自然言語処理を利用した
放送メディアアクセス技術に関する研究

後藤　淳

博士（情報学）

総合研究大学院大学
複合科学研究科
情報学専攻

平成25年度
（2014）

2014年3月

本論文は総合研究大学院大学複合科学研究科情報学専攻に
博士（情報学）授与の要件として提出した博士論文である.

審査委員：

山田 誠二 教授	(主査)	総合研究大学院大学／国立情報学研究所
相澤 彰子 教授		東京大学／国立情報学研究所
佐藤 真一 教授		東京大学／国立情報学研究所
相原 健郎 准教授		総合研究大学院大学／国立情報学研究所
宮尾 祐介 准教授		総合研究大学院大学／国立情報学研究所

A study on media access related to broadcasting based on natural language processing

Jun Goto

DOCTOR OF
PHILOSOPHY

Department of Informatics
School of Multidisciplinary Sciences
The Graduate University for Advanced Studies (SOKENDAI)

March, 2014

A dissertation submitted to the Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Advisory Committee

Seiji Yamada (Chair)	The Graduate University for Advanced Studies / National Institute of Informatics
Akiko Aizawa	The University of Tokyo / National Institute of Informatics
Shiin'ichi Satoh	The University of Tokyo / National Institute of Informatics
Kenro Aihara	The Graduate University for Advanced Studies / National Institute of Informatics
Yusuke Miyao	The Graduate University for Advanced Studies / National Institute of Informatics

論文要旨

近年，放送と通信の連携が急速に進んでいる．これまで放送とは放送局から一方的に送られてきた映像や音声を受信し視聴するものであった．ブロードバンド環境の普及により，この状況が変わりつつある．2013年３月には，放送と通信の連携を推進する新しい放送方式であるハイブリッドキャストの技術仕様が，放送局やメーカーなどで組織されたIPTVフォーラムより公開され，今後，この新方式の受信機が普及していくことが予想されている．これにより，放送番組に付与された情報と，通信で得られるコンテンツを利用した様々なアプリケーションをテレビ受信機上で利用できるようになる．

放送と通信の連携が進むなか，放送メディアを取り巻く環境の大きな変化の１つに，番組アーカイブと呼ばれる放送済みの番組を蓄積した巨大なデータベースの登場が挙げられる．ユーザは番組アーカイブにアクセスすることで，大量の過去の番組をいつでも視聴できるようになる．もう１つの大きな変化は，マイクロブログなどの通信コンテンツが放送の一面を見せ始めたことである．これまで，放送が伝えてきた情報，例えば，イベント，災害，事故等のより詳細な情報が，マイクロブログを通して一般の人々からリアルタイムに直接多くの人々に届くようになってきている．

本研究の目的は，視聴中の放送番組を基点に，番組アーカイブやマイクロブログなどの通信上のコンテンツから番組に関連する情報を容易に取得する効果的なメディアアクセス手段を確立することである．放送と通信の連携を進展していく上で重要となる番組アーカイブやマイクロブログのコンテンツを十分活用するには，大量のデータから必要な情報を効果的に引き出すことができる手段が不可欠である．情報を使いこなせる人とそうでない人との間に起こるデジタルデバイドが問題となるなかで，放送と通信の連携時代のメディアアクセスにおいても，誰でも容易に利用できる環境の構築を支援することは急務である．

本論文では，まず第１章で，本研究の背景としてこれまでの放送メディアの進展について整理し，近年出現した番組アーカイブやマイクロブログなどの通信を用いた新しいメディアの現状についてまとめる．これらのサービスに効率的にアクセスできる手段を提供することの社会的なニーズについて述べる．さらに，番組アーカイブとマイクロブログへのアクセスに関する現状の技術についてまとめ，取り組むべき課題を確認する．

第２章では番組アーカイブ内に格納されている大量の番組の中からユーザが見たい番組を取得するメディアアクセス技術を検討する．家庭用レコーダなどで現在利用されている番組検索には，番組に付与されているジャンルや出演者などのメタデータが利用されている．しかし，番組アーカイブのように検索対象の番組数が大規模となると，既存の番組検索では，検索結果を適量に絞り込めず，ユーザが意図する番組を取得することは難しい．本章では，まず番組検索に関する先行研究について紹介し，本研究の対象を明確にする．次に，放送番組に付与されている番組概要の類似度に基づく関連番組検索手法を提案し，提案手法を実装した実験システムを用いてその性能を評価する．さらに提案手法の拡張として，Wikipediaの更新履歴など外部の情報を利用した注目度による番組検索結果のリランキングについて，その有効性を考察する．最後に，番組オンデマンドサービスへの適用事例や，閲覧中のウェブページなどの外部コンテンツから関連性のある番組を提示する提案手法の応用について述べる．

第３章では，内容を考慮した番組検索の実現のために，番組概要からエンティティ間の関係を取得することを検討する．特に番組において重要なエンティティである人物に着目する．まず人物間の関係取得に関する先行研究について紹介し，本研究の対象を明確にする．次に，番組概要から人物間の関係を取得し，複数の関係を接続した関係グラフを生成する手法を提案する．続いて，関係グラフ生成で用いられている人物表現抽出処理，人物表現間の共参照解析処理，人物表現間の関係抽出処理の性能を評価するとともに，関係グラフを用いることで，構文解析結果から直接取得できない人物間の関係を獲得できることを検証し，エンティティ間の関係を利用した番組検索への応用について考察する．また，この関係グラフを用いて，登場人物相関図を生成する番組内容の可視化への応用を検討する．

第４章では，新しいメディアとしての存在感を増している Twitter などのマイクロブログから，テレビなどの既存メディアが伝えきれない情報を効果的に取得するための質問応答を用いたメディアアクセス技術について述べる．世界中の様々な人々から時々刻々投稿されるマイクロブログの膨大な数の文書から必要な情報を網羅的に取得し，その全体像を把握することは容易ではない．本章では，まず質問応答技術についての先行研究を紹介するとともに，マイクロブログの大量の書き込みから情報取得を行う際の課題を明確にする．次に，マイクロブログ全体の情報を網羅的に取得するための質問応答に基づく大規模データからの情報取得手法を提案する．続いて，提案手法の部分処理として用いられている言い換えパターン作成処理，回答インデックス作成処理，地名補完処理，質問応答処理の各処理についてそれぞれ説明する．最後に，東日本大震災の際に実際に発信された Twitter の大規模データを用いて提案手法の性能を評価し，課題と今後の展望を述べる．

第５章では，本論文の成果をまとめる．

本論文の成果は，放送と通信の連携に伴い登場した新しいテレビ視聴環境における課題を確認し，その課題の解決手段として番組アーカイブやマイクロブログなどの通信上のコンテンツから，放送の情報を基点に必要な番組や情報を取得するメディアアクセス技術を提案するとともに，その有効性を検証し実用化への道筋を示したことにある．本研究により，放送と通信の連携・融合時代の放送メディアへのアクセス方法の確立に向けて大きな貢献を行えたと考えている．

目次

図目次

表目次

第１章 序論

1.1 研究の背景

近年，放送と通信の連携が急速に進んでいる．これまで放送は放送局から一方的に送られてきた映像や音声を受信し視聴するものであった．ブロードバンド環境の普及により，この状況が変わりつつある．2013年３月には，放送と通信の連携・融合を推進する新しい放送方式であるハイブリッドキャストの技術仕様がIPTVフォーラム[1]より公開され，今後，新方式の受信機が普及することが予想されている [藤沢13]．これにより，放送で送信される番組関連情報と，通信で得られるコンテンツとを連携した様々なアプリケーションをテレビ受信機上で利用できるようになる．例えば，インターネット上の地図表示サービスを利用して紀行番組の場所を地図上に表示したり，Twitterなどのマイクログログへの書き込みを番組と連動して表示することなどが可能となる．

放送と通信の連携が進むなかで，これまでの放送を取り巻く環境が大きく変わったことがある．その１つとして，番組アーカイブと呼ばれる放送済みの番組を蓄積した巨大なデータベースの登場が挙げられる．インターネットなどの通信を利用して番組アーカイブにアクセスすることで，これまでは放送で送られてきていた番組を，ユーザの好きな時間に視聴できるようになる．例えば，2003年に設立されたNHKアーカイブスでは，過去に放送された数十万以上の番組を保存しており，その一部の番組はオンデマンドサービスを通して視聴することができる．多くの人手と時間を掛けて制作された良質の番組を蓄積した番組アーカイブは，巨大な知識ベースとも言え，ユーザが必要な情報を得る上でもその利用価値は大きいと言える．

もう１つの大きな変化は，Twitterなどのマイクロブログが放送メディアの側面を持ち始めたことである．マイクロブログの登場により，ユーザ発信のコンテンツが，既存メディアが伝えきれない情報を補完し始めている．これまでの既存メディアのコンテンツは，放送局やプロダクションなどが制作し，それらを電波やケーブル等で放送すること

[1] http://www.iptvforum.jp/about-iptv/member.html

で，ようやく視聴者に届くものであったが，マイクロブログにおける個人や企業が作成したコンテンツが，ダイレクトに多くの人々に届くようになってきている．ビジネス分野やエンターテイメント分野でのマイクロブログの活用例がメディアで度々取り上げられ，その存在が一般に認知され始めたことから，Twitterなどのマイクロブログを利用した個人や企業発のコンテンツの影響力はますます大きくなってきている．既存のメディアでも，その発信力は無視できず，マイクロブログのコンテンツを積極的に取り入れた番組も増えてきている．またマイクロブログ側でもテレビに連動して番組に関連する情報を発信するユーザが増加するなど，両者のコンテンツレベルでの連携も進んでいる．

このような放送と通信の連携の進展により，ユーザはこれまで受動的に既存のメディアから受け取っていた情報をはるかに凌ぐ大量の情報を，通信コンテンツから取得できるようになる．しかしながら，これまでの既存メディアの視聴スタイルに慣れたユーザにとっては，大規模な番組アーカイブから見たい番組を探したり，大量に発信されるマイクロブログから自分が知りたい情報や番組に関連する情報を引き出したりすることは，容易な作業ではない．そのため，ユーザが意図する情報を大量のデータから探し出すことをサポートする仕組みが必要である．そこで，本論文では，放送と通信の連携により登場した新しいメディアである番組アーカイブやマイクロブログの情報に効果的にアクセスできる手段を提案し，その有効性を検証する．

1.2 放送と通信の連携

本節では，本研究が目指す放送と通信の連携サービスを明確にする．まず，1.2.1で日本における放送と通信の連携の歴史を述べ，1.2.2および1.2.3で放送と通信の連携において重要な通信上のコンテンツである番組アーカイブとマイクロブログの概要についてそれぞれ述べる．1.2.4で本研究が目指す放送と通信の連携のサービス像について述べる．

1.2.1 放送と通信の連携の進展

日本の放送メディアにおける放送と通信の連携の歴史を振り返ってみる．日本における放送と通信の連携が注目され始めたのは，BSデジタル放送が開始された2003年に遡る．放送のデジタル化に伴い，BML（Broadcasting Markup Language）で記述された情報を送信するデータ放送が開始された．データ放送を利用することで，番組に関連する情報や放送局からのお知らせを文字情報として表示したり，番組コンテンツと連動した簡易なアプリケーションの実行が行えるようになった．データ放送の開始は，視聴者が放送中のクイズ番組に参加して，その回答結果を電話回線を利用し放送局に送るといった，放送と通信を連携した新しい放送コンテンツを制作するきっかけとなった．しかし，電

話回線を用いた放送局とユーザ間での通信による情報のやり取りは，送受信できる情報量の制約から限定的なものとなっていた．

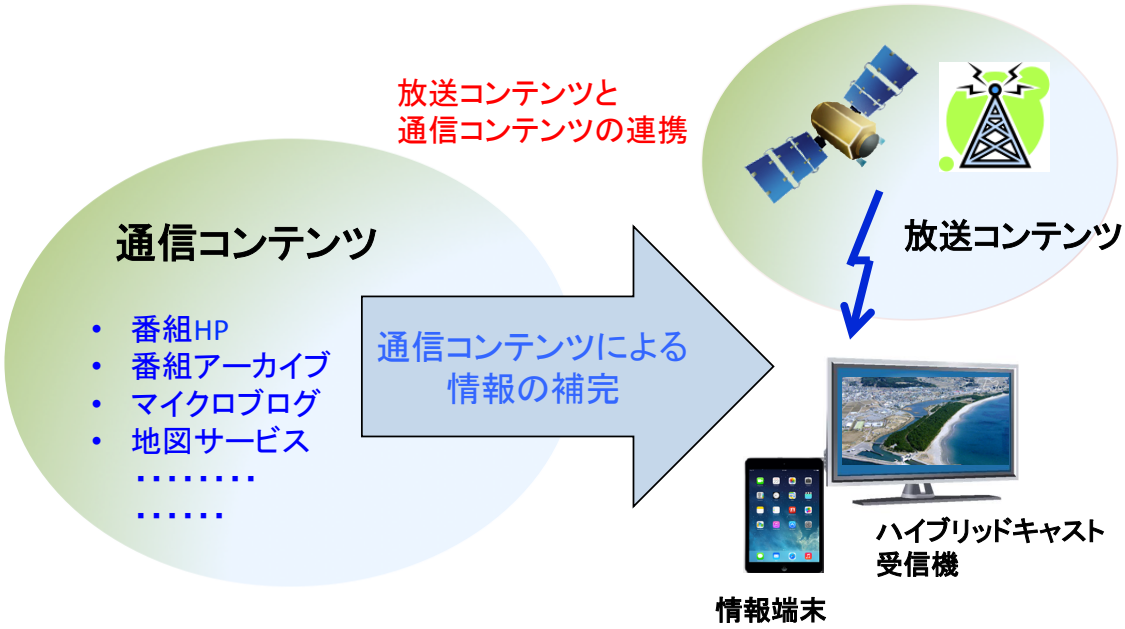


図　1　放送と通信の連携

各家庭へのブロードバンドの普及が進むにつれて，インターネットへの接続が可能となる受信機も増え始めた．この環境の変化により，データ放送に番組のホームページのURLなどの通信上のコンテンツの保存情報を含めておくことで，通信上のコンテンツにアクセスし，それを提示することができるようになった．さらに，2006年のワンセグ放送の開始に伴い，携帯電話などの情報端末でも放送番組を視聴できるようになり，テレビ番組とともにインターネット上の情報をよりシームレスに閲覧する環境が整ってきた．また，2008年には，番組オンデマンドサービスをNHKが開始し，ネットに接続された専用端末，高機能テレビ，パソコンなどでオンデマンドサービスを楽しむことができるようになった．

そして，次世代の放送と通信の連携として，2013年にHybridcastの技術仕様が公開され，受信機やPCを含む情報端末上で放送と通信のコンテンツをより連携させた様々なアプリケーションを利用できる環境が整うことになる（**図 1**）．今後，このハイブリッドキャストが普及していくためには，番組アーカイブやマイクロブログをはじめとする通信コンテンツを効果的に利用する手段が必要である．

1.2.2 番組アーカイブ

番組アーカイブとは, 放送済みの大量の番組を再利用可能な形式で保存しているデータベースである．歴史的に日本では放送番組の保存は積極的に行われず，番組のアーカイブの歴史は長くはない．江原 [江原07] の報告によると，番組の保存が行われてこなかった理由は，古くから放送番組の保存を進めてきたヨーロッパなどに比べて，日本では放送コンテンツの文化的価値を重要視していなかったことや, VTRテープなどの保存メディアのコストが非常に高価であったことが挙げられる．そのため，一度収録されたテープに別の番組を上書きして再利用するという運用が行われていた．その結果，当時に制作された番組のほとんどが，放送局にVTRでは残されていないこととなった．ただし，フィルムで制作されていた番組は上書きできないため，その多くは残されており，VTR時代の番組でも視聴者が個人的に録画していたテープが提供されたというケースもある．このようにして集められた番組をデジタル化し保存して出来上がったものが番組アーカイブである．近年，ハードディスクなどの蓄積メディアのコストが低下したため，デジタルテープからハードディスクへの移行も進んでいる．

現在，多くの放送事業者が，アーカイブした番組をインターネットやケーブルテレビを通して視聴者に配信する番組オンデマンドサービスを開始している．例えば，NHKでは，NHKアーカイブスに70万本以上の番組を保存しており，その一部のコンテンツ（約4000本）を番組オンデマンドサービスで配信している．番組オンデマンドサービスで見たい番組を探すには，一般に，番組へのアクセス頻度によるランキング，ジャンルによる分類, タイトルや番組内容の全文検索などの手段を用いることができる．しかし，これらの現状のアクセス手段だけでは, アーカイブされた大量の番組の中から見たい番組を探し出すことが容易ではないという課題がある．

1.2.3 マイクロブログ

マイクロブログは, インターネットを利用して情報を投稿できるサービスの総称であり，一般に一度に発信できる情報量が少量に制限されている．そのため，ブログなどに比べて手軽に投稿ができ，頻繁な投稿がなされている．マイクロブログを代表するサービスとして，2006年に開始されたTwitterがある．Twitterでは，ユーザからの投稿はtweetと呼ばれ，一度に投稿できるコンテンツの量が140字以内に制限されている．また注目するユーザを登録してそのユーザのtweetを自動で取得するためのfollowと呼ばれる手段や，共感を得たtweetの複製をそのまま発信するretweetという手段も提供されている．このfollowやretweetの機能により，一人の発信したtweetが様々なユーザを介して，最終的に，何万，何十万人にも届くという既存の放送メディアに匹敵する発進力を持つ場合

がある．一方，受信側では，世界中の発信者から届く大量のtweetから自分が知りたい情報を選び出すことは労力のいる作業である．Twitterでは，情報をフィルタする一つの手段として，tweetに付与されたHashtagと呼ばれる情報（「#」から始まる文字列で表される）を利用できる． Hashtagを指定することにより興味のあるtweetだけを取得することが可能となる．ただし，Hashtagが付与されているtweetは全体から見るとごく一部であり，膨大なtweetから必要な情報を取得するためにはHashtag以外の手段が必要である．

マイクロブログのサービスと，ネットに接続可能な携帯電話や通信端末の普及によって，ユーザは時間や場所を選ばず，目の前で起こった出来事の情報をその場で発信できるようになった．マイクロブログを閲覧する側からすると，情報にたどり着く手段さえ整えば，世界中で起こっている様々な出来事についての情報をリアルタイムに取得できるようになったと言える．一般に，テレビなどの既存のメディアは，電波という限られたリソースを用いているため，それぞれのメディアの判断で情報に優先順位をつけている．そのため，少数の人のみに役立つ情報よりも，より多くの人に役立つ情報が優先されて放送される傾向がある．マイクロブログのコンテンツはメディア自体がその情報を取捨選択している訳ではなく，受信側でその情報の選択ができるため，既存のメディアのコンテンツからでは得ることができない貴重な情報を取得できる可能性がある．マイクロブログに存在する貴重な情報を活用するためには，大量の書き込みから必要な情報を取得する効果的な手段を提供することが課題となる．

1.2.4 目標とする放送と通信の連携サービス

本研究の目標は，放送と通信の連携サービスの普及のために，放送のコンテンツを基点として，関連する通信上のコンテンツを容易に取得することができるメディアアクセス環境を構築することである．ハイブリッドキャストの仕組みを利用することにより，受信機上で通信上のコンテンツを用いた様々なアプリケーションを動作させることができる．本研究では，通信から取得する対象のコンテンツとして，1.2.2および1.2.3で説明した番組アーカイブとマイクロブログを用いる．

放送を基点に関連する通信コンテンツの情報を簡単に取得できるようになれば，**図 2**に示すような放送と通信を連携したサービスが可能となる．例えば，ユーザが話題となっている「伊勢神宮の式年遷宮」を取り上げた紀行番組を視聴していたとき，番組アーカイブにアクセスして遷宮に関する過去の番組を再生したり，マイクロブログから番組では紹介されていない伊勢神宮に関する詳細な情報や，地元の人しか知らない近くの美味しいレストランなどのローカル情報を得ることが可能となる．また，災害や事故が起きた場合には，番組オンデマンドにアクセスして見逃した重要なニュースや特別番組を視聴したり，個人から発信されるtweetなどのマイクロブログへの書き込みを，放送で

は伝えきれない詳細な情報や局所的な情報を取得するための手段として活用することができる．

図 2 放送と通信の連携のサービスイメージ

1.3 関連技術と本研究の位置づけ

本節では，番組検索とマイクロブログからの情報取得についての関連技術，および，それらの課題についてまとめ，本研究の位置づけを明確にする．

1.3.1 番組検索技術

番組の検索では，放送局が付与したメタデータを利用した検索手法が多く提案されている．現在の放送で用いられているメタデータは，ARIB（社団法人電波産業会）の規格 [ARIB13] に準拠しており，放送形態，番組識別ID，放送日時，放送チャンネル，番組タイトル，出演者，番組ジャンル，番組概要などの情報が含まれている（**表 1**）．例えば，メタデータの一つである番組ジャンルは，ニュース，ドラマ，映画などのジャンル大分類とより詳細な分類のジャンル中分類が番組毎に設定されている．**表 2**にジャンル大分類の一覧を示す．番組ジャンルや出演者などのメタデータを利用した番組検索は，小規模な番組を対象としては既に実用化されており，ハードディスクレコーダや一部の高機能テレビなどのコンシューマー機器に取り入れられて番組の選択や自動録画などに活用されている [村上07].

