が持つある特定のトピックにより検索が行なわれ，ある程度取捨選択が行なわれた後の複数文書を対象とし，原文の代わりとなる要約を提示することを目的とする．そのためには「**原文の主要内容の網羅性の高さ**」ならびに「**可読性の高さ**」という二つの尺度において高い評価を受ける要約であることが重要である．

先行研究としては，可読性を考慮しつつ，内容の網羅性に中心を据えて議論を行なっているものが多い [RJB00, SSW99, GMCK00]．それらの手法は類似した文書グループの発見において，クラスタリングを用いているが，その一方で，クラスタ間の関係や，各クラスタ内部のより詳細な部分クラスタ構造は利用していない．一方，我々は，対象文書集合のより詳細な類似性構造を階層的なクラスタリングにより得ることで，各クラスタでの共通事項のみならず，クラスタ間の差異も考慮できると考えた．そして，この情報を要約の手がかりとすることができれば，より肌理の細かい重要箇所の抽出ができるのではないかと考える．

本稿では，この点を主要な問題意識としつつ，内容の網羅性ならびに可読性の二点について，次のように近接する．森 [森 02] は，検索結果文書の間に存在する類似性構造を階層的クラスタリングによりクラスタ構造として抽出し，これを語の重みとする手法を提案している．この手法ではクラスタ間に存在する構造，すなわち，差異や共通点を同時に考慮して語の重みに反映できる．この手法は検索結果文書の各々を個別に要約する際に有効に機能することが示されている．しかし，複数文書から単一の要約文書を生成する場合には，文書間での内容の重複があり得る．そこで，我々は，元来，検索質問文とパッセージ群の関連度とパッセージ間の冗長性を同時に扱う要約手法として提案されている MMR[CG98] を，単文の重要度と冗長性を同時に考慮する手法として改訂した．

2.2　提案手法

図 1 に提案手法の概略を示す．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

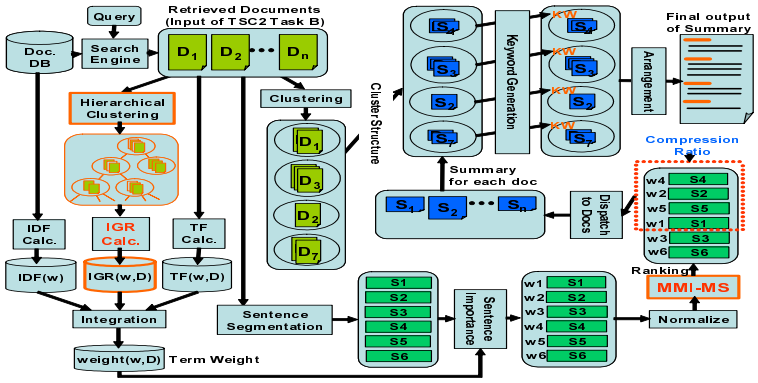


図 1: 情報利得比に基づく語の重みと MMR に基づく重要文抽出

約文中の冗長性の制御」の二点を統合した重要文抽出手法である．

2.3　評価実験

NTCIR3 TSC2 課題 B[TSC 02] に参加し，評価をおこなった．TSC2 タスクオーガナイザから与えられるトピック情報に従い，要約を生成し，その生成結果をタスクオーガナイザに提出する．同 Formal Run は，30 のトピックから構成され，各トピックは一つの情報検索結果に相当し以下の情報などから構成される．なお，対象となる文書は毎日新聞 98 年，99 年の記事である．各トピック，各要約文書長に対して，以下の 4 種類の要約が用意される．

人間による自由作成要約 (upperbound)，評価対象システムによる要約，lead 法による要約 (baseline1)，stein 法による要約 (baseline2)

次に，要約評価者に，トピック毎に原文書集合と各要約出力を読み比べてもらい，内容の網羅性 (C)，可読性 (R) の 2 つの観点から要約文書に順位をつけてもらう．

2.4 実験結果と考察

提案システムをトピックごとに baseline，upperbound のそれぞれとと比較した時の優劣を表 1 に示す．これによると，我々のシステムは対象とする文書集合の数が少ない時により評価が高い事がわかった．そこで 7 文書以下の文書集合から構成されるトピックに限定した評価も行なった．この結果を表 2 に示す．