表 1 メタデータの種類

メタデータの種類	内容
放送形態	地上波/BS/CS などの ID
番組識別ID	放送日における番組固有の ID
放送日時	放送日，開始時刻，放送時間
放送チャンネル	放送局の固有の ID
番組タイトル	番組のタイトル，新番組，再放送などのラベルも付与
出演者	出演者，声優，監督などを含む.
番組ジャンル	ジャンル大分類，中分類情報が最大３種類
番組概要	番組の概要を説明したテキスト
番組内容	番組の内容を説明したテキスト．番組概要と同じテキストの場合も多い

表 2 ジャンル（大分類）の一覧

ID	ジャンルの分類	ID	ジャンルの分類
0x0	ニュース／報道	0x7	アニメ／特撮
0x1	スポーツ	0x8	ドキュメンタリー／教養
0x2	情報／ワイドショー	0x9	劇場／公演
0x3	ドラマ	0xA	趣味／教育
0x4	音楽	0xB	福祉
0x5	バラエティ	0xC - 0xD	予備
0x6	映画	0xE	拡張
		0xF	その他

これまでの番組検索に関する多くの研究において，上記で説明したメタデータが利用されている．住吉らは，メタデータと人手で作成したテンプレートを利用して番組を推薦するシステムを提案している [住吉03] . Gotoらは，エージェントとの対話に基づき，EPG情報の出演者，番組ジャンル，放送チャンネルなどのメタデータを利用して，録画している番組や放送中の番組から，ユーザが見たい番組を検索するエージェントシステムを提案している [Goto04]．溝口らは，メタデータと独自のオントロジに基づく類似度を利用した関連番組推薦システムを提案している [Mizoguchi07].

また，メタデータの間接的な利用として，ユーザが見ていた過去の番組の履歴のメタデータを取得してユーザプロファイルを作成し，これから放送するユーザの好みの番組を提示する手法も提案されている．隆らは過去に視聴した番組の視聴履歴に含まれるジャンル情報や検索に利用したキーワードを利用して番組を検索・提案するシステムを提案している [隆01]．AliらはTiVoと呼ばれる協調フィルタリングに基づく番組推薦システムを提案している [Ali04]．協調フィルタリングを利用したシステムでは，十分な履歴がない場合，例えば，開始されて間もないサービス初期や，放送されたばかりのコン

テンツについては, 既存の履歴を利用した推薦は行えないというコールドスタート問題がある．そのため，TiVoでは出演者やジャンルなどのメタデータを利用したコンテンツの類似度による推薦を併用している．

これまでの番組検索技術で用いられていたジャンルや出演者などのメタデータによる絞り込みは, 基本的に録画された番組やケーブルテレビの放送番組などの比較的少量の番組のデータベースを対象に検索するためのものであった．しかし，検索対象の番組数が，番組アーカイブのように膨大になると，これまでの方法では，ユーザが期待する検索結果が得られなくなる．対象の番組の数が膨大になると，ユーザが指定したメタデータにマッチする番組の数や種類が多くなり, なかにはユーザの意図しない番組まで検索結果として出力されてしまうためである．一方，見たい番組が明確に決まっている場合は，番組名やシリーズ名等のメタデータによる検索で番組を探すことはできるが，ユーザが探している番組以外にも, 別の観点でもっと興味がある番組が巨大な番組アーカイブには眠っているかもしれない．番組アーカイブへのアクセス手段としては，ユーザが有益と思える番組に接触できる機会を提供することも重要である．

本研究では，上記の問題点に鑑みて，番組アーカイブの規模でも適量の結果を出力することができ, その上でユーザの要求と関連性のある様々な番組を取得することができる番組検索手法を検討する．

1.3.2 マイクロブログへのアクセス技術

マイクロブログから必要な情報を取得するとき, 数億規模の書き込みから欲しい情報を如何に引き出すことができるかが問題となる．マイクロブログは，リアルタイム性と網羅性というメリットを持っている反面，その全体の情報量が膨大であるため，必要な情報に辿り着くことが難しいと言える．例えば，東日本大震災では，「既存メディアから得られない現地の局所的な情報を被災者や支援者に提供する手段として, マイクロブログが存在感を示した」とマスコミ等でその有用性が頻繁に取り上げられたが，橋元の報告によると，大量の情報を整理し情報を探し出す手段が十分でなかったために，事例によっては必ずしも必要な情報を得ることができなかったとの問題点も指摘されている [橋元13] ．

このような問題を解決するために, マイクロブログから必要な情報を取得する手法について, 多くの研究開発がなされている. アメリカ国立標準技術研究所（National Institute of Standards and Technology, NIST）が開催している情報検索をテーマとした国際ワークショップTRECでは，2011年からマイクロブログから必要な情報が書かれている投稿を

取得するマイクロブログトラック[2]が行われている．このトラックでは，2011年の１月から２月にかけての２週間分の1600万tweetを対象とし，リアルタイムに情報をフィルタリングするタスクが行われている. 多くのシステムでは, 単語の頻度に基づく類似度, 時間情報による選別, シソーラスやRelevance feedbackによるキーワード拡張などを利用して, 必要なtweetを取得するシステムを開発している [Gurini12] [Miyanishi12] [Gao12].

日本語を対象としたtweetからの情報取得の研究開発の事例としては，Yahoo Japanが，Twitterのリアルタイムの投稿をキーワードでフィルタできるサービス「Yahooリアルタイム」[3]を2011年より開始している．このサービスでは，Tweetに含まれるキーワードを入力することで，５秒おきにそのキーワードが含まれるtweetのみを表示することができる．現在のところ，外部アプリケーションがAPIなどを通してこのサービスの結果を利用することはできない．また，青島らは，tweetのタイムスタンプと単語の共起情報を利用し，同じ話題に触れているtweet特定し，それらをマージして表示することで，大量の情報をわかりやすく表示するシステムを提案している [青島13]．北口らはクエリとしてtweetを選択することで，類似するtweetを取得するシステムを開発している [北口13].

上記に述べたマイクロブログ上の情報へのアクセスに関する研究開発の事例では, キーワードや時間情報により刻々変化するマイクロブログをフィルタリングし, 情報の取得や閲覧を行い易くすることが試みられている．しかし，時間やキーワードだけでの制約だけでは，数億規模の情報を俯瞰的に把握することは難しい．たとえ，適合するtweetが得られたとしても，その取得したtweetが大量にあれば，それらをすべて読むことは事実上不可能であり, ユーザが本当に知りたい情報に辿り着くことができないかもしれない．また，既存のWeb検索エンジンのように，マイクロブログを全文検索してそのランキング上位のみを取得し閲覧することは可能であるが，その場合，ロングテール部分に存在する貴重な情報を持つかもしれない書き込みは無視されてしまう.

本研究では，ユーザの様々な欲求をシステムに自然言語で的確に伝えられ，その質問に対する回答だけを対象の文書から抜き出し提示できる質問応答技術を利用して, マイクロブログの膨大な書き込みから必要とする情報のみを自動で取得し, それらの情報を俯瞰的に提示するアクセス手段を検討する.

[2] http://trec.nist.gov/data/microblog.html

[3] http://search.yahoo.co.jp/realtime

1.4 本論文の構成

本節では，次章以降の本論文の構成について述べる．第２章では番組アーカイブ内に格納されている大量の番組からユーザが見たい番組を取得するメディアアクセス手段を検討する．まず番組検索に関する先行研究について紹介し，本研究の対象を明確にする．次に，メタデータに含まれる番組内容を示した概要の類似度に基づく関連番組検索手法を提案し，その性能を評価する．さらに提案手法の拡張として，Wikipediaの更新履歴など外部の情報を利用した注目度による番組検索結果のリランキングについて，その有効性について考察する．最後に，番組オンデマンドサービスへの適用事例や，閲覧中のウェブページなどの外部コンテンツから直接関連性のある番組を提示する応用例について述べる．

第３章では，内容を考慮した番組検索の実現のために，番組概要からエンティティ間の関係を取得することを検討する．特に番組において重要なエンティティである人物に着目する．まず人物間の関係取得に関する先行研究について紹介し，研究の対象を明確にする．次に，番組概要から人物表現間の関係を取得し，複数の関係を接続した関係グラフを生成する手法を提案する．続いて，関係グラフ生成で用いられている人物表現抽出，人物表現間の共参照解析，人物表現間の関係抽出の各処理の性能を評価するとともに，関係グラフを用いることで，構文解析結果から直接取得できない人物間の関係を獲得できることを検証し，取得した関係を利用した番組検索への応用について考察する．また，この関係グラフを用いて，登場人物相関図を生成する番組内容の可視化の応用を検討する．

第４章では，新しいメディアとしての存在感を増しているマイクロブログから，テレビなどの既存メディアが伝えきれない情報を取得するための質問応答技術に基づくアクセス手段について検討する．まず，質問応答技術についての先行研究を紹介するとともに，マイクロブログからの情報取得に適用する際の課題を明確にする．次に，マイクロブログの情報を網羅的に把握するための質問応答に基づく情報取得手法を提案する．続いて，提案手法の部分処理として用いられている言い換えパターン作成処理，回答インデックス作成処理，地名補完処理，質問応答処理の各処理についてそれぞれ説明する．最後に，東日本大震災の際に実際に発信されたTwitterの大規模データを用いて提案手法の性能を評価し，課題と今後の展望を述べる．

第５章では，本論文の成果についてまとめる．

第２章 言語情報を利用した関連番組検索

2.1 はじめに

デジタル放送の開始以降，放送された番組をデータベース（番組アーカイブ）に蓄積し，インターネット等を通じて番組を提供するサービスが普及し始めている．データを蓄積するためのストレージの大規模化により，今後も番組アーカイブに保存される番組の量は飛躍的に大きくなることが予想できる．ユーザは，放送されている番組に加え，通信を利用して番組アーカイブから自分が興味のあることや調べたいことを取り上げた番組を視聴したいときに利用できるようになる．

地上波放送，BS/CS衛星放送，ケーブルテレビなどを通して多くのチャンネルで番組を視聴できる環境が整い，自動で家庭用ハードディスクレコーダに録画した番組を視聴することも当たり前になってきた近年，見たい番組を選ぶための番組検索技術は既に多くの家庭でも利用されている．現在，一般に販売されている多くのテレビ受信機や家庭用レコーダでは，番組に付与されたジャンルや出演者，放送時間などの番組に付与されているメタデータを指定することで番組を検索することができる．家庭用レコーダの番組検索のように検索対象が数十から数百程度の規模の場合で，出演者や番組ジャンル（ドラマ，報道など）の検索のための条件が明確に決まっている場合には，これらのメタデータを用いた番組検索は番組にたどり着くための有効な手法のひとつである．しかし，放送と通信の連携により，検索の対象が放送中やレコーダの番組から，番組アーカイブに蓄積されている番組が対象となり，その数が数千から数万以上の規模になった場合，ジャンル情報や出演者などのメタデータによる絞り込みだけでは，十分に番組を絞り込むことができず，意図している番組にたどり着くことは難しい．

番組アーカイブの検索の手段として，番組に付与されたメタデータのテキストを対象として，フリーキーワードによる全文検索を利用することが考えられる．しかし，現状の番組アーカイブの規模が大きくなっているとは言え，Webを対象とした全文検索とは違い，キーワードを入力しても何も結果が得られない場合も多い．このような状況では，ユーザが検索のたびにヒットしそうなキーワードを試行錯誤して考えなくてはならず，

情報検索などに慣れていない人にとっては極めてハードルが高い. 情報機器に精通している人とそうでない人でサービスの質に大きな差がでることは, 情報バリアフリーの観点からも適切ではない. 特にテレビのように誰もが利用する機器は特に注意が必要である.

そこで, 本研究では, 誰にでも簡単に番組アーカイブから見たい番組を取得することができる環境の実現を目指し, ユーザが選択した番組の情報を基に関連がある番組を自動で検索するメディアアクセス手段を提案する. 本章では, まず次節で番組検索の関連研究について述べる. 次に, 2.3節で情報検索などでのスコアリングに利用されるBM25をベースに複合語や固有表現を考慮した関連番組検索手法を提案する. また2.4節で提案手法を実装した実験システムの概要について説明し, 2.5節で実験システムを用いた提案手法の評価実験の結果について述べる. さらに2.6節で提案手法の拡張として番組アーカイブ以外の外部情報を利用して番組検索結果のリランキングの可能性について考察する. 2. 7 節で提案手法を用いた実用化例を紹介し, 2.8節でまとめる.

2.2 関連研究

番組を検索する際に利用する番組の情報として, 放送番組に付与される字幕や番組概要を用いる研究が多く進められている. 本節では番組検索の関連研究として, これらの番組に付与されている情報を利用した研究について述べる.

番組を探す際の手がかりとして, まず番組に付与された字幕を利用することが考えられる. 日本のデジタル放送では, 番組に付与される字幕は大きく２つに分類することができる. 一つはオープンキャプションと呼ばれ, 出演者名や地名などの文字列が映像の一部として番組に付与された字幕であり, 文字情報として取得するためにはOCR技術などで抽出する必要がある. もう一つの字幕のクローズドキャプションは, 聴覚障害者の方々のためなどに, ナレーションや出演者などの番組の音声を文字化した字幕であり, 通常は画面上に表示されておらず, 受信機のモードを変更することで表示することができる. 国の方針もあり, 音声認識技術等を利用してクローズドキャプションをすべての番組につける取り組みがなされている [Ando03].

番組検索に関する多くの研究で, 文字情報として取得が容易なクローズドキャプションが用いられている. 例えば, 番組のシーンを検索するために必要なメタデータをクローズドキャプションから自動・半自動で作成する研究が行われている. 山田らは, クローズドキャプションを解析し, 教師あり学習によりサッカー番組のシーンにゴールなどのイベントタグを付与している [山田06]. また, 場所紹介などの定型的な文章区間を特定する手法を提案している [山田07]. 柴田らは, 番組本編の映像と字幕の言語情報を利用し, 料理番組のトピックを推定する手法を提案している [柴田07]. またMiuraら

は，自然番組において，被写体とその動作を示す字幕区間を特定する手法を提案している [Miura08]．このようにクローズキャプションを解析してシーンを示す詳細メタデータが付与されると，それを利用したシーン検索やダイジェスト番組の自動作成などの応用 [Miyazaki08] が期待できる．

しかしながら，上記で提案されている各手法により詳細なメタデータが付与することができる番組の種類は限られており，網羅性の点で問題がある．現状，すべての番組のシーンにこれらのメタデータを付与するには，非常に多くの異なる手法を併用することが必要であり，コスト面などから実用化は難しい．そのため，より汎用的で網羅性のあるメタデータの付与が必要とされている．また，現状の多くの番組アーカイブでは，クローズドキャプション情報が格納されておらず，クローズドキャプション情報を利用した番組検索の実現には時間を要する．

序論でも紹介したが，デジタル放送開始以降に放送される番組には，番組を特徴づけるメタデータが付与されている．そのメタデータには，番組の内容を自然言語でコンパクトに記述している番組概要が含まれる．山口らは，Webの検索キーワードと閲覧したページからtf-idfの高い単語を抽出し，番組概要から抽出したキーワードとのシンプソン係数により関連する番組を取得する手法を提案している [山口10]．澤井らは，Twitterでフォロー関係のあるユーザや代表的なユーザのtweetを利用してプロファイルを作成し，プロファイルと検索対象の番組の単語ベクトルの類似度計算により番組を推薦する手法を提案している [澤井10]．ベクトルの重みとして，番組概要に含まれる単語のidf値とジャンルのidf値を利用している．これらの手法では，番組概要から取得した単語ベクトルを番組アーカイブ以外の外部サービスのAPI情報により重み付けを行い，類似度を計算している．しかし，これらの手法では，外部サービスの結果に依存しているため，APIの利用回数の制限などから大規模な検索には適用することができなかったり，サービスが出力する仕様が変更されるたびに，システムの挙動が大きく変わるため，実用的な利用は難しい．

本研究では，現状，番組アーカイブにクローズドキャプションのデータが付与されていない状況を考慮し，番組の内容を自然言語で記述した番組概要を利用した番組検索手法を検討する．番組概要は，放送されるほとんどの番組に付与され，番組アーカイブにも再利用されているため，番組アーカイブを対象とした番組検索を早期に実現することが期待できる．番組概要を利用した上記の先行研究では，検索エンジンやTwitterなどの外部サービスから得られる情報による重みに焦点が当てられており，番組概要そのものの解析にはシンプルな手法が用いられている．本研究では，放送局がコストをかけて人手で作成し，番組の内容をコンパクトにまとめている質の良い要約とも言える番組概要の特徴を分析し，番組概要の関連性に基づく関連番組検索を提案する．また，外部サー

ビスの解析結果を必要としないWikipediaの更新履歴を用いた関連番組検索のリランキング方法についても検討する．

2.3 番組概要に基づく関連番組検索

これまでにも述べたとおり，番組を検索するとき，検索対象の番組数が多くなってくると，出演者やジャンルによるメタデータでの検索では適量の結果を得ることが難しい．その一因として，最近の多くのテレビ番組では，スポーツ選手や俳優が情報番組に出たり，歌手やお笑いタレントがドラマに出演するなど，多くの出演者が様々なジャンルの番組に出演しており，出演者と番組の内容の相関が小さくなってきていることがあげられる．ジャンルによる検索も一つのジャンルが取り扱う範囲が大きいため，検索対象の番組数が増加するにつれて非常に多くの番組が検索結果に出力されることになる．このようなことから，番組アーカイブを対象とした番組検索を考える場合，出演者や番組ジャンルのメタデータによる検索とは異なる別の手段で番組を検索することを検討する必要がある．

近年，フィルタバブルと呼ばれる問題が取り上げられている [Pariser11]．これは個人化技術によって，ユーザが接することができる情報が制限されて，本来，接触できるはずの情報を見ることができなくなるという，Pariserが提起した問題である．番組検索においても，個人ごとの視聴履歴を基に協調フィルタリング技術などを利用し，システムがユーザが興味がある可能性が高い番組のみを推薦することは可能である．しかし，数多くの優れた番組が登録されている番組アーカイブにおいて，特定の種類の番組だけに提示を限定することは，ユーザの貴重な番組視聴の機会を奪うことに成りかねない．そのため，ユーザの要求に関連する番組を提示することができる番組検索の技術が必要である．検索により意外性のある番組が提示されれば，これまで興味がなかった番組を見る機会をユーザに提供することができる．

本研究では，上記の問題を踏まえ，番組概要の関連性に基づき番組を検索する手法を検討する．番組概要とは，放送番組に付与される，番組内容を自然言語で記載した情報である．ドラマやドキュメンタリーの番組では，そのあらすじを記述した概要が付与される傾向がある．本研究では，以後，NHKの番組に付与されている番組概要を用いる．**図** 3にNHKの自然番組「ダーウィンがきた！」の番組概要の例を示す．また，番組概要は，電子番組表などで視聴者が番組を選ぶ際の有力な情報となるため，放送事業者も積極的に番組に付与している．また，多くの場合，番組オンデマンドにおける番組の紹介文として再利用されることもあり，その質も徐々に向上してきている．NHKオンデマンドに登録されている2218番組の番組概要を調べたところ，その文字数は平均183（最大200）であった．

アフリカ中部のカメルーンに世界最速のサルがいる．足が長くスタイル抜群のパタスモンキーだ．草原を時速50キロで走り抜ける．全身の筋肉をバネのように使った走りは，チーターそっくり．しかもこのサル，短距離走だけでなく省エネ走法でマラソンも得意．さらにひと跳びで3メートルの高さまで届くほどのジャンプ力．パタスモンキーはなぜ走り続けるのか？　サルのトップアスリートの運動能力の秘密やユニークな子育てに迫る．

図 3 番組概要

本研究で提案する関連番組検索の動作概要について説明する．図 4にその動作イメージを示す．例えば，歴史ドラマを視聴していて，その背景や登場人物の詳細を知りたいとき，ユーザは，簡単な操作（検索ボタンの押下等）により登場人物や同時代を取り上げた特集番組など視聴中の番組と関連している番組を取得することができる．明示的に検索などの複雑な操作を行わなくても，見ている番組の番組概要から番組を特徴づける言語情報をシステムが自動的に取得し，番組アーカイブに登録されている関連する番組の検索を行ってくれる．システムは，関連する番組を検索するため，視聴中の番組と番組アーカイブの番組概要の言語情報を用いてそれらの関連度を計算し，その上位の番組を取得している．また関連番組の検索結果を提示する際に，なぜその番組が選ばれたのかを示すための情報が提示される．

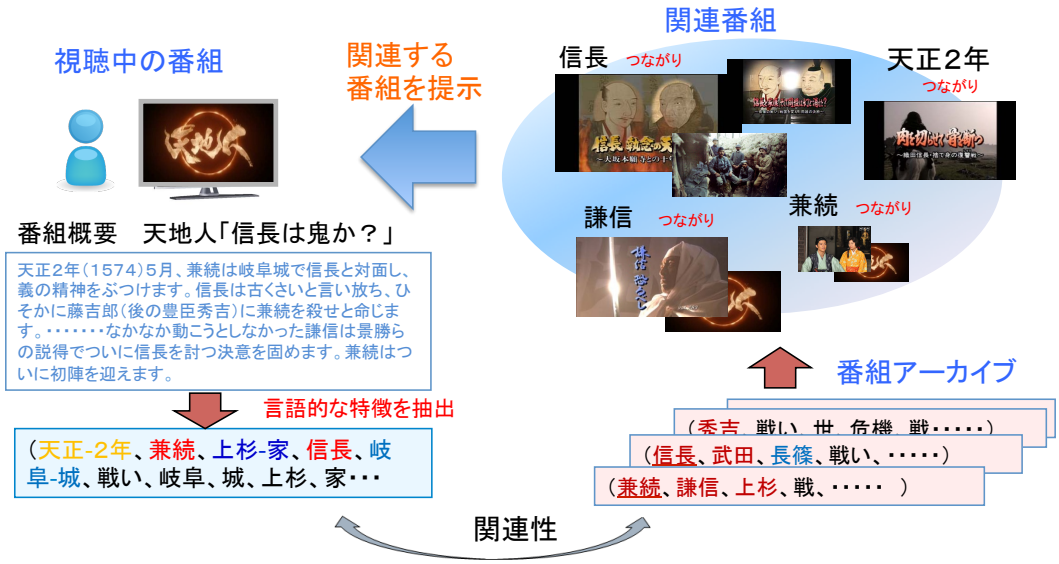


図 4 関連番組検索の概要

以下，2.3.1でBM25を利用した番組間の関連性導出について，2.3.2で複合語を考慮するためのn-gramへの拡張について，2.3.3で固有表現の重み付けについて説明する．さらに，2.3.4では，得られた番組検索結果がなぜ選ばれたかを示す関連ラベルの生成について述べる．

2.3.1 BM25を利用した番組間の関連度

本研究での番組間の関連度の導出には，情報検索や質問応答などのタスクで高い性能を示しているBM25を利用する[Robertson99] [Aramaki06]．BM25は，DARPA主催の評価型ワークショップTRECの情報検索タスクにおいてエントリーしたシステムOkapiに用いられた手法であったため，Okapi BM25としても知られている．BM 25は，質問文などのクエリにおける語の頻度term frequency（tf）と，データベースに格納されている文書におけるtf とinverse document frequency（idf）を用いて，クエリに対する各単語のスコアを取得することができる．tfとはある文書に含まれる単語の出現頻度であり，idfは，ある単語が含まれるデータベース中の文書数を示すdocument frequency（df）の逆数に対数をとったものであり，ある語のデータベースにおける希少性を表す指標である．

番組アーカイブに含まれる全番組の番組概要に現れる語 t_i の総数をIとすると，検索のクエリとなる選択された番組の番組概要Qから取得する特徴ベクトルQ，番組アーカイブに保存されている番組Dの特徴ベクトルDは，それぞれ式(1) (2)で表わされる．Q, Dともに，番組アーカイブの全単語数の次元をもったベクトルとなる．ただし，１つの番組概要に現れる単語がアーカイブに含まれている全単語数に比べて非常に小さいため，ほとんどの次元が０となる疎なベクトルとなる．

$$Q = {}^{t}\left(q(t_1), q(t_2), \ldots, q(t_i), \ldots q(t_I)\right) \tag{1}$$

$$D = {}^{t}\left(d(t_1), d(t_2), \ldots, d(t_i), \ldots d(t_I)\right) \tag{2}$$

クエリ番組の番組概要Qから取得する特徴ベクトルQの要素$q(t_i)$と，番組アーカイブに含まれるある番組の番組概要Dから得られた特徴ベクトルDの要素$d(t_i)$は，BM25に基づき，式(3)，(4) により定義する．

$$q(t_i) = \frac{(k_3 + 1)tf_q(t_i)}{k_3 + tf_q(t_i)} \tag{3}$$

式(3)において，$tf_q(t_i)$ はユーザが選択したクエリ番組における語t_iの頻度を示す．

$$d(t_i) = \frac{(k_1+1)tf_d(t_i)}{k_1((1-b)+b\cdot {}^{\varepsilon}/_{\eta})+tf_d(t_i)} log\left(\frac{M-m(t_i)+0.5}{m(t_i)+0.5}\right) \qquad (4)$$

式(4)において，$tf_d(t_i)$ は番組アーカイブ中のある番組における語t_iの頻度を示し，M, $m(t_i)$ はアーカイブ全体の番組数，t_iを含む番組数をそれぞれ示す．また，εは番組アーカイブに格納されているある番組概要の長さを示し，ηは番組アーカイブに含まれる番組概要の平均の長さを示す．$k1$，$k3$ はクエリと番組アーカイブの間で重視する比を決め，b は番組概要の長さの正規化に関する調整用パラメータである．

$$Score(Q,D) = \sum_{i=1}^{I} q(t_i)\cdot d(t_i) \qquad (5)$$

視聴番組Qと番組アーカイブに含まれるある番組Dの関連度$Score(Q,D)$は，式(5)により示すとおり，式(3)と(4)をすべての単語t_iにおいて求め，それらの積を総計することより取得する．

2.3.2 複合語を考慮したn-gramへの拡張

複合語は，単語が複数連鎖して一つの意味を持つものである．例えば，「日本-銀行」と「日本-の-銀行」は異なる意味の表現であるが，この２つの表現は共通の単語の「日本」と「銀行」を含んでいる．そのため，「日本銀行」を含む番組をクエリとして単語ベースで検索した場合，「日本-銀行」という複合語を含む番組と「日本-の-銀行」を含む番組のスコアに大きな違いがなくなる．直感的には，同一の単語の連鎖を持つ複合語が共起する場合の関連度が，その一部の単語が個別に離れて共起している場合よりも高くなるほうが望ましい．

そこで，2.3.1で説明したBM25に基づく関連番組検索手法において，解析の単位となるt_iを単語からn-gramに拡張し，複合語が共通している番組間のスコアを上昇させることを検討する．単語の連鎖数が1からnまでのすべてのn-gramを要素とし，頻度に基づく重みを付与したベクトルを生成する．検索のキーとなる選択された番組Qに含まれている単語の集合から得られた特徴ベクトルをQ_{ng}と，番組アーカイブに保存されている番組Dの特徴ベクトルD_{ng}は，式 (6) (7) で表わされる．それぞれのベクトルの次元は全n-gramの総数Nとなる．

$$Q_{ng} = {}^{t}\big(q(t_1), q(t_2), \dots, q(t_i), \dots q(t_N)\big) \tag{6}$$

$$D_{ng} = {}^{t}\big(d(t_1), d(t_2), \dots, d(t_i), \dots d(t_N)\big) \tag{7}$$

n-gramに拡張したベクトル間の演算を行うことで，n-gramに拡張していない場合に比べて，複合語が含まれる番組のスコアを相対的に上げることができる．しかしながら，あるn-gramが共起すれば，その (n-1)-gramから 1 -gramも共起し重複して計算されるため，連鎖数nが大きくなると，複合語を含む番組のスコアが過剰に上昇してしまう．そのため，各n-gramの値に対して調整用の重み$w_{ng}(t_i)$を導入し，過剰なスコアの上昇を緩和する（式(8)）．なお，$g(t_i)$はn-gramの連鎖数を返す関数である．調整要重み$w_{ng}(t_i)$を考慮したn-gramに拡張した関連度$Score_{ng}$は最終的に式(9)となる．

$$w_{ng}(t_i) = 1/g(t_i) \tag{8}$$

$$Score_{ng}(Q_{ng}, D_{ng}) = \sum_{i=1}^{N} w_{ng}(t_i) \cdot q(t_i) \cdot d(t_i) \tag{9}$$

対象とする複合語には, 品詞によるフィルタリングを行い, 助詞, 助動詞を含むn-gramは除外する．形容詞や動詞などはその対象としているため，「美しい国日本」「振り込め詐欺」などの名詞の連続でない複合語についても扱うことができる．

2.3.3 拡張固有表現に基づく重み

番組概要に含まれる表現の意味を反映した番組の検索を実現するため, これまでに提案した単語やn-gramの頻度に基づく手法に，番組概要から抽出した固有表現（Named Entity）による重みを導入することについて検討する．

固有表現とは, テキスト内の内容を理解する上で重要である人名や組織名といった言語表現である．自然言語処理分野において，固有表現の概念が最初に登場したのは Defense Advanced Research Project Agency（DARPA）主催のMessage Understanding Conference（MUC） [Grishman96]においてであると言われている．MUCでは，新聞記事などから国防に関する情報を構造化して抽出する評価型プロジェクトが実施されており，その際，固有の名称（人名，組織名，地名）と，特定の数値表現（時間，日時，金額表現，割合表現）の抽出が独立のタスクとして取り上げられていた．日本では，評価型ワークショップのIREX [Sekine00] で，MUCの 7 種類の固有表現に「人工物（Artifact)」を加えた 8 種類の固有表現の抽出が行われた．日本語を対象とした固有表

現の抽出手法としては，人手で作成した規則により固有表現を得る手法 [内元00]や，固有表現抽出を形態素へのラベル付与の問題と捉え，Support Vector Machine（SVM）などの機械学習を用いた手法が提案されている [山田02]．その後，構文解析器などのツールのオプションに固有表現抽出処理が実装されて，これまでに様々な研究でIREX準拠の固有表現タグが利用されている．

番組検索に用いる固有表現を考えた場合，IREX の固有表現は，数値表現を除くと，人名，地名，組織名，人工物の 4 種類であり，その適用範囲が十分とは言えない．例えば，**図 5**（固有表現に相当する部分に下線）では，アフリカ，カメルーンなどは IREX の地名タグを用いることができるが，動物名のチーター，パタスモンキーや職業や称号を示すアスリートなどの表現を区別できないため，固有表現の種類を増やす必要がある．

アフリカ中部のカメルーンに世界最速のサルがいる．足が長くスタイル抜群のパタスモンキーだ．草原を時速50キロで走り抜ける．全身の筋肉をバネのように使った走りは，チーターそっくり．しかもこのサル，短距離走だけでなく省エネ走法でマラソンも得意．さらにひと跳びで3メートルの高さまで届くほどのジャンプ力．パタスモンキーはなぜ走り続けるのか？サルのトップアスリートの運動能力の秘密やユニークな子育てに迫る．

図 5 拡張固有表現を付与した番組概要

上記のような問題点を踏まえ，Sekine は，これまでの固有表現を拡張した関根の拡張固有表現（Sekine’s Extended Named Entities）[Sekine08] を提案している．拡張固有表現は，MUC，ACE，IREX で定めた固有表現の定義に加え，百科事典などの項目を参考にした 4 階層から成る 200 種類の固有表現タグセットである．また，橋本らは拡張固有表現タグを新聞記事と白書コーパスに付与した拡張固有表現コーパス [橋本 08] を作成している．これら 2 種類のコーパスを比較した結果，拡張固有表現の分布はコーパスの種類によりかなり異なることが報告されている．

本研究において番組検索で拡張固有表現を利用するには，番組概要に現れる拡張固有表現の状況を把握することが重要である．そこで，対象とする拡張固有表現を決定するために，番組概要における拡張固有表現の出現頻度を調べた．まず，2218 番組の番組概要に対し，数値表現を除く拡張固有表現の第 3 階層タグ 50 種類を付与した番組概要コーパスを作成した．その中で少なくとも 2 つ以上の番組に現れる 2289 の拡張固有表現タグを取得し，その頻度が高い順に上位 10 位（全体の 86%）を取得した（**表 3**）

表 3 番組概要における拡張固有表現の出現頻度

順位	拡張固有表現タグ名	割合（固有表現の頻度/全固有表現数）	
1	人名（Person）	0.354	(635/2289)
2	政治的地名(GPE)	0.150	(269/2289)
3	称号名(Titile)	0.132	(236/2289)
4	生物名(Living_Thing)	0.056	(100/2289)
5	地形名(Geological_Region)	0.036	(64/2289)
6	芸術作品名(Art)	0.034	(61/2289)
7	生物部位名(Living_Thing_Part)	0.030	(53/2289)
8	食べ物名(Food)	0.025	(44/2289)
9	地理的組織名(GOE)	0.022	(40/2289)
10	地域名(Region)	0.014	(33/2289)

得られた上位10位までの拡張固有表現のうち，タイトルが同じである同一のシリーズの番組内にしか出現しなかった拡張固有表現タグの食べ物名，芸術作品名，生物部位名については除外し，政治的地名，地形名，地域名については地名としてまとめた．このようにして最終的に得られた人名(PER)，地名(LOC)，地理的組織名(GOE)，称号名(TIT)，生物名(LIV)の拡張固有表現タグを関連番組検索における関連度導出に利用する．

そこで，上記で定めた５種類の拡張固有表現を，番組検索での関連度（式(9)）を改善するための重みとして用いる．まず，機械学習を用いて番組概要から対象の拡張固有表現を抽出する．次に，n-gramの t_i が抽出された拡張固有表現に一致するかを判定し，一致すれば拡張固有表現による重みを付与する（式(10)）．αは拡張固有表現の種類により定められる定数である．

$$w_{ne}(t_i) = \begin{cases} \alpha & (t_i\text{が拡張固有表現の場合}) \\ 1.0 & (\text{それ以外の場合}) \end{cases} \tag{10}$$

前節で定義したn-gramに拡張した関連度（式(9)）に，拡張固有表現の重み$w_{ne}(t_i)$を付与した最終的な番組間の関連度は以下のようになる．

$$Score(Q, D) = \sum_{i=1}^{N} w_{ne}(t_i) \cdot w_{ng}(t_i) \cdot q(t_i) \cdot d(t_i) \tag{11}$$

2.3.4 関連ラベルの提示

関連番組検索により結果が複数出力された場合，最終的にユーザは自分の意思で視聴する番組を選ぶ必要がある．その際，番組検索の結果数が非常に多くなると，番組選択のタスクそのものがユーザに負担をかけることになる．そのため，システムがなぜその番組を薦めているのかを示す情報を提示し，複数の番組検索結果から見たい番組を選び易くすることが必要である．

本研究では，選ばれた番組がクエリ番組から見て，どのような観点でシステムにより出力されたかを示すラベルを提示することとする．以後，このラベルを関連ラベルと呼ぶ．関連ラベルは，クエリ番組から見た検索結果の各番組の関係性を示す情報であるため，クエリ番組が変われば，検索結果にある同じ番組でもその関連ラベルも変化する．関連ラベルはクエリ番組の番組概要に含まれているn-gramのいずれかから選ばれ，そのn-gramの選択には，前節で説明したクエリ番組と検索結果の番組の関連度の計算過程で得られる各n-gramのスコアが用いられる．関連ラベルとして取得されるn-gramは，以下の計算により求められる．

$$label(Q,D) = arg\max_{t_i}(w_{ne}(t_i)\cdot q(t_i)\cdot d(t_i)) \qquad (12)$$

計算式からもわかるように，関連ラベルは，番組アーカイブ内での頻度が少ない連鎖数nが大きい複合語や，重みを付与される固有表現が選ばれる傾向がある．式(12)においてトップのn-gramだけでなくスコア上位のn-gramを複数取得すれば，一つの番組に対して複数の関連ラベルを付与することも可能である．

図 6は，NHKの自然番組「ダーウィンがきた！」をクエリ番組として選んだときの関連番組検索の結果の一部を示したものである．例の検索結果には４つの関連番組が出力されており，結果①の番組には，国名の「カメルーン」という関連ラベルが提示されている．また結果②と③の番組は，「サル」「チーター」という動物名の観点で関連であることがわかる．結果④では，固有表現ではない一般名詞の複合語が関連ラベルとして取得されており，両番組が「運動能力」という観点で関連性があることを示している．同じ関連ラベルを持つ検索結果が複数ある場合には，それらを同じ関連ラベルのグループにまとめて，検索結果をクラスタリングすることもできる．図の例では番組間の関連ラベルは１つだけであるが，先に述べたとおり，より多くの検索結果がある場合などについては，複数のラベルを出力することができる．