表 1: baseline，upperbound との比較による優劣 (30 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	7	22	1	8	20	2	9	21	0	7	21	2
v.s. Lead 法	**15**	**12**	**3**	6	21	3	**17**	**13**	**0**	10	20	0
v.s. Stein 法	**16**	**12**	**2**	7	22	1	11	18	1	3	27	1

W:win L:lose T:tie

表 2: baseline，upperbound との比較による優劣 (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

我々のシステムは baseline 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,「要約率が小さい時」「対象とする文書集合の数が少ない時」という状況下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 の統合が重要文抽出に効果的である事を示す．

2.5 本部分課題のまとめと今後の課題

本稿では，文書分類をされた後の複数文書を対象とし，内容の網羅性と可読性を併せ持つ原文の代わりとなる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，対象とする文書集合の数が小さい時 (7 文書以下の時) に効果的である事が示された．今後の課題として，文書集合が大きい場合についての改善を検討したい．さらに現在のシステムに文間の結束性を保つ機構や言い換え手法などを採り入れる事による，可読性向上を考慮した要約生成も考えたい．

3 文章の構造解析による新聞記事からの事件情報抽出

3.1 概要

この課題では，一連の出来事において関連する人間の相互関係として意味構造を定義し，特に新聞の事件記事から意味構造（事件スキーマ）を抽出する手法について検討する．

必要な情報をすばやく効率よく手にいれるために，それらのドキュメントに対する自動要約や情報抽出などの自然言語処理への要求が高まってきている．このような現状において，パターン駆動による表層処理的な自然言語処理技術は，実装が容易なこととそれでいてある程度実用的な結果を得られることから，意味解析，文脈解析による「深い解析」が「王道」とは思われつつも，多くのシステムで採用されている．意味解析，文章解析とは，割りきってしまえば，文章を構成する文字の一次元的な配列を使用目的に応じて，定義された構造へ変換することである．抽出しようとする情報は，使用目的に応じてその「意味」の形式が変わりうる．

そこで，我々は，ある程度実用規模での文書理解，情報抽出を前提とし，文章要約や二次利用可能な情報蓄積を利用目的と想定し，意味構造を検討する．実際には,「犯罪」,「事件」ついて書かれた新聞記事 (事件記事) を対象とし,「犯罪」,「事件」の意味を表現する「犯罪スキーマ」を提案する．

本研究では，事件記事に対し，既存の文章内に内在する意味関係の解析を行い文章を構造化し汎用的内部表現を得る．そして，得られた汎用的内部表現から犯罪スキーマの抽出を行う．ある種の情報の抽出には，深い解析を行うよりも，むしろ，表層的な手がかり表現やパターンマッチングを用いることにより，容易に行うことができるものもある．一方，深い解析が必要な情報抽出もある．我々の解析システムでは，抽出する要素の性質に応じて，両者を使い分ける．また，解析により得られた構文木の部分木に対応する表層文についてパターン駆動で抽出を行うことにより，両者を組み合わせた解析が実現できる．

3.2 犯罪スキーマ

3.2.1 犯罪スキーマの定義

すべての事件記事は，罪状，動機，供述，人物の 4 つの要素で表現できると仮定する．そこで，我々は事件記事をこれらの要素をもつ犯罪スキーマとして定義する．以下は，犯罪スキーマ中のスロットである．

> **罪状スロット** (記事中で犯人が問われている罪状を値として持つ), **動機スロット** (犯人が犯行に至る理由を値として持つ), **供述スロット** (犯人の取り調べ中に述べている言動を値として持つ), **人物スロット** (記事中での役割を示すロール，経歴であるプロフィール，その人物のとった行動を示す行動という要素を持つサブスキーマで表現される)

3.3 文章の汎用的意味処理と犯罪スキーマ

汎用的意味処理は，入力テキストに対しまず構文解析を行い，その結果に対し，複文の関係解析を行う．各意味構造抽出部で意味構造を抽出し，それらの結果を統合した汎用的内部表現を出力する (図 2)．形態素解析には日本語形態素解析ツール JUMAN，構文解析には日本語構文解析ツール KNP を用いる．