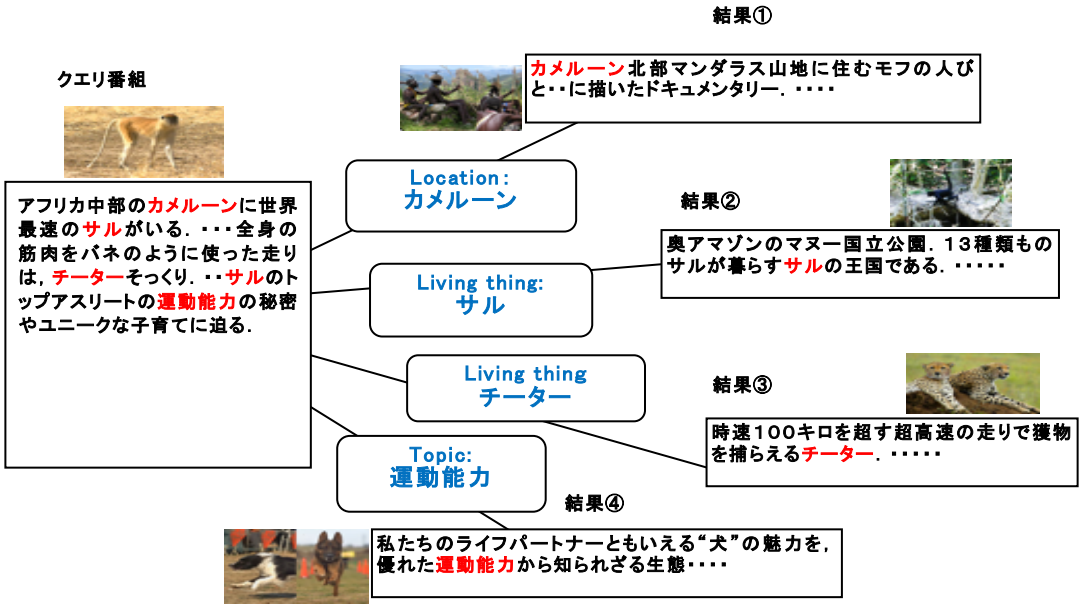


図 6 関連番組間の関連ラベル

2.4 実験システム

提案手法の有効性を検証するため，提案手法を実装した関連番組検索の実験システムを構築した．実験システムでは，番組アーカイブから取得したすべての番組概要の間の関連度および関連するパラメータを格納したインデックスをシステム起動時に作成することで，高速に番組検索を行うことができる．検索により得られた関連番組は，関連ラベルとともにユーザインタフェースに提示される．以下に，実験システムの処理概要と実験システムを操作するための関連番組検索インタフェースについて述べる．

2.4.1 実験システムの概要

実験システムでは，高速に番組を検索するため，あらかじめ各番組の番組概要に含まれているn-gram毎のスコアを計算したインデックスを作成している．

実験システムは，最初の起動の際に，高速に関連番組を取得するためのインデックスを作成する．そのために，番組アーカイブに登録しているすべての番組の番組概要に対して，形態素解析により各番組概要に含まれる形態素のn-gram（n=1-3）を生成し，それに対応するBM25に基づく関連度を計算するための各パラメータを求める．番組アーカイブ全体に関するパラメータとして，番組数M，あるn-gram t_iを含む番組数$m(t_i)$，番組概要の平均の文字列の長さη，番組毎に異なるパラメータとして，各番組概要のn-gram t_iの頻度，番組概要の文字列の長さεを求める．

次に，n-gramの連鎖数に対する重み$w_{ng}(t_i)$と，拡張固有表現のn-gramに重み$w_{ne}(t_i)$を導出する．拡張固有表現の重み$w_{ne}(t_i)$には，開発用の15番組をクエリとした予備実験でMean Average Precision（MAP）が最大となる重みのPER: 1.0, GOE:1.1, LOC:1.1, TIT:1.2, LIV:1.3（重みの範囲1.0-1.5の場合）を用いる．MAPについては次節で改めて説明する．重みを付与する対象の拡張固有表現の特定には, 6400番組の番組概要の拡張固有表現コーパスをConditional Random Fields（CRF）[Lafferty01] で学習した認識器を用いる．学習の素性には，各形態素及び前後の形態素の表層，形態素の文字n-gram，読み，品詞，前後の品詞，EDR概念辞書[4]の概念IDを用いた．この認識器の抽出精度は，200番組（学習データには含まれない）の番組概要に適用した場合，F値で0.909であり，拡張固有表現毎の内訳は，PER: 0.956，GOE:0.783，LOC:0.894，TIT:0.772，LIV:0.586である．

上記で説明したパラメータを予め番組アーカイブ内のすべての番組に対して取得しインデックスに格納しておく．これにより，選択した番組がアーカイブ内の番組であった場合には，高速に番組間の関連度を取得できる．ただし，放送中の番組のように，まだ番組アーカイブには登録されていない番組がクエリとして指定された場合には, その番組概要に対して, インデックスを作成したときと同様のパラメータ取得のための処理をリアルタイムに行い，関連度の計算に必要なパラメータを取得する．また，生放送のニュース番組のように番組概要がほとんど付与されず, クローズドキャプションが付与されている番組がクエリ番組となる場合は, クローズドキャプションを番組概要として用いる．このようにして，クエリ番組とインデックスに格納しているアーカイブ内の番組のパラメータを用いて，式(11)で各番組の関連度を取得する．同時に，式(12)の値が最大となるn-gram t_iを関連ラベルとして取得する．最終的に関連度のスコア順に指定された順位までの番組を検索結果として取得し, 関連ラベルにより検索結果の分類を行い, 後述する関連番組検索インタフェースにより提示する．

2.4.2 関連番組検索インタフェース

開発した実験システムは， Webブラウザ上で動作する関連番組検索インタフェースを備えている．このインタフェースを用いると，ユーザがある番組を選択すれば，その番組に関連する番組が自動で提示される．検索結果の表示には，関連度にスコア順に並べたランキング形式と関連ラベル毎に番組を集めた表示形式を選択できる．

関連ラベル毎に結果を表示した動作例を**図 7**で説明する. 番組を視聴中のユーザが画面をクリックすると，図のようにインタフェースの中心に視聴中の番組が表示され，その周りに「南極」や「エスペランザ基地」「流氷」などの関連ラベルごとに関連番組が

[4] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html

分類されて表示される．もし周りに興味のある番組があれば，ポインタをかざすと，その番組の概要が提示され番組冒頭の映像が再生される．その番組が気に入った場合にはそれをクリックすると，選ばれた番組がインタフェース画面の中心に移り再生される．さらに，今度は選んだ番組を起点として，新しい関連番組の検索結果が表示される．このように本インタフェースを用いることで，次々と関連する番組を検索していくことができる．

図 7 実験システムのインタフェース

2.5 評価実験

提案手法の有効性を検証するため，構築した実験システムを用いて，関連番組検索の評価実験および番組間の関連ラベル付与に関する評価実験を行った．以下，2.5.1 で関連番組検索の評価実験の結果について，2.5.2 で関連ラベルの評価実験の結果について述べる．

2.5.1 番組検索の評価

関連番組検索の評価実験は，以下に述べる条件で実施した．実験システムが検索対象とする番組概要として1783番組の番組概要を用いた．これは，NHKオンデマンドに同一時期（2009年7月）に登録されていた2218番組からシリーズ番組などで全く同じ番組概要が付与されている番組を除去したものである．また検索において選択されるクエリ番組としては，ジャンルの異なる20番組を用いた．

関連番組検索の性能を検証するため，検索結果を10名の被験者（20-30代，男性5名，

女性5名）により評価した．評価結果のうち，過半数の6名以上が「関連あり」と判断した結果を正解データに用いた. 検索結果の評価法として, Mean Average Precision（MAP）を利用する．MAPは，再現率を考慮した順位付き検索結果を評価する尺度Average Precision（AP）（式(13)）のクエリQの平均で，式(14)のように表される．なお，rは関連番組の総数，Nは検索結果の総数，$pr(r)$はr位における適合率，$rel(r)$はr位の番組が「関連あり」の場合に1, それ以外で0となる関数である．$N = 20$として評価を行った.

$$AP = \frac{1}{N}\sum_{r=1}^{N} \mathrm{pr}(r) \cdot rel(r) \qquad (13)$$

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|} AP_j \qquad (14)$$

比較手法として，以下の６つのベースライン手法を用いた．実験では，検索結果の上位20番組について評価を行った.

ベースライン１：語の重みにtf-idfを用い，類似度の計算にコサイン尺度を利用.
ベースライン２：ベースライン１をn-gram (n=3)に拡張.
ベースライン３：Latent Semantic Indexing（LSI）[Deerwester88] を利用し，tf-idfで重み付けした番組-単語行列を特異点分解により番組-概念行列に変換する．類似度計算には，概念ベクトル間のコサイン尺度を利用する.
ベースライン４：BM25に基づく関連度のみを使用
ベースライン５：ベースライン４をn-gramに拡張
ベースライン６：ベースライン４に固有表現の重みを付与

実験の結果，すべてのベースライン手法に対して，提案手法の性能が上回った．それぞれの手法の実験結果を**表** 4に示す．まず，ベースとなる手法の比較では，提案手法で用いているBM25に基づく関連度を用いたベースライン４が, ベースライン１に比べて，高い評価が得られている．これは，基本的にBM25とtf-idfコサイン尺度による性能の違いに起因している．具体的には，ベースライン４では，クエリ番組から得られるパラメータと番組アーカイブに登録されている検索される側のデータを区別した計算をおこなっていることや, 番組概要の長さなどのベースライン１で利用していないパラメータ

の利用に効果があると考えられる．

n-gramの拡張は，BM25に基づくベースライン４と，それをn-gramに拡張したベースライン５との比較においてベースライン５が高い評価が得られ，その拡張による効果が確認できる．また，ベースライン１とベースライン２の評価の違いからも，n-gramの拡張による効果が見て取れる．これは，n-gramの連鎖数が大きい複合語が共通の番組が上位になり，その番組が評価されたため，評価結果が上昇したと考えられる．

表 4 関連番組検索の結果

Method		Mean Average Precison
提案手法	BM25 + n-gram + NE	0.721
ベースライン１	tf-idf	0.656
ベースライン２	tf-idf + n-gram	0.671
ベースライン３	LSI	0.661
ベースライン４	BM25	0.700
ベースライン５	BM25 + n-gram	0.712
ベースライン６	BM25 + NE	0.704

拡張固有表現の重み付与の効果は，提案手法とベースライン５，ベースライン４とベースライン６との比較により確認できる．具体的には，n-gram拡張なしの場合で0.700から0.704に上昇し，n-gram拡張した場合も0.712から0.721に上昇している．これにより，n-gram拡張の有無によらず，拡張固有表現による重みの効果が確認できる．ここで，n-gram拡張なしの場合に上昇幅が僅かであるのは，そもそも形態素１つで固有表現の絶対数が少ないことと，今回の実験では，ある形態素が固有表現の一部だった際に重みを付与するなどの対応を行っていないためである．そのため，一つの単語が固有表現の場合にしか効果がなかったと考えられる．また，実験では，固有表現の抽出に機械学習による結果を利用したため，その抽出精度が評価結果に影響していると考えられる．組織名，職業名，生物名は人名や地名に比べて精度が低いため，学習データを増やしたり，他の言語資源を利用した素性を追加するなどして精度を向上させる必要がある．

今回の実験では，人名（PER）の重みを1.0となっている．これは，パラメータ設定のための予備実験で，人名の重みを上げると，姓のみや名のみが一致する別人物が出現する番組の順位が上がり，総合的には評価が下がったためである．しかし，直感的には，同一人物が出現している２つの番組概要があれば，その番組間になんらかの関連がある

はずである．そのため，同一人物の判定処理を導入し，同一人物と確認できた表現だけに適切な重みを付与することで，固有表現の効果は更に上昇すると考えられる．

LSIによる関連番組検索のベースライン３では，特異値分解による概念と単語とのマッピングによる実質的なキーワード拡張の効果により他の手法では得られない結果が取得できていた．しかし，評価結果では，そのようなLSIでのみ得られた検索結果があまり評価されず, tf-idfコサイン尺度によるベースライン１および２を僅かに上回るに留まった．これは，対象とする番組概要の単語数や，番組アーカイブの対象番組数が十分でなく, 特異値分解による概念と単語とのマッピングが十分得られなかったためと考えられる．そのため，人間が判断すると，クエリとは内容の関連の薄いように判断される番組が多く結果に出力され，全体として評価が伸びなかったと考えられる．しかしながら，LSIに限らず，同義語辞書や上位下位概念辞書などの言語資源を用いたキーワード拡張の手法をうまく取り入れることで, 提案手法では現状推薦できていないパラフレーズを含む番組を取得できる可能性があり，今後検討していく必要がある．

2.5.2 関連ラベルの評価

提案手法では, 検索結果とクエリ番組との関連性を示唆する関連ラベルを出力している．この関連ラベルが適切に付与されているかを調べるための評価実験を行った．実験では，関連番組検索の評価実験で「関連あり」と判断された結果の関連ラベルに対し，３段階（適切, ほぼ適切, 不適切）で評価を行った．関連ラベルの主観評価の結果を**表 5** に示す．評価実験の結果，被験者が適切と判断したラベルの割合は0.558であり，ほぼ適切と判断したラベルが0.404となり，これらを合わせると0.962と高い割合のラベルが評価されていた．高い評価が得られた理由として，多くの番組で得られたラベルが，クエリ番組と検索結果に共通して出現する, 固有表現や複合語などの文章中の特徴的な複合語が選ばれたため，評価者にとってラベルの意味を容易に把握でき，それを番組の関係として推測できたためと考えられる．

表 5 関連ラベルの評価

適切	ほぼ適切	不適切
0.558	0.404	0.038

また，番組検索の結果を提示する際に，その関係性を示す関連ラベルの提示が関連性の評価に与える効果を検証するための実験を行った．実験では，結果の番組概要を提示

した場合と，番組概要に加え関連ラベルを提示した場合の違いを調べた．実験に用いたクエリ番組として20番組，検索対象とする番組アーカイブとして1783番組を利用した．また，提案手法によりクエリ１番組につき20番組の検索結果を取得した．評定者は２名とし，予めラベルなしの評価をした上で，ラベルありの結果を再評価した．順序効果の影響を軽減するため，ラベルなしの評価とラベルありの評価は一週間の時間をあけて同じ評定者がおこなった．２人の評価者の結果が一致した番組を正解データとして用いた．

表 6 関連ラベル提示による効果

関連ラベルなし	関連ラベルあり
0.618	0.692

実験の結果，**表 6**に示すように，MAPによる評価結果がラベルを提示することにより上昇している．この結果は，ラベルなしの際に「関連なし」と評価されていた番組が，ラベル提示時には，「関連あり」と評価されたことを示している．このような結果が得られた理由としては，関連ラベルが与えられることで，２つの番組概要における関連の観点が明確となり，「関連あり」と評価された番組が複数あったためと考えられる．評価者への実験後のアンケートでも，「ラベルを見ることにより，評価が行い易かった」との内観を得ることができた．このことからも，関連ラベルの提示はユーザが番組を選択する際に有効な手がかりとなる可能性がある．

2.6 外部知識を利用したリランキングの検討

これまでに番組概要に含まれる n-gram と拡張固有表現の重みを用いて番組を検索する手法について述べてきた．本節では，関連番組検索の結果を，世の中で注目されている表現を用いて補正することを検討する．例えば，最近話題になって注目されている人物 A と名前をよく知らない人物 B が共に出ている番組を見ていたときに，関連番組検索の結果として A を取り上げた番組と B を取り上げた番組が提示されれば，ユーザーは人物 A を取り上げた番組をより見ていた番組と関連があると感じるかもしれない．

番組概要に含まれている表現（例えば人名）がどのくらい注目されているかは，番組アーカイブ内の語の頻度や固有表現の種類からだけでは知ることはできない．注目度を取得するために Google などの商用検索エンジンの検索履歴や語の頻度を用いることが考えられるが，現状公開されている商用検索エンジンの API の使用回数には上限があり，実用的に番組検索などで利用することは難しい．また，ショッピングサイトで利用されているようなクリックスルーログを利用することも考えられるが，これらのデータを利

用できるユーザは限られている上，コールドスタート問題があり，かならずしも注目されている語を取得することができない．そこで，ある表現の注目度を取得するための外部の知識として，誰でも利用可能な Wikipedia の更新履歴に着目し，関連番組の検索結果をリランキングする手法の可能性について検討する．

2.6.1 Wikipedia 変更履歴

Wikipedia は世界中のユーザが自由に編集できるオンライン百科事典であり，2010 年 8 月時点で，英語版で 337 万件，日本語版で約 69 万件の記事が登録されている．これらの記事は，様々なユーザが加筆や修正を行いながら作成されている．また変更された記事のすべて履歴は記録されており，誰でもすべての変更履歴の推移を閲覧することができる．Wikipedia では，コンテンツのクロールを禁じているため，その代替手段として，変更履歴を含めコンテンツの内容を一括ダウンロードできる手段を提供している．変更履歴は，XML 形式で記述されており，記事ごとの変更内容と時間，編集したユーザ名を取得することができる．

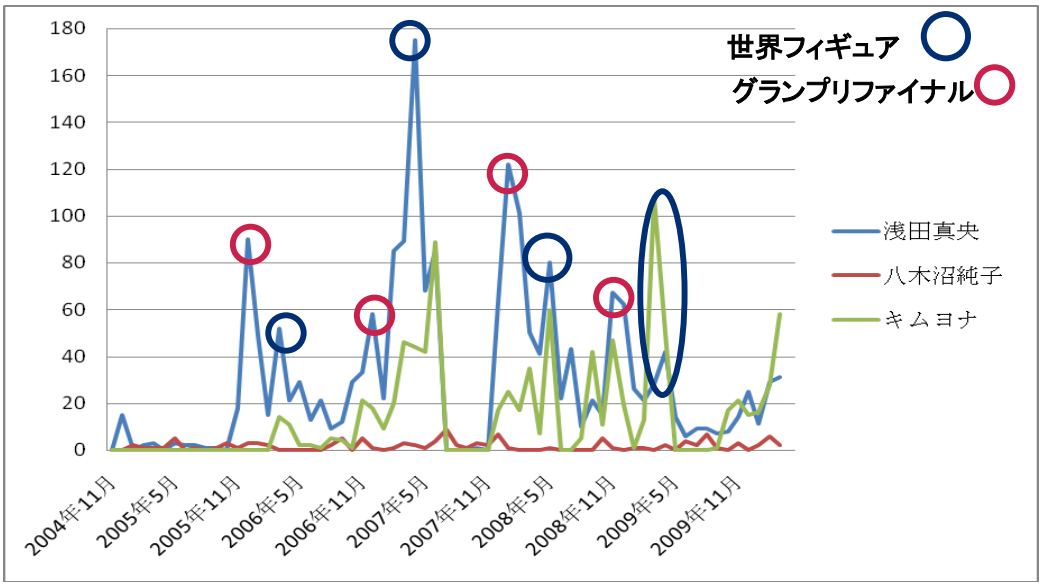


図 8 更新履歴の推移

Wikipedia のタイトルごとに，変更頻度を一定期間で集計し，その合計をタイトルの語の注目されている尺度（注目度）として利用することを検討する．**図 8** に，スケート番組の番組概要に出現する 3 人の人名，浅田真央選手，キムヨナ選手，八木沼純子解説者の変更履歴の推移（月単位）を示す．大きな国際試合がある 3 月，6 月，12 月など

注目されている時期には，浅田選手やキム選手の変更の頻度が上昇する傾向が読み取れる．一方で解説者の変更履歴の頻度に変化はない．このように更新履歴と注目度は一定の相関があると考えられる．

ただし，変更履歴の頻度を取得する際に注意すべき点がある．変更履歴には，同一 ID による短期間の集中的な書き込みや，ユーザ間での編集合戦（Edit War）など，本来，注目度としては複数回カウントすべきでないログが含まれている．そのため，一定期間に集中する同一ユーザによる書き込みを更新数から省くこととする．さらに 1 日毎にユニークな編集ユーザ数をカウントすることにより，ノイズの除去を行う．

編集合戦や不正な更新が続くと，管理者により誰も編集ができなくなる保護期間が設けられる場合がある．**図** 8 の 2007 年 06 月から 11 月までが浅田選手，キム選手の両記事ともに保護期間にあたる．保護期間中の変更の頻度は 0 となるが，本来であれば多くの変更が行われる可能性がある．そこで，保護期間の直前と直後の平均を頻度とする補完措置を行う．

2.6.2 更新履歴に基づく注目度の効果

関連番組検索手法の固有表現による重み $w_{ne}(t_i)$の代わりに，更新頻度に基づく注目度 $w_{wiki}(t_i)$を重みとして利用し，注目度によるリランキングを行う効果を検証する．注目度 $w_{wiki}(t_i)$は，各記事の更新頻度 $freq(t_i)$をそのまま注目度とすると，更新頻度の幅が大きいため，式(12)を用いる．

$$w_{wiki}(t_i) = \log_r(r + freq_{wiki}(t_i)) \tag{12}$$

$$Score_{wiki}(t_i) = \sum_{i=1}^{N} w_{wiki}(t_i) \cdot w_{ng}(t_i) \cdot q(t_i) \cdot d(t_i) \tag{13}$$

Wikipedia の記事の更新では，人名や組織名などのカテゴリの種類により，その傾向が異なる．そのため，底 r をパラメータとして用いて更新頻度をどの程度を考慮するかを定める．$q(t_i)$, $d(t_i)$ にはそれぞれ式(3)と式(4) を用い，$w_{ng}(t_i)$には式(8)を用いる．

更新頻度に基づく注目度の効果を確認するため，番組アーカイブの 9471 番組を対象に，注目度を利用した手法の評価を行った．クエリとした番組は，現代ドラマ，時代劇，スポーツ，報道番組の 4 種類を選び，それぞれ 20 位までの関連番組を検索した．注目度に利用する表現は人名を対象とした．式(12)の対数の底 r は 2 とし，最新の 2 年の変更履歴を注目度として利用した．

提案手法の有効性を示すため，比較手法として，以下の３つのベースライン手法を用いた．

ベースライン１ ： n-gram 拡張を行った BM25 に基づく関連度に対して，拡張固有表現の重みを付与

ベースライン２:ベースライン１で拡張固有表現重み付与なし

ベースライン３: n-gram 拡張を行った tf-idf 重みによるコサイン類似度

評価手法には，Discounted Cumulative Gain（DCG）[Jarvelin02] を利用した（４段階評価０-３）．各ベースライン手法と提案手法について，３位，５位，10 位，20 位での各 DCG のスコアを**表 7** に示し，**図 9** に各順位での評価結果の推移を示す．これによると，提案手法によるリランキングの評価が，すべての順位の DCG でベースラインの各手法を上回っていることがわかる．これは，導入した Wikipedeia の注目度の重みにより該当する番組の順位が上昇したためと考えられる．

表 7 リランキング結果の評価

	Proposed	ベースライン１	ベースライン２	ベースライン３
DCG_3	9.32	8.46	8.58	8.50
DCG_5	12.93	11.58	11.72	11.48
DCG_{10}	18.04	16.60	16.53	15.41
DCG_{20}	24.05	22.34	22.32	21.57

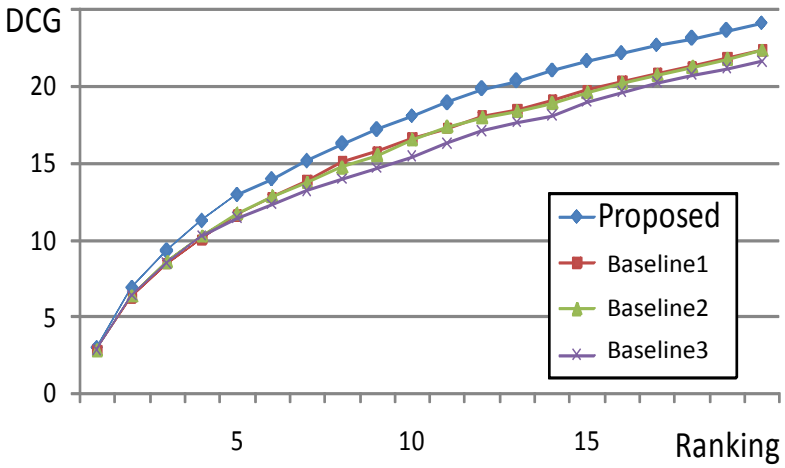


図 9 リランキング結果の推移

将軍・家茂の元に，上洛と攘夷実行を求めて京から勅使が訪れます．・・・・和宮は家茂の身を案じて上洛に反対し，後押ししたのが天璋院だと知って強い敵対心を抱きます．勝麟太郎を斬るためにやってきた坂本龍馬は，勝の進歩的な考えに感銘を受け，弟子になりたいと志願します．

図 10 クエリ番組の番組概要

例えば，**図 10** の番組概要（下線は人名）を持つ番組「篤姫」をクエリとした際の番組検索の結果では，Wikipedia の更新履歴において書き込みが多く注目度の高い「坂本龍馬」の重みにより，リランキングの結果で「龍馬伝」や「そのとき歴史が動いた」などの番組がその順位が上昇している．

実験結果から Wikipeia の更新履歴に基づく注目度を利用することにより，番組アーカイブ内の番組概要のみを利用した番組検索結果の評価を改善できる可能性を確認できた．今後，Wikipedia 以外の外部情報を利用した，検索結果をリランキングする手法についても検討する必要がある．

2.7 提案手法の応用

本研究で提案した番組概要に基づく関連番組検索手法は，EnVison[5]と呼ばれる言語情報と映像情報の解析技術を利用した番組検索用のライブラリに採用され，現在，このライブラリを通して様々なサービスに用いられている（ただし，Wikipediaの更新履歴に基づく注目度は現状使用されていない）．EnVisionが導入された事例としては，NHKがサービスしている番組オンデマンド[6]や番組アーカイブ[7]において，関連する番組の提示機能のための処理に利用されている．別の事例としては，クリエイティブライブラリ[8]での映像素材の検索にも活用されている．クリエイティブライブラリは，様々な番組に用いられている素材を教育用等のために提供するサービスである．クリエイティブライブラリの各映像素材には，その映像の内容や使用された番組を記述した説明文が付与されているため，この説明文を番組概要の代わりに利用し本検索手法を動作させている．

[5] http://www.nes.or.jp/transfer/catalog/

[6] https://www.nhk-ondemand.jp/

[7] http://www.nhk.or.jp/archives/

[8] http://www1.nhk.or.jp/creative/

図 11と**図 12**に，番組アーカイブとクリエイティブライブラリでの導入例をそれぞれ示す．

図 11 番組アーカイブの検索への応用

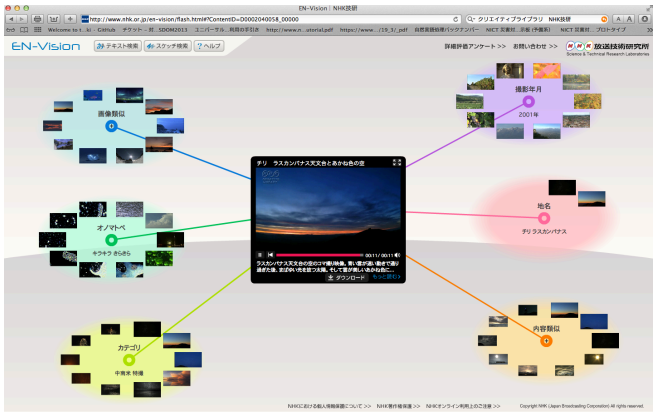


図 12 映像素材の検索への応用

上記で紹介した応用例は，放送された番組やその素材が検索の基点となっていたが，通信上のWebコンテンツを基点とした番組検索への応用として，ツールバー形式の実験用アプリケーションを開発した（**図 13**）．このアプリケーションを利用すると，ブラウザでWebページを閲覧している際に，その内容に関連する番組を簡単に検索できる．常時利用しているブラウザにこのツールバーをプラグインし，ツールバー上に現れる「TV Search」のボタンを押すだけで，システムが自動でWeb ページのコンテンツと番組アーカイブ内の番組の概要との関連度を計算し，見ているWebページの内容に関連している

番組を提示してくれる．例えば，ニュースサイトを閲覧していて，気になる記事の詳細を知りたいときに１クリックで関連ニュース番組をすぐに再生したり，「月食」などのWikipediaの記事から関連する特集番組やドキュメンタリー番組を探したりすることが可能となる．このツールバーを利用することで，Web上のコンテンツ閲覧とのシームレスな番組視聴環境を実現することができる．

このように，本研究で提案した関連番組検索手法は，自然文の概要が付与されている様々なコンテンツ検索に対して導入することが可能である．

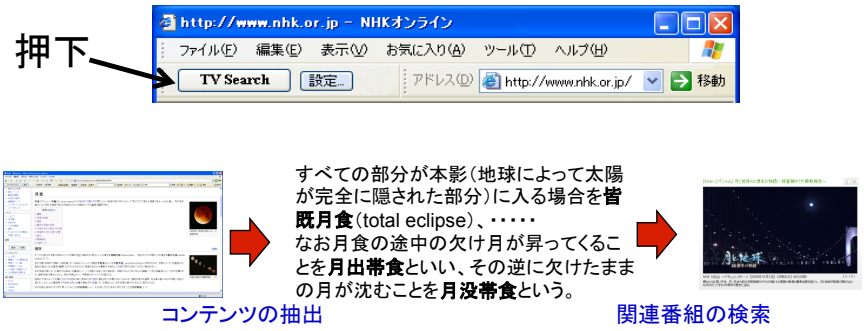


図 13 閲覧中のWebコンテンツからの番組検索

2.8 まとめ

本章では，番組アーカイブへのアクセス手段として，番組に付与されている番組概要を利用した関連番組検索手法を提案した．提案手法を用いることにより，放送される番組や番組アーカイブに登録されている番組を選ぶだけで，関連する番組を検索することができるようなる．番組オンデマンドのコンテンツに対して行った評価実験の結果，比較手法に比べて，拡張固有表現や複合語を考慮した提案手法が，関連番組を高精度に検索できることを示した．また，提案手法が出力する番組間の関係性を示す関連ラベルが，検索結果が提示された理由を推測する手掛かりをユーザに与え，検索結果の評価を向上させることを示した．また，Wikipediaなどの誰もが利用可能な外部知識の更新履歴に基づく注目度を利用し，検索結果をリランキングする手法の検討を行った．評価実験の結果，提案するリランキング手法が更新履歴による注目語を含む番組の順位を上昇させ，全体の検索結果の評価を向上させる可能性を示した．また，本研究で提案した手法の種々の実用化例を示し，本研究の成果が社会に還元されていることについて述べた．今後，普及が進むことが予想されるハイブリッドキャストを用いて，本研究をさらに発展させた様々な応用を行う予定である．

第３章 番組内容把握のための人物表現間の関係抽出とその応用

3.1 はじめに

番組間の関連性を考える際, その番組に登場する人物, 国, 組織, イベントなどの様々なエンティティ間の関係を把握することは重要である．前章では，番組の内容を記述した番組概要から n-gram を取得し，その n-gram の頻度や固有表現の有無に基づき番組アーカイブから関連する番組を検索する手法について述べた．検索された番組は，多くの場合, イベントや人物などのエンティティを通しての関連性がある番組を取得できている．しかし，エンティティ間にある関係性までは考慮することができないため，構成やエンティティ間の関係性がよく似ている番組を提示することができない．例えば，単語ベースでの手法を用いると，「ＡがＢと争っている」「C が D と争っている」のように同じ関係性を持っていてもそのエンティティが異なっていると高い関連性を得ることができない．逆に「ＡがＢと争っている」と「ＡがＢと協力している」いう文は，意味的には，反対のことが述べられているが，ＡとＢが共通しているため，エンティティ間の関係を考慮せずに関連性ありと提示してしまう．人が文章を読み理解するように番組概要におけるエンティティ間の関係の構造を把握することができれば, より関連性の高い番組を取得できる可能性がある．つまり，エンティティそのものよりも，それらの関係性に重きを置いた検索を行うことも可能となる．

本章では，内容を考慮した番組検索の実現のために，番組概要からエンティティ間の関係を取得することを検討する．特に番組において重要なエンティティである人に着目し，登場人物間の関係を取得する．本章の構成は，まず次節において，人間関係抽出の関連研究について述べる．次に 3.3 節で人物表現間の関係抽出の流れについて説明し，3.4 節から 3.6 節で関係抽出の各処理について述べる．続いて，3.7 節で実験システムの構成，3.8 節でその評価実験の結果について述べる．さらに，3.9 節で本手法による結果を利用した応用例について紹介し，3.10 節でまとめと今後の課題について述べる．

3.2 関連研究

人間関係をテキストから取得する研究として，Web 文書から人名の共起頻度を取得し，それらの人物間の人間関係を取得する研究が行われている．松尾らは Web に出現する研究者の共起頻度から関係の有無を求め，特定の語彙との共起頻度に基づき 4 種の関係ラベルを付与し，人間関係のネットワークを生成している [Matsuo06]．Mika らは，Web 文書に加えて，メール，出版物などのタグ付けされた情報を用いて，人名の共起やリンク構造を解析することで，人間関係を示すネットワークを生成している [Mika05]．これらの研究では，関係の有無の判定に，文や文章内での人名の共起頻度を用いており，対象とするコーパスが一定の規模より小さい場合や，人名が多頻度でコーパスに登場しない場合，共起する人名のペアを得ることが難しくなる．また，人物間の関係はあらかじめ人手で作成した語との共起により取得しており，特定のドメインの関係性以外は取得することはできない．

また，小説などのストーリー把握のために，実際の小説の全文章を解析し，その登場人物間の関係を取得する研究が行われている．馬場らは，小説のテキストデータから登場人物とその属性を抽出するとともに，人名が出現する「場面」という概念を用いて，場面に共起する人物には関係があると捉え，人物相関図を生成している [馬場 07]．神代らは，小説のテキスト中の会話に着目し，会話の前後に出現する人名を抽出し相関図を生成する手法を提案している [神代 08]．これらの研究においても，人間関係はその共起頻度により取得されている．そのため，人物間にどのような関係があるかまでは取得できない．また，対象としている人物表現は人名のみであり，代名詞や一般名詞が同じ人物を指すことまでは考慮されていない．

番組概要などの小規模なテキストから人物間の関係を抽出するには，小説や Web などの大規模なテキストの解析のように，単に人名の共起頻度を利用することは難しい．番組の要約である番組概要に人名が複数出現することは多くないためである．番組概要のような小規模なテキストから人物間の関係を抽出するには，番組概要を構成する文の構造を把握し，また異なる文の間で同じ人物を指す表現を特定し，人物表現間にある関係を確実に取得する必要がある．小規模な文書から確実に人間関係を抽出する研究は，筆者の知る限り，文章理解の困難さを理由にあまり行われていない．そのため，本研究では，小規模な文書から幅広い登場人物の関係を取得するため，人名以外の表現についても取り扱い，同じ人物を指す表現の共参照の関係を特定した上で，対象の文書に記述されている人間関係を獲得する．

3.3 番組概要からの関係グラフの取得

本節では，番組概要からエンティティとその関係を取得するために，エンティティ間の関係を示す情報である関係グラフを取得する手法について述べる．番組概要から関係グラフを取得する処理の流れを**図 14** に示す．番組概要とは，前章でも説明したとおり，番組に付与されている内容を簡潔に示したテキストである．本研究では，NHK の映画番組およびドラマ番組の番組概要を用いている．

提案手法では，まず，番組概要から人物を示している言語表現を抽出する．人を示す表現には「オードリー・ヘップバーン」などの人名のほかにも人物を示す表現が含まれる．提案手法では，後述する４種類の人物表現を抽出する．本研究では，人を示す表現を番組概要から抽出する処理を人物表現抽出処理と呼ぶ．

次に，複数の箇所で特定された人物表現が同じ人物を指しているかどうかの判定をおこなう．人物表現抽出処理で得られた表現は，表層的には異なるものも多く含まれるが，そのいくつかは同じ人物を指している場合がある．例えば，文章中に「オードリー」という人名と「彼女」という代名詞があった場合に２つの表現が同じ人物を指すかどうかの判定を行う．以後，上記で説明した処理を共参照解析処理と呼ぶ．

さらに，番組概要に対して構文解析を行い，文節間の係り受け関係を示す構文木を取得する．人物表現抽出処理で得られた人物表現と構文木を利用して，人物間の関係を取得する．例えば，「ブラッドリーはアンと遭遇した」という文からは，「A は B と遭遇した」という関係が得られる．このように２つの人物表現の間の関係を取得する処理を関係抽出処理と呼ぶ．

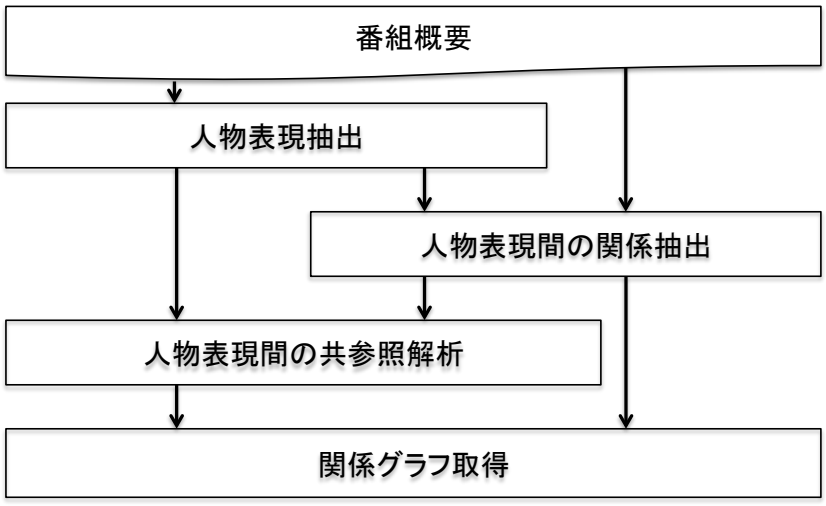


図 14 関係グラフ取得の流れ

最後に，共参照解析処理で同じ人物と判定された人物表現の結果を用いて，関係抽出処理で得られた複数の人物間の関係を接続する．これにより，人物表現をノード，人物間の関係をエッジとしたグラフ構造を持つ人物表現間関係の情報である関係グラフが取得できる．関係グラフの結果からノード間の関係を取得することで，構文解析の結果から直接得られない人物表現間の関係も獲得することができる．上述した関係グラフ取得のための各部分処理の人物表現抽出処理，共参照解析処理，関係抽出処理については，3.4 節から 3.6 節で詳細に述べる．

3.4 人物表現抽出

番組概要のようなテキストから人間関係を抽出する場合に，テキスト上に現れる登場人物を示す表現を特定する必要がある．先行研究 [Mika05] [Matsuo06] では，人物関係の抽出は，特定のコミュニティでの関係を抽出する目的のため，人名リストが与えられた条件で行われている．番組概要に現れる登場人物の言語表現は多種多様であり，事前にそのリストを用意することはできない．そのため，本研究では，人物を示す表現を番組概要から自動で抽出する．以下に，本研究で取り扱う４種類の人物表現について説明する．これらの人物表現の種類は，NHK の映画番組 207 本と連続ドラマ８シリーズの番組概要を人手で分析し決定した．

3.4.1 人物を表す固有表現

番組概要においても，人物を示す表現としては人名を使用することが多い．しかしながら，人名以外の表現が代わりに用いられることがある．例えば，**図 15** に示す映画の番組概要例では，職業や地位を表す「王子」や「首相」が人物を示す表現として用いられており，文章中に王子や首相の人物の名前は現れない．そのため，本研究では，人名（PERSON）に加えて，２章でも利用した拡張固有表現の職業名・称号名（TITLE）を人物表現の固有表現として抽出する対象とする．これらの表現の認識には，後述する教師あり機械学習により作成した認識器を利用する

Ｂ・バルドーが，フランス首相の一人娘ブリジットという令嬢役にふんしたラブ・コメディー．官房長のミシェルに恋焦がれるブリジットは，初めは相手にされなかったものの，ちょっとした作戦が功を奏し，見事ミシェルと結婚．しかし，彼の派手な女性関係にうんざりした彼女は，国賓の王子との浮気を宣言する．王子役はアカデミー主演男優賞に４度ノミネートされたフランスの２枚目スター，シャルル・ボワイエ．

図 15 映画の番組概要例

3.4.2 一般人物表現

人物を示す表現として, 固有表現ではない一般名詞が用いられる場合がある. 例えば, 「男」「人」「女性」などの表現である．番組概要のなかには，これらの人を示す一般名詞のみが使用され, 固有表現である人名がまったく出現しない場合もある. そのため, 人物を示す一般名詞を一般人物表現（G-PERSON）として定義し，これらを人物表現の一つとして抽出する.