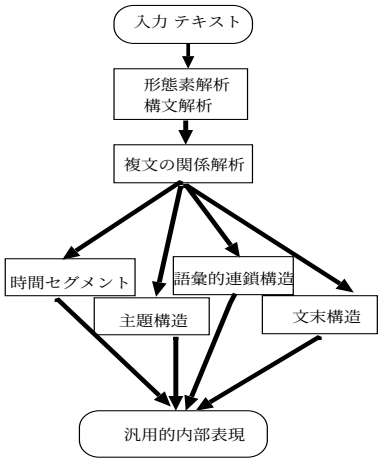


図 2: 意味構造の解析

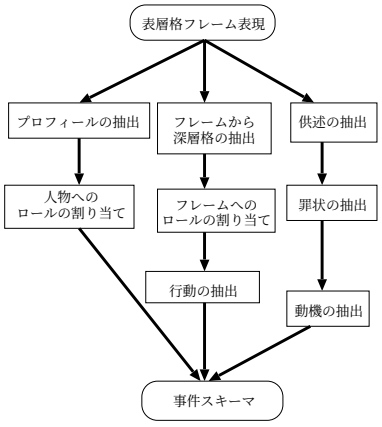
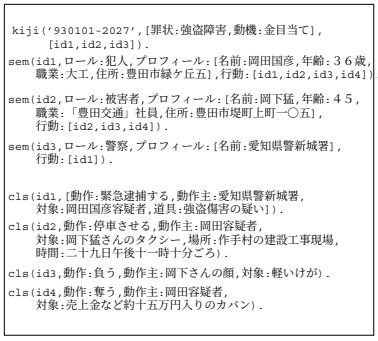


図 3: 犯罪スキーマの抽出アーキテクチャ



```
kiji('930101-2027',[罪状:強盗障害,動機:金目当て],
      [id1,id2,id3]).
sem(id1,ロール:犯人,プロフィール:[名前:岡田国彦,年齢:３６歳,
    職業:大工,住所:豊田市緑ケ丘五],行動:[id1,id2,id3,id4]).

sem(id2,ロール:被害者,プロフィール:[名前:岡下猛,年齢:４５,
    職業:「豊田交通」社員,住所:豊田市堤町上町一〇五],
    行動:[id2,id3,id4]).
sem(id3,ロール:警察,プロフィール:[名前:愛知県警新城署],
    行動:[id1]).


cls(id1,[動作:緊急逮捕する,動作主:愛知県警新城署,
    対象:岡田国彦容疑者,道具:強盗傷害の疑い]).
cls(id2,動作:停車させる,動作主:岡田容疑者,
    対象:岡下猛さんのタクシー,場所:作手村の建設工事現場,
    時間:二十九日午後十一時十分ごろ).
cls(id3,動作:負う,動作主:岡下さんの顔,対象:軽いけが).
cls(id4,動作:奪う,動作主:岡田容疑者,
    対象:売上金など約十五万円入りのカバン).
```

図 4: 犯罪スキーマの例

3.3.1 主題構造解析 (主題の抽出と主題の連鎖関係)

主題構造解析をするために，トピックと主題の抽出を行う．トピックと主題の定義を以下に示す．

- トピック：本研究では，新聞記事の見出しに出現する**名詞句**をすべてトピックと定義する．
- **主題**と**題述** [Hal01]：各文は，主題構造を持つと仮定し，各文は主題と題述とから構成されているとする．具体的には，**は格**，もしくは初出現の**が格**を主題と定義し，文の主題以外の残りの**名詞句** を，

題述と定義する．

記事中の各文間が，**A. 主題維持**，**B. 主題変化**，**C. 主題回復**，**D. トピックの導入**，**E. 主題派生**，**F. 主題の導入** のうちで少なくとも１つを満たすものとし，何らかの結束性を持っているとする．

原則として，結束関係の強さはA＞B＞C＞D＞Eとし，可能な限り結束性の高い連鎖を採用する．ただし，主題を抽出する際，主題が省略されている文に関しては，省略 (ellipsis) により結束構造 (cohesion)[Hal01] があるものとして，主題の維持と見なす．

3.3.2　犯罪スキーマの抽出アーキテクチャ

図 3 に示す手順で犯罪スキーマを抽出する．

3.3.3　犯罪スキーマの抽出例と各スロットの評価

現在実装されているスロットについて評価を行った．表 3 に評価結果を示す．この表からみても分かるように概ね 8 割程度の正解率が得られている．今回，目的とすることは犯罪スキーマを用いて，事件記事を構造化することにあるため，抽出精度は，現段階では十分であると考えている．犯罪スキーマの結果を図 4 に示す．

表 3: 30 記事に対する抽出結果

	ロール	名前	年齢	住所	職業	罪状	供述	動機
全出現数	51	64	64	44	35	30	7	30
システム	50	51	51	50	44	25	7	20
正解率	98.0%	79.7%	79.7%	88.0%	79.5%	83.3%	100.0%	66.7%

3.4　本部分課題のまとめと今後の展望

事件記事を対象とし，文章全体を構造化する人物の相互関係および時間的進行の観点から犯罪スキーマを提案した．実際に意味解析部は新聞記事 1601 記事対し動作を確認した．犯罪スキーマ抽出部もプロフィール，供述，動機，罪状については自動で抽出し評価を行った．

今後は，深層格フレーム抽出，フレームの人物の同定の自動化を行っていく予定である．

参考文献

[CG98] Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

[GMCK00] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-Document Summarization by Sentence Extraction. In *Proceedings of ANLP/NAACL Workshop on Automatic Summarization*, pp. 40–48, 2000.

[Hal01] M.A.K. Halliday. *An Introduction to Functional Grammar*. くろしお出版, 2 edition, 2001.

[RJB00] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of ANLP/NAACL Workshop on Automatic Summarization*, 2000.

[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 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.

[森 02] 森辰則. 検索結果表示向け文書要約における情報利得比に基づく語の重要度計算. 自然言語処理, Vol. 9, No. 4, pp. 3–32, 7 月 2002.

研究成果

研究発表

- 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. 9, No. 6, pp. 掲載予定, 2003.

- 森 辰則，國分 智晴，田中 崇: “空間分割型 CL-LSI による大規模言語横断情報検索”, 情報処理学会論文誌:データベース, Vol. 43, No. SIG 2(TOD 13), pp.27–36, 2002.
- 森 辰則: “検索結果表示向け文書要約における情報利得比に基づく語の重要度計算”, 自然言語処理, Vol. 9, No. 4, pp.3–32, 2002.
- Tatsunori Mori: “Information Gain Ratio as Term Weight — The case of Summarization of IR Results ”, —Proceedings of the 19th International Conference on Computational Linguistics (COLING 02), pp.688–694, 2002.
- Hiroshi Nakagawa and Tatsunori Mori: “Simple but Powerful Automatic Term Extraction Method”, Proceedings of the second International Workshop on Computational Terminology (COMPTERM 02), pp.29–35, 2002.
- 浅野秀胤，田村直良: “文章セグメントの単一化による多文章自動要約”, 言語処理学会第 8 回年次大会, pp.551–554, 2002.
- 大村高史，田村直良: “主題構造解析による新聞記事からの気象情報の抽出と応用”, 言語処理学会第 8 回年次大会, pp.623–626, 2002.
- 吉田和史，塩田好伸，森辰則: “情報利得比に基づく重要語抽出による情報ナビゲーション”, 言語処理学会第 8 回年次大会,pp.475–478, 2002.
- 辻克俊，権瓶竹男，森 辰則: “共起性を考慮した素性集合分割による Co-Training”, 言語処理学会第 8 回年次大会, pp.5–8, 2002.
- 森 辰則，太田 知宏，藤畑 勝之，公文 隆太郎: “質問応答システムにおける最良優先探索制御”, 情報処理学デジタル・ドキュメント研究会報告 2002-DD-33, 2002.
- Tatsunori Mori, Tomohiro Ohta, Katsuyuki Fujihata and Ryutaro Kumon: “A* Search Algorithm for Question Answering”, Proceedings of NTCIR Workshop 3 Meeting — Part IV: question Answering Challenge (QAC1), pp.39–46, 2002.
- 金山 淳一，北條 孝，田村 直良: “文章の構造解析による新聞記事からの事件情報抽出”, 情報処理学会自然言語処理研究会 2002-NL-152,pp.1–6, 2002.
- 佐々木 拓郎，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”, 情報処理学会自然言語処理研究会報告 2002-NL-152,pp.63–70, 2002.
- Tatsunori Mori and Takuro Sasaki: “Information Gain Ratio meets Maximal Marginal Relevance — A method of Summarization for Multiple Documents ”, —Proceedings of NTCIR Workshop 3 Meeting — Part V: Text Summarization Challenge 2 (TSC2), pp.25–32, 2002.

公開ソフトウェア

- 中川 裕志，森 辰則: 専門用語抽出システム

  URL: http://www.forest.eis.ynu.ac.jp/TermExtraction/

  与えられた特定分野のコーパスのみから，その中に現れる専門用語を特定し，抽出することができる.