行くところのない男は休暇の１０日間だけ，シナリオライターを目指して執筆に励む彼女と同居することになるが...

上記の例では，人名は文中に一切登場せず，「男」という一般名詞で登場人物を表現している. この番組概要から主要な登場人物が得るためにも一般人物表現を抽出する必要がある．一般人物表現は「アメリカ-人」「構成-員」などのように複数の形態素からなる複合語の表現があり，そのバリエーションは多いため，固有表現の抽出と同様に教師あり機械学習により作成した認識器により抽出を行う.

3.4.3 関係人物表現

一般人物表現のなかには, 人物を表すとともに他の人物との関係性を示す特殊な役割の一般名詞がある．例えば，「父」「一人娘」「上司」「部下」「雇用者」「雇い主」などのような表現である．これらの表現では，その表現自体が人物を示しているのと同時に，別の人物表現間との関係を示す．そのため，このような表現を通常の一般人物表現と区別し，関係人物表現（R-PERSON）として抽出する．また関係人物表現は，人物間の関係を抽出する際にも利用することができる．例えば，「ボブの部下は・・・」という表現では，「部下」は人物表現であると同時に，ボブとの関係性も示す．関係人物表現についても機械学習による認識器による抽出を行う.

3.4.4 人称代名詞

日本語では，人名が出現した後の文においても，その人名を使用しつづけることが多い．例えば，「安倍首相は国際オリンピック総会への参加を決めた．そこで，安倍首相は東京招致のための最終プレゼンテーションを行う予定である．」というように，特にニュース記事等では，「安倍首相」を代名詞の「彼」などに置き換えることは極めて少ない．しかしながら，映画やドラマの番組概要の文章では，その文字数に制限があるため，「彼」「彼女」などの人称代名詞が比較的よく使用される．映画番組 378 本の概要を調べたところ，158 本（41.8%）で人称代名詞が用いられていた．そのため，番組概

要においては，人物を示す表現の１つとして，人称代名詞の抽出が重要であると考え，人称代名詞も人物表現として抽出を行う．なお，人称代名詞はその表現のバリエーションは限られているため，形態素列のパターンマッチングで抽出する．

3.5 人物表現間の共参照解析

番組概要中の人間関係を把握するためには，得られた人物表現のうち，どの表現が同じ人物を示すのかを特定する必要がある．これまでに述べたとおり，番組概要では，人物を示す表現は人名以外にも一般名詞や代名詞も用いられるため，文字列の類似度だけでは，その共参照を判断することが難しい．表層上は異なる表現であっても，同一の人物を示す場合や，逆に同一の表現でも異なる人物を指す場合もあるからである．そのため，同じ人物を指す表現を特定する必要がある．

飯田らは，最尤の照応先を決定するトーナメントモデルによる共参照解析手法を提案している [飯田 05]．この手法では，対象をあらゆる名詞としているため，エンティティ固有の素性は取り入れていない．本研究では，その目的が人物間の関係を取得することであるため，共参照解析の対象についても人物表現抽出により得られた表現に限定する．これにより，名詞全体の集合から共参照先を選ぶタスクに比べ，大幅に候補の対象を削減でき，精度の高い共参照の関係を取得することが期待できる．また，性別などの人物の表現固有の素性を利用できる．以下に，本研究で提案する人物表現間の共参照解析処理について説明する．

3.5.1 代名詞の利用

共参照の解析では，一般に照応する元となる照応詞の選択とその照応先（先行詞）の決定を行う処理に分けられる．本研究では，照応詞として，「彼」，「彼女」等の代名詞と，連体詞に続く一般人物表現（「その男」，「その人」等）を用いる．先行詞は後方照応は考慮せず前方照応のみを対象とし，照応詞より前に出現する人物表現を対象とする．先行詞の判定は，照応詞が含まれる文の前文から始め，文内の照応順序は，以下に示すセンタリング理論に基づく表層格を利用した優先度 [Walker94] を利用する．

主題（ハ）＞ 主語（ガ格）＞ 間接目的（ニ格）＞ 直接目的（ヲ格）＞ その他

対象の文で主語などが省略され先行詞の候補がない場合は，順に前の文を対象としていく．先行詞を決定する際には，性別と単数複数の属性による制約を設ける．つまり，照応詞の属性と異なる人物表現は先行詞にはならない．例えば，照応詞が「彼」であった場合，性別が違う女性名や，複数表現である「兵士達」などは先行詞とは判定されない．

性別の判定は，一般名詞については人手で作成した性別辞書により判定し，男性，女性名，不明を判定する．例えば，「父」「母」「叔父」「妹」「息子」などの関係人物表現や，「看護婦」「保母」「スチュワーデス」などの職業名,「皇太子」「妃殿下」などの Title でも性別が判定することができる．もちろん，多くの一般名詞は，性別を含まないため，辞書に含まれないものは特に制約は設けない．人名の性別判定については，後述する教師あり学習による判別器で判定する．

官房長のミシェルに恋焦がれるブリジットは，初めは相手にされなかったものの，ちょっとした作戦が功を奏し，見事ミシェルと結婚．しかし，彼の派手な女性関係にうんざりした彼女は，国賓の王子との浮気を宣言する．・・

上の例文を用いて，「彼」の照応先が「ミッシェル」に決まる手順を説明する．まず，「彼」の照応先の候補は，人物表現抽出処理により前の文から抽出された「ブリジット」もしくは「ミッシェル」となる．この２つの候補のうち，センタリング理論に基づく格の優先度を用いると，主題である「ブリジット」が第一候補となる．しかし，「ブリジット」は性別判定により女性名であるため，代名詞の「彼」と矛盾が生じ，次の候補「ミッシェル」となる．「ミッシェル」は，男女両方で用いる名前であるため，矛盾は生じないため「彼」の照応先となる．

3.5.2 構文解析結果の利用

構文解析結果において，ある人物表現が別の人物表現に係っており，係元の人物表現の格がノ格である場合，これらを共参照とする．例えば，「官房長のミッシェル」という表現があった場合，「官房長の」という文節が「ミッシェル」という文節に係っており，「官房長の」の文節の格助詞が「の」であるため，２つの人物表現「官房長」と「ミッシェル」を共参照とする．ただし，例外として，関係人物表現がノ格の係り先であった場合はこれらを共参照の対象から除外する．例えば，「親友のポール」と「ポールの親友」という２つの表現は，両表現ともに，１つの人物表現が他方の人物表現にノ格を通して係っている．「親友のポール」では，「親友」と「ポール」は同一人物を指すが，係先が関係人物表現である「親友」となる「ポールの親友」という表現では，「ポール」と「親友」は別の人物であり同一人物とはならない．

複数の人物表現が同一文節に含まれているとき，これらの人物表現を共参照とする．例えば「一人娘ブリジット」という表現では，その同一文節に「一人娘」と「ブリジット」という２つの人物表現が含まれているため，これらの人物表現を共参照として取り扱う．

3.5.3 特定のパターンの利用

後述する関係抽出処理により得られた関係を持つ人物表現のペアの中で，予めデータベースに登録したパターンと同じ関係を持つものを共参照とする．例えば，「ＡはＢである」「ＡがＢと呼ばれる」「ＡはＢと名付けられる」などの特定のパターンを関係としてもつ名詞を共参照として取得する．また，「A を演じる B」「A がふんする B」などの俳優と役名との関係がある場合も特定のパターンとして取り扱う．特定のパターンは開発データに含まれている全パターンから人手で抽出している．

3.5.4 文字列の類似性の利用

人物表現の文字列の類似性により共参照を取得する．ただし，これまで述べてきた代名詞による共参照や，係り受けによる共参照，特定の関係による共参照などの結果に矛盾がある場合は，表層が一致しても共参照とはしない．例えば，「ポール」という名前が複数箇所に使われていれば，それらの「ポール」は，同名の登場人物がいるレアケースをのぞいて，ほとんどの場合同じ人物を指すだろう．しかし，人名以外の人物表現については，同じ文字列でも異なる人物をさす場合がある．例えば，番組概要に「刑事のジョン」と「刑事のポール」が出現した場合を考える．係り受けによる共参照解析の結果，２カ所に出現する「刑事」はそれぞれジョンとポールという２つの人名と共参照が得られる．そのため，人名の間に矛盾が生じるため，２カ所に出現する「刑事」は文字列としては同じであるが共参照とはならない．

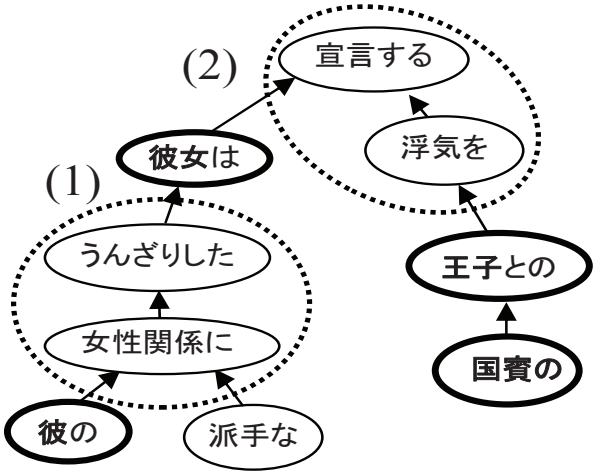
以上，これまでに述べた４種類の共参照となるケースを説明してきたが，本論文では，代名詞による共参照を Pronominal，ノ格や同一節など係り受けに基づく共参照を Dependency，特定の関係に基づく共参照を S-relation，人物表現の表層の類似度に基づく共参照を Text-Sim と表す．

3.6 人物表現間の関係抽出

番組概要では，限られた文字数で登場人物やその人間関係を簡潔に紹介しなければならないため，1 文中に複数の人物表現が出現する場合もある．関係抽出処理では，１文中に出現する人物間の関係をすべて抽出する．また，抽出される人物間の関係には，「A の父の B」などの血縁などの恒久的な関係と，「A が B に恋する」などのように人物間に起こるイベントがあるが，本研究では，これらの両方の関係を抽出すべき人物間の関係として取り扱う．

3.6.1 構文解析に基づく関係抽出

Hasegawa らは，英語の言い換え表現の獲得のため，人名や組織などの固有表現の間に出現する文字列を関係として取り出している [Hasegawa04]．日本語の解析では，英語とは語順が異なり，文字列で見ると２つの表現の間に関係が現れないことも多い．そのため，構文解析を用いて係り受けの結果（構文木）を取得し，得られた構文木と人物表現抽出の結果を用いて，人物間の関係を抽出する．



彼の/派手な/女性関係に/うんざりした/彼女は/国賓の/王子との/浮気を/宣言する．

図 16 構文木からの関係抽出

図 16 に構文解析の結果から関係を取得する例を示す．図の下部に示す文を構文解析した結果が図の上部の構文木である．まず人物表現抽出で得られた固有表現と構文木の節との対応を取る．人物表現を含む文節は，「彼の」「彼女は」「国賓の」「王子との」である．このすべての文節を結ぶパスを考えて関係を取得する．ただし，そのパス内に人物表現がある場合は関係として取得しない．例えば，「国賓の」と「彼女は」からは「王子との浮気を宣言する」という関係が取得できるが，人物表現の「王子」を含んでいるため，人物間の関係とはしない．これは，「王子との」「彼女は」の文節からより適切な関係の「浮気を宣言する」が取得できるためである．例では，最終的に「彼の」と「彼女は」の文節の間にある表現（1）「A の女性関係にうんざりした B」，「彼女は」と「王子との」の文節の間にある表現（2）「A は B との浮気を宣言する」を関係表現として取得する．また，人物表現の節に含まれる格助詞は，関係の方向性を利用する際の情報として取得される．

3.6.2 関係抽出のためのゼロ主語の補完

日本語では, 主語や主題は, 一度言及されるとその後の文では省略されることが多い. また, 複文では, ほとんどの場合, 前後のどちらかの文で主語や主題が省略されている. このような省略された語は，ゼロ代名詞と呼ばれ，前の文の主語や主題を照応していることが多い. 本研究で用いる関係抽出は１文ごとの構文解析結果を用いて人物表現間の関係を取得するため，主語など人物表現の省略があると，文章中に２つ以上の人物表現が現れず，関係を抽出できない場合がある．そのため，主語や主題が省略されている文については，以前の文から尤もらしい名詞の候補を特定し，人物表現について主語と主題の補完を行う．補完には，共参照解析処理でも利用したセンタリング理論に基づく表層格を利用した優先度を利用する．

美人を見るとめまいがして，まっとうな恋ができない大学の数学教授が，プラトニックな関係を求めて恋人を募集する．(数学教授が)応募してきてきた同じ大学の文学教授とつきあうことになるが・・・．

上記の例文では，「数学教授」と「文学教授」は，別の文に現れるため，直接の係り受け関係を得ることができない．そのため，主題の補完を行い，第１文の「数学教授」を第２文の主語として補完を行う．これにより，補完処理を行わない場合には得られない「数学教授」と「文学教授」の間の関係「A が B とつきあう」を取得することができる．

3.6.3 関係抽出のための複文の分割

関係抽出の精度は構文解析の精度に大きく依存するため，構文解析の精度を向上させることは重要である．構文解析は，使用するアルゴリズムや学習データに大きく依存するが，一般に重文や複文などの文では，その精度が低い傾向がある．そこで，複文や重文を単文へ分割することによって，構文解析の精度向上を図る．複文の分割には，機械翻訳のための連用中止や連体節のパターンマッチングによる手法 [Kim94]を用いる．[Kim94]では，ニュースを対象とした実験の結果，87.9%の分割成功率が得られたことが報告されている．また，単文へ分割した際に主語がない場合については，3.6.2 と同様にセンタリング理論による格の優先度を利用して，ゼロ主語を補完する．例えば，以下の例では，(1)の文が(2)のように２文に分割されている．さらに(2)では，分割された後ろの文に主語もしくは主題がないため，前の文から主題の「ドン」を取得して補完している．

(1) **スペイン貴族の末裔ドンは，結婚式の最中，殺人犯の汚名を着せられるが，逃げる途中に出会った牛追いのサムに用心棒として雇われる.**

(2) **スペイン貴族の末裔ドンは，結婚式の最中，殺人犯の汚名を着せられる．しかし，(ドンは) 逃げる途中に出会った牛追いのサムに用心棒として雇われる.**

上記のようにして得られた人物表現間の関係は，3.5 節で述べた人物表現間の共参照解析処理の結果により複数の関係を接続することで，グラフ構造を持つ人間関係の関係グラフを生成することができる．関係グラフは，そのノードに同じ人物を示す複数の人物表現を含み，エッジに同じ人物間の複数の関係が格納される．そのため，関係グラフを用いることで，そのノードとエッジに含まれる表現の組み合わせにより，構文解析結果から直接取得できない人物間の関係を取得することができる．

3.7 実験システム

これまでに説明した手法を利用して，番組概要などの自然文で記述された文章から人間関係グラフを自動で生成するシステムを開発した．開発データには，NHK の映画番組 207 本と連続ドラマ 8 シリーズの番組概要を用いた．また開発データに対して，人名（PERSON），職業名・称号名（TITLE），一般人物表現（G-Person），関係人物表現（R-Person）の人物表現を付与したタグ付きコーパスを作成した．

人物表現抽出処理では，上記のタグ付きコーパスを CRF により学習を行った人物表現抽出器を用いる．学習に利用した素性には，形態素の表層，読み，品詞，文字種，各単語の前後の文字 n-gram（N=1, 2, 3），活用形，EDR の概念辞書の上位概念を用いた．形態素解析器には，Chasen[9]を用いた．人物表現抽出のうち，代名詞の抽出については，その種類が少数で限定されるため，辞書とのパターンマッチングにより抽出する．

人物表現間の共参照の同定は，前節で説明した規則処理により実装した．共参照の同定で用いる人物名の性別判定では，放送番組の概要から取得した重複のない 1.5 万人の人名の性別（男性，女性，Unknown）を Multi-class SVM [Crammer01] により学習した認識器を用いた．人名は，漢字，カタカナ，ひらがなの表記のみを対象とし，アルファベットを含む人名は対象外とした．学習の素性として，表層文字列，読み，文字種，文字列長，特定文字の有無を用いた．その精度は，5000 人（学習データに含まれない）の性別の精度を測ったところ，F 値で男性が 0.899，女性が 0.881 であった．

[9] http://chasen-legacy.sourceforge.jp/

人物表現間の関係を取得する関係抽出処理では，構文解析結果から得られた構文木と人物表現抽出で得られた人物表現を用いて，人物表現間の関係を取得する．構文解析器には，CaboCha[10]を用いた．3.4.2 節で説明した複文の対応には，Kim らの手法により単文に分割する処理，ゼロ照応を補完する処理を実装した．

図 15 の番組概要（以下に再掲載）を入力とした場合のシステムの動作例を以下に示す．

Ｂ・バルドーが，フランス首相の一人娘ブリジットという令嬢役にふんしたラブ・コメディー．官房長のミシェルに恋焦がれるブリジットは，初めは相手にされなかったものの，ちょっとした作戦が功を奏し，見事ミシェルと結婚．しかし，彼の派手な女性関係にうんざりした彼女は，国賓の王子との浮気を宣言する．王子役はアカデミー主演男優賞に４度ノミネートされたフランスの２枚目スター，シャルル・ボワイエ．

表 8 人物表現抽出処理の実行結果

人物表現	人物表現の種別	人物表現	人物表現の種別
B.バルドー	PERSON	ミッシェル	PERSON
首相	TITLE	彼	PRONOUN
一人娘	R-PERSON	彼女	PRONOUN
ブリジッド	PERSON	国賓	TITLE
令嬢	R-PERSON	王子	TITLE
官房長	TITLE	王子	TITLE
ミッシェル	PERSON	２枚目スター	TITLE
ブリジッド	PERSON	シャルル・ボワイエ	PERSON

表 8 は人物表現抽出処理の実行結果とその人物表現の種別である．**表** 9 は，人物表現のうち，関係が抽出できたペアとその関係（人物表現の文節の格助詞は表示していない）を示している．**表** 10 は，共参照の関係にある人物表現のペアとその共参照の種類を示している．このようにして得られた人物表現間の関係抽出と共参照解析の結果を用いて取得できた関係グラフを**図** 17 に示す．なお，この結果は関係グラフを Graphviz[11] のフォーマット dot で記述し，独自開発のソフトウエアで描画したものである．関係グラフの各ノードは人物の性別により異なる色（女性：桃，男性：青，性別不明：灰）で表示され，ノード中の人物名の表示では，B・バルドーなどの俳優名を赤字で表示している．

[10] http://code.google.com/p/cabocha/

[11] http://www.graphviz.org/

表 9 関係抽出処理の実行結果

関係表現	人物表現のペア	
ふんする	令嬢	B.バルドー
という	ブリジッド	令嬢
一人娘	首相	ブリジッド
恋い焦がれる	ミッシェル	ブリジッド
結婚	ブリジッド	ミッシェル
女性関係にうんざり	彼	彼女
浮気を宣言	彼女	王子
役は	王子	シャルル・ボアイエ

表 10 共参照解析処理の実行結果

人物表現のペア		共参照の種類
ミッシェル	彼	Pronominal
ブリジッド	彼女	Pronominal
ブリジッド	ブリジッド	Text-Sim
ミッシェル	ミッシェル	Text-Sim
王子	王子	Text-Sim
B.バルドー	令嬢	S-relation
ブリジッド	令嬢	S-relation
王子	二枚目スター	S-relation
一人娘	ブリジッド	Dependency
官房長	ミッシェル	Dependency
二枚目スター	シャルル・ボアイエ	Dependency
国賓	王子	Dependency

図 17 関係グラフ取得の実行結果

3.8 評価実験

提案手法の性能を検証するため，実験システムを用いて関係グラフを生成し，その評価を行った．実験のテストデータとして，映画番組 90 本の番組概要（開発データには含まれない）を用いた．このテストデータに対して，人手で人物表現，その人物表現間の関係，人物表現間の共参照の情報を付与した正解データを作成した．この正解データとシステムが出力した結果を比較し評価した．以下に，提案手法を構成する各部分処理について，3.8.1 で人物表現抽出処理，3.8.2 で人物表現間の共参照解析処理，3.8.3 で人物表現間の関係抽出処理ついてそれぞれ評価を行う．また 3.8.4 で人物表現間の関係を利用した番組検索への応用について検討する．

3.8.1 人物表現抽出の評価

生成した人間関係グラフのノードに含まれる人物表現が正しく抽出できているかを検証した．そのため，番組概要から人物表現を抽出する人物表現抽出処理の結果を評価した．評価実験の結果を**表 11** に示す．人物表現グラフのノードについての結果は，適合率 0.912，再現率 0.903，F 値 0.907 が得られた．再現率，適合率の定義は以下の通りである．

$$適合率 = \frac{\text{正しく抽出できた人物表現の形態素数}}{\text{システムが抽出した人物表現の形態素数}}$$

$$再現率 = \frac{\text{正しく抽出できた人物表現の形態素数}}{\text{人手で抽出した人物表現の形態素数}}$$

共参照解析や関係抽出などの後段の処理のベースとなる結果であるため，その性能は重要である．人物表現の種類ごとに個別にその精度を検証すると，G-PERSON と TITLE の F 値がそれぞれ 0.845，0.844 となっており，他の人物表現に比べると低い結果となった．これは，素性として用いた EDR シソーラスの分類とテストデータの文中に出現した語彙が一致しないものがあり，意味の素性の効果が十分得られなかったためと考える．TITLE については，「運び屋」「新入り」「出納係」など語彙の種類が多岐にわたっているが，今回用意した学習データは十分でなかったためと考えられる．そのため，TITLE を含む学習用コーパスを増やすか，他の言語資源から取得した職業リストなどを利用した素性を加えることにより改善が可能と考えられる．

表 11 人物表現抽出の結果

人物表現の種類	適合率	再現率	F 値
PERSON	0.940(202/215)	0.918(202/220)	0.929
TITLE	0.855(141/165)	0.834(141/169)	0.844
G-PERSON	0.842(117/139)	0.848(117/138)	0.845
R-PERSON	0.975(119/122)	0.992(119/120)	0.983
PRONOUN	1.000(62/62)	0.984(62/63)	0.992
TOTAL	0.912(641/703)	0.903(641/710)	0.907

3.8.2 人物表現間の共参照解析の評価

人間関係グラフの生成では, 同一人物を示す人物表現を１つのノードに結合させている．この関係グラフにおけるノード結合の性能を調べるため，共参照解析による結果の評価を行った. 実験の結果, 適合率 0.892, 再現率 0.725, F 値 0.800 が得られた（**表 12**）. 再現率，適合率の定義は以下の通りである.

$$適合率 = \frac{\text{正しく同定できた人物表現間の共参照の関係数}}{\text{システムが同定した人物表現間の共参照の関係数}}$$

$$再現率 = \frac{\text{正しく同定できた人物表現間の共参照の関係数}}{\text{人手で同定した人物表現間の共参照の関係数}}$$

表 12 人物表現間の共参照解析の結果

ノードの結合の種類	適合率	再現率	F 値
Dependency	0.857(60/70)	0.732(60/82)	0.789
Text-Sim	0.955(64/67)	0.821(64/78)	0.883
S-relation	0.917(33/36)	0.600(33/55)	0.725
Pronominal	0.845(49/58)	0.710(49/69)	0.772
Total	0.892(206/231)	0.725(206/284)	0.800

共参照解析処理の精度は，共参照の対象となる人物表現の抽出精度に大きく依存する．そのため，人手で付与した人物表現を入力としたときの精度を調べてみた．その結果を**表** 13 に示す．その結果，適合率が 0.892 から 0.928，再現率は 0.725 から 0.859 へと大きく上昇する．再現率が上昇する理由は，人物表現抽出処理で抽出できていない人物表現への共参照の関係が取得できるためである．また，適合率が上昇する理由についても，人物表現抽出処理では適合率 0.907 であり 1 割程度の誤りがあるため，人手で作成した人物表現では，それらの誤りが除外されるため，共参照解析の結果が改善するものと考えられる．

表 13 人物表現間の共参照解析の結果（人物表現を与えた場合）

ノードの結合の種類	適合率	再現率	F 値
Dependency	0.952(79/83)	0.963(79/82)	0.958
Text-Sim	0.960(72/75)	0.923(72/78)	0.941
S-relation	0.953(41/43)	0.745(41/55)	0.837
Pronominal	0.839(52/62)	0.754(52/69)	0.794
Total	0.928(244/263)	0.859(244/284)	0.892

今回の実験で得られた共参照解析処理の結果をその種類別に検証すると，S-relation の再現率が，他の共参照に比べ低い結果となっていることがわかる．その理由として，開発データの規模が十分でないために本来は共参照とすべき関係の表現を予め作成した S-relation の語彙として登録できていなかったことが考えられる．そのため，次章で提案する大規模な含意パターン対による言い換えを利用するなどして，現在の S-relation の語彙として登録されている関係表現を拡張することが必要である．また S-relation では，後述する関係抽出の結果を利用しているために関係抽出の精度に大きく依存するため，関係抽出の再現率が十分でないために共参照が特定できなかった場合もあったと考えられる．

代名詞の共参照である Pronominal では，センタリング理論による格の優先度よりも，意味の近い先行詞を取るべきである事例があった．例えば，次の番組概要の文章では，本手法では，照応詞「その選手」の参照先を，前文の主題である「エディ」としてしまう．「ボクサー」を照応先とするためには，センタリングよりも意味を優先する必要がある．

エディは，プロモーターから新人ボクサーを売り出すための依頼を受ける．その選手は・・・・

このような先行詞の選択の誤りは，「選手」の概念が「ボクサー」を含むという事前知識がないと対応は難しい．この種の問題を解決するためには，同義語や上位下位の概念などによる意味情報を素性として取り入れる必要がある．

また，共参照の対象を人物表現に限定しているために，その精度は人物表現抽出の結果に左右される．本実験の結果でも，人物表現抽出の結果を用いると，単体の評価に比べ再現率が 10%以上大きく低下している．これは，関係抽出と同様に人物表現抽出の再現率が低いことが原因である．特に Dependency と S-relation の精度が低い要因は，人物表現（TITLE，G-PERSON）の抽出精度が低かったことに起因している．

本研究の共参照解析処理の比較のため，Iida らが [Iida11]を基に実装したツール Syncha[12]を用いて，本タスクと同条件で結果を取得した．その結果，適合率 0.938（30/32），再現率が 0.106（30/284）の結果が得られた．Syncha により取得できた共参照は，表層的に類似しているものだけであり，本手法のテキストの類似性に基づく Text-Sim による適合率 0.955 と同程度の適合率であった．また，Syncha では，以下のような文章において，２カ所に出現する「彼」を共参照としていた．例では，前方の「彼」は「サニー・フーパー」を指し，後方の「彼」は「若手スタントマン」を指す．本手法では，このような場合でも，それぞれの他の共参照と矛盾が生じるため，２つの「彼」は共参照とはならない．

バート・レイノルズ演じるベテラン・スタントマン，サニー・フーパーは，一流ともてはやされていたものの，年齢とともに衰えを感じていた．そんな彼の前に現れたのが，ジャン・マイケル・ビンセント演じる若手スタントマン．彼の出現にあせりながらも，谷を車で越える大アクション・シーンに挑む…．

Syncha の再現率 0.106 と低かった理由は，ツールが用いているモデルが新聞記事に基づき学習されているため，番組概要に出現する固有表現の抽出精度が十分でなかったこと，「彼女」や「彼」などの代名詞の共参照をうまく取得できなかったことが考えられる．また，キャッシュモデルに基づく手法 [Iida09] による共参照の対象となる名詞の制限が効きすぎていることも考えられる．

3.8.3 人物表現間の関係の評価

生成した人間関係グラフにおいて，そのエッジにあたる人物間の関係が正しく取得できているかを評価した．実験で利用した正解データは，評価者がテストデータの番組概

[12] https://www.cl.cs.titech.ac.jp/~ryu-i/syncha

要の２つの人物表現の間に何らかの関係があると判断した表現を抜き出したものである．評価では，人物間の関係とその２つの人物表現がすべて合っている場合のみ正解とした．ただし，文字列が完全に一致していない場合でも評定者が同じ関係であると判断すれば正解とした．例えば，人手作成の正解が「派手な女性関係にうんざり」で，システムの抽出した「女性関係にうんざり」であっても，評価者が２つの人物表現間の関係と判断できれば，正解と評価されている．再現率，適合率の定義は以下の通りである．

$$適合率 = \frac{\text{正しく抽出できた人物表現間の関係数}}{\text{システムが抽出した人物表現間の関係数}}$$

$$再現率 = \frac{\text{正しく抽出できた人物表現間の関係数}}{\text{人手で抽出した人物表現間の関係数}}$$

評価結果は，適合率 0.722，再現率 0.694，F 値 0.708 が得られた．エラー解析の結果，人物表現の誤りに起因して，その関係が取得できていない例が多くあった．人手で付与した人物表現の結果を入力した場合の精度は，適合率 0.742，再現率 0.779，　F 値 0.760 が得られた．人物表現抽出の結果を利用した場合では，得られた関係は 179 個であったのに対して，人手で付与した人物表現の結果を入力した場合は，201 個の関係が取得できている．この差は関係抽出処理が人物表現抽出処理の結果に依存しているため，人物表現抽出ができていないノード間の関係が取得できていなかったことを示している．また，適合率の差は，人物表現抽出で誤って抽出された人物表現を含む関係が出力されていたと考えられる．

人物表現抽出の再現率は 0.903 であり，全人物表現 710 個のうち，69 個の人物の抽出ができていない．また，適合率は 0.912 であり，58 個の抽出の誤りがある．人物表現抽出の評価でも述べたとおり，TITLE や G-PERSON は改善の余地がまだ残されているため，これにより，関係抽出の結果も改善できることが期待できる．

表 14 人物表現間の関係抽出の結果

	適合率	再現率	F 値
人物表現抽出の結果を利用した場合	0.722(179/248)	0.694(179/258)	0.708
関係抽出単体	0.742(201/271)	0.779(201/258)	0.760

その他の適合率の低下の要因として，複文の分割処理の誤りに起因しているケースがあげられる．複文の分割処理では，分割後の文で，主語や主題のある文から，主語や主題が省略されている文に対して，主語や主題を補完する処理を行っているが，現アルゴリズムでは，センタリング理論に基づく格情報のみを利用しており，人物を主語としない動詞に対してもゼロ主語を補完し，結果的に誤った人物表現間の関係を抽出する場合があった．誤った補完を改善するためには，格フレーム [Kawahara06] などを利用し，人物表現が取りうる述語とその格情報などの知識を素性に導入する必要がある．さらに，ゼロ主語の補完以外にも，所有格や目的格が省略されている場合もあり，関係抽出の再現率向上にはゼロ代名詞全般の補完を行う必要がある．

本手法による関係グラフの生成により，ノードに含まれる表現間の関係を取得することで，単に構文解析結果の構文木から関係を取得するだけでは得られない人名間の関係を取得することができる．例えば，「彼の派手な女性関係にうんざりした彼女は，国賓の王子との浮気を宣言する．」という文の構文解析結果からは，人間関係（彼，彼女，女性関係にうんざりする）の３つ組が得られるが，関係グラフの結果(**図 17**)では「彼」のノードには{ミッシェル，官房長}，「彼女」のノードには{ブリジット，令嬢}が含まれている．それぞれのノードで他の表現と置き換えることで，**表 9** では得られていない３つ組を新たに取得することができる．**表 15** に（彼，彼女，女性関係にうんざり）の３つ組をもとに，関係グラフの結果から新たに得られる関係を示す．

表 15 関係グラフから得られる人物表現間の関係

人物表現１	人物表現２	関係
ミッシェル	彼女	女性関係にうんざりする
官房長	彼女	女性関係にうんざりする
彼	ブリジット	女性関係にうんざりする
ミッシェル	ブリジット	女性関係にうんざりする
官房長	ブリジット	女性関係にうんざりする
彼	令嬢	女性関係にうんざりする
ミッシェル	令嬢	女性関係にうんざりする
官房長	令嬢	女性関係にうんざりする

この新しい３つ組の獲得の結果から，代名詞や職業名などの人名ではない人物表現を含む人物表現間の関係をもとに，（ミッシェル，ブリジット，女性関係にうんざりする）という人名を用いた３つ組を取得できることがわかる．

実験で用いた90番組のうち，番組概要に人名が２種類以上現れる24番組を対象に，構文解析結果から直接取得した人名間の関係と，関係グラフを基に得られた人名間の関係を比較した．その結果，構文解析結果から直接得られる人名間の関係数は21個であったのに対して，関係グラフの結果からは36個の人名の関係（うち１つの関係は誤り）を取得できた．これは，番組概要で２つの人名が同一の文に現れていなかったり，人名の代わりに代名詞や一般人物表現，関係人物表現など人名以外の人物表現が用いられるためである．関係グラフの結果を用いることで，人名と人名以外の人物表現をひも付けることができ，より多くの人名間の関係の取得が可能となる．このように人名間の関係を網羅的に取得できることで，例えば，連続ドラマ等で，ある特定の登場人物間の関係の推移を取得することなどが可能となる．

3.8.4 人物表現間の関係を利用した番組検索への応用の検討

番組概要から関係グラフを生成することにより，人物間の関係性を考慮した番組検索を行うことが期待できる．グラフ構造のデータ間の類似度はその処理時間が膨大となるため，例えば，２章で提案した番組検索手法に，関係グラフから取得した関係表現に重みを付与することにより，エンティティ間の関係性に着目した関連番組検索の結果を出力することが考えられる．また，番組検索において関係グラフのノードつまり人物表現をマスクし関係表現のみで検索を行えば，同じ人間関係を持つドラマ等を検索することができる．

関係抽出で得られる表現の種類は多岐に渡り，この関係をそのまま番組検索に適用したのでは，検索結果が十分に得られない場合がある．例えば，クエリとして指定した「AとBが別れる」という関係は，ほかの番組の番組概要には「AとBが離婚する」「AからBが離れる」などの表現で記述されているかもしれない．そのため，クエリに用いる人物間の関係を，含意パターンを利用するなどして別の様々な表現に言い換えることを必要である．含意パターンを利用したエンティティ間の関係の言い換えについては次章で詳しく検討することとする．

3.9 関係グラフの応用

番組概要から取得した人物間の関係を利用した応用として，番組内容の可視化を行うシステムを開発した．開発したシステムでは，番組概要から作成した関係グラフの結果から登場人物の人間関係相関図を生成する（**図 18**）．システムで図示された人間関係では，ノードの持つ格情報を利用して方向性を付与し，その関係を矢印で示している．基本的に，主題の「ハ」やガ格を持つ人物表現を基点に他の人物表現へ関係の方向性が生成される．また，**図 19** に示す音声クラスタリングと字幕に付与されている話者情報を

取得する特許技術 [後藤 12] を利用することで，その登場人物の顔画像と登場人物名を取得し，その人間関係相関図に付与することも可能である．

図 18 関係グラフから生成したドラマの相関図

開発したシステムでは，この相関図をユーザインタフェースとして利用し，特定の人物が登場するシーンの検索を行う．相関図上で人物間の関係を指定すると，その関係を持つ人物達が共に登場しているシーンが再生される．相関図の人間関係とシーンの関連付けには，[後藤 12] により得られた登場人物の顔画像とその人物名から登場しているシーンを特定することができるため，この情報を利用している．

NHK 放送技術研究所オープンハウス「技研公開 2008」で，NHK 連続ドラマ「瞳」の番組概要と本編から人間関係相関図を生成した本デモシステムの展示を行った．来場者から「直感的にストーリーを把握できる」「出演者と役がよくわかる」「複数の相関図を見ることで，主人公とほかの人の関係の変化がわかる」などの感想を得ている．

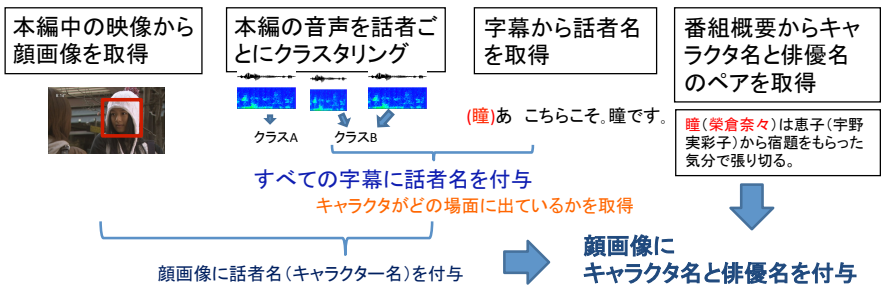


図 19 登場人物の顔画像と出演シーンの特定

このシステムを用いることで，連続ドラマなどで，過去の回を見逃したとしても，相関図を見るだけで，簡単にストーリーの把握を行うことができる．また，相関図だけではわかりにくい場合は，番組オンデマンドに対して過去の番組をリクエストして特定のシーン，例えば，主人公とその恋人がでてきているシーンばかりを集めて視聴することも可能となる．また人名間の関係を網羅的に取得することにより，歴史ドラマなどの時間経過で関係が変化するような場合にも，人名の間の関係を比較することで，人物間の関係の推移データを作成することができる．

3.10 まとめ

番組の内容を考慮した関連番組検索の実現のため，番組概要から人物表現間の関係を取得する手法を検討した．番組概要から人物を示す表現を抽出し，人物表現間の共参照解析処理，人物表現間の関係抽出処理の結果を統合することで，グラフ構造をもつ人物表現間の関係の情報を示す関係グラフを生成する手法を提案した．関係グラフ生成の評価実験の結果，関係グラフのノード抽出の精度で適合率 0.912，再現率 0.903，ノード結合の精度で適合率 0.892，再現率 0.725 が得られ，人物表現のノード間の関係取得で適合率 0.722，再現率 0.694 となる結果が得られた．エラー解析の結果，人物間の関係抽出を向上させるためには，主に，(1) 人物表現の抽出精度の向上，(2) 構文解析の誤り削減のための複文の分割処理の改善，(3) 省略されているゼロ主語の補完精度の向上などの課題が確認できた．また，関係グラフを用いることで，構文解析結果から直接得ることができない人物表現間の関係を抽出できることを示し，関係を考慮した関連番組検索への応用の可能性について述べた．また得られた関係グラフを用いて，番組の内容を可視化する登場人物の相関図を生成できることを示した．

本章で提案した人物表現間の関係取得手法を，時間の情報を持つ新聞やニュース原稿などの大規模なコーパスに適用することにより，時間的に特定の人の関係がどのように変化するかを取得することができる．また，本研究では，対象とするエンティティを人物表現に絞ったが，企業などの組織間の関係性の変化の分析などに応用することも可能である．また，文章には直接記述されていない人物表現とその関係を共参照解析により取得することで，質問応答などの知識としても利用することが期待できる．

第４章 質問応答に基づくマイクロブログからの情報取得

4.1 はじめに

放送は，これまで放送局から電波を通して広範囲にコンテンツを発信するものであった．インターネットに接続できる携帯端末やスマートフォン，Twitter や facebook などのマイクロブログの登場により，個人からでもリアルタイムに情報を発信できるようになった．マイクロブログは，個人発のコンテンツを広範囲の不特定多数の人々に配信できるため，一種の放送メディアの側面を持ち始めている．東日本大震災においても，Twitter などのマイクロブログが，既存メディアでは伝えられていない情報を多く発信していたとの報告がある．例えば，被災地で不足しているものとして，平時では予想が困難な物資の情報が様々な状況で個人から発信されていることが判明している．

一方，既存メディアが伝えた情報と現地の実際の状況との間にミスマッチが生じていたという事例が報告されている．例えば，テレビや新聞が報じた「被災地で防寒着が不足している」という情報に呼応して，多くの善意の人から防寒着の上着が大量に現地に送られたが，津波被害のなか，泥水の中で復旧作業をする必要のあった人々がより切実に求めていたのは，防寒のズボンであったという．また別の例では，全国から支援物資として届けられた多くの衣類はどれも通常サイズのものばかりで，4L サイズなどの大きな衣類が必要な人が一月以上も被災時の衣類を着続ける必要があった．しかしながら，これらの局所的な情報まで網羅的に既存メディアが伝えることは，その限られたリソースを考えると実現は難しい．

このような問題を解決するために，マイクロブログのコンテンツを局所的および詳細な情報の取得に用いて，既存メディアが伝えきれない情報を補完することが考えられる．ハイブリッドキャストの登場により，テレビを見ながらその補完情報をマイクロブログから取得するアプリケーションをテレビ受信機上で動作させることも可能となる．例えば，ニュース番組では十分に伝えられない状況の詳細や，情報番組で放送時間内に伝えきれていないローカル情報等を取得することもできるようになる．これにより，番組制

作者の取材では明らかになっていない，現地の一部の人しか知らない有用な情報を取得できるかもしれない．

上記で述べたことを実現するために問題となるのは，マイクロブログにおいて多くの人から発信される大量の情報の全体像を把握することが容易ではないということである．事実，東日本大震災のときにも，マイクロブログに投稿された被災地からの切実な要望や貴重な情報が，政府，地方自治体，NPO などの救援団体に必ずしも届かず，救援活動や復興支援が最大限の効率で進展しなかった事例も報告されている．要請や要望に対して何の反応もなかった Twitter への書き込み（tweet）も数多く存在した．

こうした状況に対応するため，東日本大震災の際，自然言語処理を用いて Twitter 上の安否情報を整理することを目指した「ANPI_NLP」の取り組みが行われたが，開発の速度や多数のボランティアの組織化に課題があったことが報告されている [Neubig11]．このように，大量の情報を整理する状況が発生してから，新たに情報を取得し整理する技術を開発するのは困難である．災害時に限らず，マイクロブログで発信される大量の情報を整理し，上述した想定外の要望も含めて，必要な情報を必要な人に把握が容易なフォーマットで届ける技術の開発を予め行っておくことが重要である．

以上のような点に鑑みて，本章では，既存メディアからの情報を補完する手段として，ユーザが必要な情報を網羅的にマイクロブログから取得する手法を提案する．その手法として，必要としている情報が何であるのかを質問という形式により明確にシステムに伝えられ，その質問に応じて与えられたデータベースやコーパスから回答を探し出すことができる質問応答技術を用いる．例えば，ニュース番組で報道されたイベントの詳細な情報を聞く質問をすれば，現地から投稿されたマイクロブログの書き込みから様々な種類の回答を取得することで，より的確にその状況を把握することができるようになる．多様な質問に回答できる質問応答手法を開発することによって，tweet 等のテキストデータが人手での処理が不可能な量となっても，そこに現れる多様で大量の事象を意味的観点から分類・抽出を可能にし，さらに回答の地図上への表示や，回答に時間的な制約をかけることのできるインタフェースも合わせて提供することにより俯瞰的把握を容易にする．

本章の構成は以下の通りである．まず，次節において本論文で利用する質問応答に関する関連技術について述べ，4.3 節で提案する質問応答に基づくマイクロブログからの情報取得手法について述べる．次に 4.4 節で質問応答のための言い換えパターン作成，4.5 節で回答を高速に取得するための回答インデックス作成，4.6 節で tweet の不足する情報を補う地名補完処理，4.7 節で質問応答処理の各処理について説明する．4.8 節において，東日本大震災時の実際の tweet を対象としたシステムの評価について報告する．最後に 4.9 節にて本章の結論を述べる．

4.2 関連研究

膨大な情報を持つマイクロブログから必要な情報を取得するためには，システムにどのような種類の情報が欲しいかを明確に伝えられ，大量の文章から回答のみを取得して出力することができる質問応答技術の利用は有力な解決手段である．

質問応答システムとしては，IBM社のWatson がクイズ番組の人間のチャンピオンに圧勝し一躍有名になった [Ferrucci10]．Watsonは，Wikipediaを含む大規模な辞書，辞典や台本などJeopardyというクイズ番組の分野に関連する確かな知識を予め選別し，データベース化している．このシステムは一種のエキスパートシステムと言えるため，専門性の高い医療分野に応用が進められている．

また，日本で実用化された質問応答システムのひとつに，「しゃべってコンシェル」と呼ばれるシステムがある [吉村12]．しゃべってコンシェルは，いわるゆホテルのコンシェルジェのように，ユーザが携帯電話を通してシステムに依頼をすると，対応する結果を出力してくれるサービスである．意図解析エンジンと呼ばれる技術を用いて，質問が端末機能や専門の検索エンジンのどのサービスで回答できるかを選択する．天気予報やニュース，レシピ検索などの専用のサービスを用いて回答できない質問の場合，質問応答技術を利用して，回答を抽出する．知識検索エンジンは，予め質問と回答に関する情報を整理して格納したデータベースを利用したDB型と，Webなどを検索し，その結果から回答を検索する検索型により実現されている．検索型では，質問から取得した回答のタイプを固有表現タイプとして取得し，同一の固有表現を検索結果から回答として取得している．

テレビの番組に対する質問応答として，Gotoらが提案したTV Agentシステムがある [Goto06]．このシステムでは，ユーザから得られた質問を複数の知識ごとにエージェントを配置し，各エージェントが受け持つ知識の信頼性の重みを付与し，最終的な回答を出力するというものである．各エージェントが回答を取得する対象として，出演者やジャンルなどの放送波に重畳されている情報から作成した構造化知識と，インターネットや新聞，その他の百科事典などの非構造化知識を用いている．放送に含まれている情報は，正確な反面，その知識の量が限られているため，幅広い質問には対応できない．一方で，インターネットに掲載されている情報などの非構造化知識は巨大で網羅性はある反面，その情報は確かとは言えないため，これらを統合することで網羅性と回答精度を両立させている．

日本語における質問応答の進展には，国立情報学研究所（NII）が主催する評価型ワークショップであるNTCIRが大きな役割を果たしている [森08]．NTCIR-3，NTCIR-4，NTCIR-5では，QACと呼ばれる質問応答タスクが，新聞２年分の記事を対象に実施され

ている．例えば，固有表現や数値表現を答える質問応答タスク（QAC-1, 2），回答の数が複数となる質問応答タスク（QAC-1, 2），複数の連続した質問群に回答するシリーズ質問タスク（QAC-1, 2, 3）などを行っており，多くのシステムが参加している．多くの参加システムのアプローチは，質問に含まれるキーワードで文書を検索し，その結果から固有表現抽出器の結果を特定し，その固有表現の回答らしさにより，回答をランキングするというものであった．日本語のFactoid型のタスクは2005年に終了しており，その後，Why型，How型，定義を問うnon-factoid型の質問応答や，多言語質問応答（CLQA, ACLIA）などの応用のタスクが行われている．

上記に紹介した既存のFactoid型の質問応答システムは，回答が一意に定まる質問を対象としている．マイクロブログから情報を抽出する場合，その回答はあらかじめ定まっておらず，いくつあるかさえ，わからない場合がある．例えば，「被災地で不足しているものは何ですか」などの質問では，投稿者の置かれた状況によってその回答は異なり，「ミルク」から「手話通訳」まで様々なものが回答となりえる．時間や場所でフィルタせずに，マイクロブログ上の全データを対象とした場合，回答の種類は，一つの質問に対して数百から数千以上に及ぶこともある．これらの回答は常に固有表現で表される訳ではないため，従来のFactoid型の質問応答システムのように，質問タイプから回答の固有表現を決定し，情報検索の結果から同じ固有表現を回答候補として抽出するというアプローチでは得られる回答が限られる．また，回答が記載されている文書が少数だった場合でも，特定の人々には非常に重要な情報であるかもしれないため，質問のキーワードによる全文検索結果の上位のみから回答を抽出する方法も取りづらい．そのため，本論文では，固有表現抽出に基づくアプローチはとらず，大規模な含意パターンを利用した質問文の言い換えにより，様々なマイクロブログ文書に書かれている回答を網羅的に取得することにより，ロングテールに存在する情報をも獲得することを目指す．

4.3 質問応答に基づくマイクロブログからの情報取得

マイクロブログの大量の情報から，ユーザが必要としている情報を網羅的に把握するためには，対象のマイクロブログをフィルタし投稿記事を絞り込むだけでは不十分である．マイクロブログの投稿記事一つ一つの情報量は少ないため，その対象が少なければ容易に情報を取得することができる．しかし，マイクロブログに様々な人から時々刻々と投稿される膨大な情報を網羅的に取得し，その全体像を把握することは難しい．その点で，質問応答技術はユーザの要求を的確に自然言語でシステムに伝えられ，欲している情報を最小限の表現で返すことができるため，このようなタスクには有効であると考えられる．そこで，本研究では，質問応答技術を用いて，ユーザが知りたい情報を聞く質問からその要求内容を特定し，対応する回答を大量のマイクロブログ上の情報から取

得する手法を検討する. 質問応答に基づくマイクロブログからの情報取得の処理の流れを**図** 20 に示す．処理内容は大きく分けて，(1) 言い換えパターン作成処理，(2) 回答インデックス作成処理，(3) 地名補完処理，(4) 質問応答処理から構成される．

(1) 言い換えパターン作成処理は，大規模なコーパスから含意関係にあるパターン（含意パターン）を自動的に抽出し, 含意パターンデータベースを作成する処理である．含意パターンとは，あるパターンが成り立った場合に成り立つパターンのことである．言い換えパターン作成は，マイクロブログから逐次獲得したパターンを利用して含意パターンデータベースの更新を行うことも可能であるが，より大規模なコーパスである Web 6 億文書などから含意パターンを一度取得するとその語彙数は大きく変わらないため, その実行はオフラインで行う. 4.4 節でその詳細について述べる．
(2) 回答インデックス作成処理は，マイクロブログの書き込みを継続的に取得し，回答を高速に取得するための回答インデックスをリアルタイムに作成する処理である．回答インデックスでは，マイクロブログを構文解析の結果を利用して，回答を取得に必要な名詞とパターンの組み合わせを格納している. 4.5 節でその詳細について述べる．
(3) 地名補完処理は, 回答インデックス作成の際に問題となる地名の省略等に対応するために, マイクロブログの書き込みの構文解析結果に対して重要な地名や場所名を補完する処理である．4.6 節でその詳細について述べる．
(4) 質問応答処理は，次の３つの部分処理で構成されている．その一つは，ユーザからの質問文を解析し得られるパターンと, (1)で作成された含意パターンデータベースを用いて，回答検索を検索するための多彩な表現のクエリを生成する質問解析処理である．もう一つの処理は，質問解析処理が出力したクエリを用いて，(2)で作成された回答インデックスを検索し，回答集合を出力する回答検索処理である．最後の入出力処理は，ユーザからの質問を質問解析処理に伝え，回答検索処理からの回答リストを整理して表示するための処理を行う．入出力処理には，大量の回答をユーザに俯瞰的に見せるための回答提示処理も含まれている. 4.7 節でその詳細について述べる．

提案手法を用いることで，「宮城県で何が不足しているか」のような自然言語の質問に対して，マイクロブログの大量の書き込みから「ズボン」「ミルク」「手話通訳」などの既存メディアでは取得できない局所的な情報を取得できるようになる．なお，本手法は対象とする情報としてマイクロブログを主たる情報源とするが，掲示板や一般の Web 文書などへの適用も可能である．

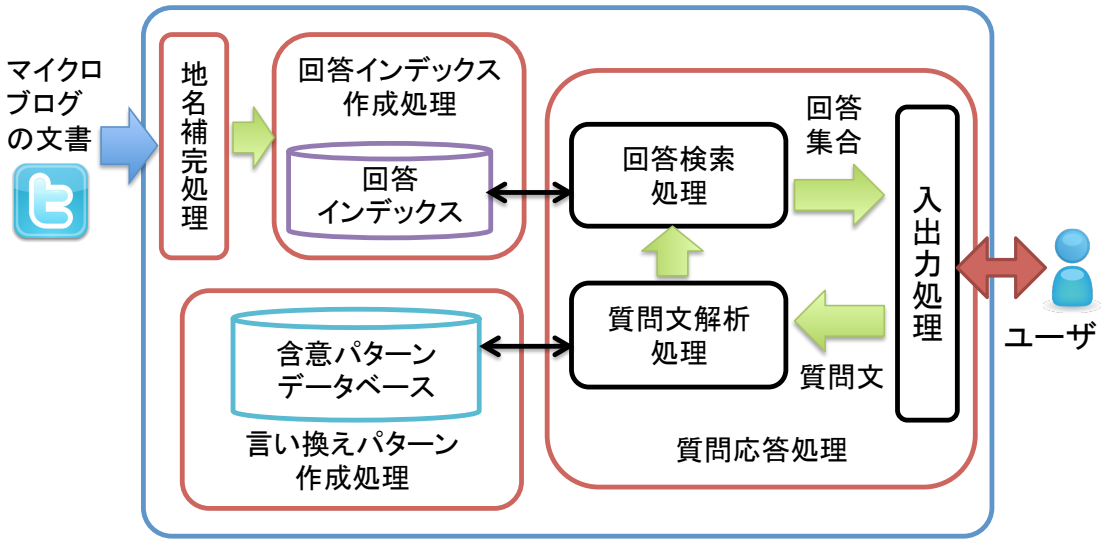


図 20 質問応答に基づくマイクロブログからの情報取得の概要

4.4 質問応答のための言い換えパターン作成

ユーザから入力される質問文から取得したパターンを，同じ意味を持つ様々な表現に言い換えることができれば，幅広い回答を獲得することが可能となる．Negriらは人手で作成した含意パターンデータベースを用いて，ユーザから入力された質問文と近いパターンを獲得し，その含意パターンにより回答を取得するシステムを構築している[Negri09]．しかし，システムは，映画や展示会などの文化的なイベントに関するコーパスに出現する，人手で抽出した449のパターンのみを利用しており，その対象とするドメインや含意パターンの言い換えは限定的である．

本研究では，Web 6 億文書からなる大規模なコーパス [Shinzato11] から含意パターン対を獲得し，それらを利用してオープンドメインで幅広い言い換えを可能とする．含意パターン対とは，例えば，あるパターンの「XからYまで移動する」とそれを含意する「XからYまで歩く」のようなパターンのペアのことであるが，含意が成立するための名詞句X，Yにある制約等を考慮するといくつか種類が考えられる．ここでは，クラス依存バイナリーパターン，クラス非依存バイナリーパターン，ユーナリーパターンという三種類の構文パターンの含意パターン獲得について説明する．

4.4.1 クラス依存バイナリーパターン獲得

クラス依存パターンとは，パターン中の変数に対応する名詞の意味クラスに制約を掛けた構文パターンであり，パターンに２つの変数が含まれているものを特にクラス依存

バイナリーパターンと呼ぶ. 構文パターンにクラス制約を掛けることでパターンの多義性が解消できる．例えば，「ＹのためのＸ」という構文パターンは「Ｙ:病名のためのＸ:薬品」のように，Ｙが病名，Ｘが薬品となる意味クラスの単語の場合は，ＸとＹの治療関係と呼べる関係を表し，上記のパターン「Ｘ:薬品で Y:病名が治る」の含意パターンと見なせる．一方，「Ｘ:作業のためのＹ:道具」」の場合は手段または道具という意味的関係を表現する. このようにして構文パターンと共起する単語を特定の意味クラスに限定することで，構文パターンの曖昧性が大きく減らされ，高頻度で曖昧なパターンが活用可能になり，より大量の回答を獲得できる [De Saeger09]．意味クラスには，Kazama らが提案した単語クラスタリング法 [Kazama08] によって自動獲得した結果を用いる．この手法では大規模 Web コーパスから得られる名詞と動詞の係り受け関係の統計データを用いて，名詞の隠れクラスへの事後確率の分布を求める．ある名詞の所属確率が 0.2 以上の隠れクラスを，その名詞の意味クラスとする．本研究では，名詞 100 万個を 500 クラスに分類したクラスタリング結果を用いる．

クラス依存の含意パターンの認識には Kloetzer らが提案したクラス依存パターン間の教師付きの含意獲得手法を用いる [Kloetzer12]．含意パターンを認識する SVM 分類器は主に次の 3 種類の手がかりを用いる．

1. パターンの表層的素性（表層／構造を考慮した素性）
   これらの素性は，表層上似ているパターンは含意関係にある可能性が高いという前提で，パターンに含まれる形態素，内容語，構文木の部分木などの bag of words 表現を基に計算した様々な類似尺度から成る．
2. 分布類似度に基づいた素性
   ある構文パターンとその含意パターンの候補に関しては, 6 億ページの日本語 Web 文書からパターンの変数に当てはまる名詞句対を検出し，それらの名詞句対の相対的なオーバーラップを計算する．例えば，「ＸでＹを提供する」と「ＸでＹを配っている」という 2 つのパターンはＸとＹの変数に頻出する共通の単語対（例えば，「石巻市，救援物資」）が多いほど，これらの構文パターンがお互いの言い換え表現となっている可能性が高いと考えられる．類似する文脈に出現する語は類似する意味をもつことは，分布仮説 [Harris54]として言語学ではよく知られている．これらの素性はクラス依存のパターンの意味クラスに属する単語対に基づいて計算した類似尺度から成る．

3.　言語資源に基づいた素性

高度言語融合フォーラム ALAGIN[13] で公開された動詞含意関係データベース（ALAGIN リソース A-2), 日本語異表記対データベース（ALAGIN リソース A-7), 基本的意味関係の事例ベース（ALAGIN リソース A-9）と日本語形態素解析器 JUMAN の辞書から得られた異表記と反対語データを言語資源として参照し, 両パターンに含まれる内容語が同義語，異表記，含意関係，対義関係にある場合など，これらの言語資源に含まれる意味的関係にある時にその情報を素性に加える．

学習データとして 51,900 のパターン対を用いて，SVM での学習には 2 次の多項式カーネルを用いた．**図 21** は，学習データとは異なる 5,338 のパターン対を評価した本分類器から得られるクラス依存パターン含意の認識精度である．

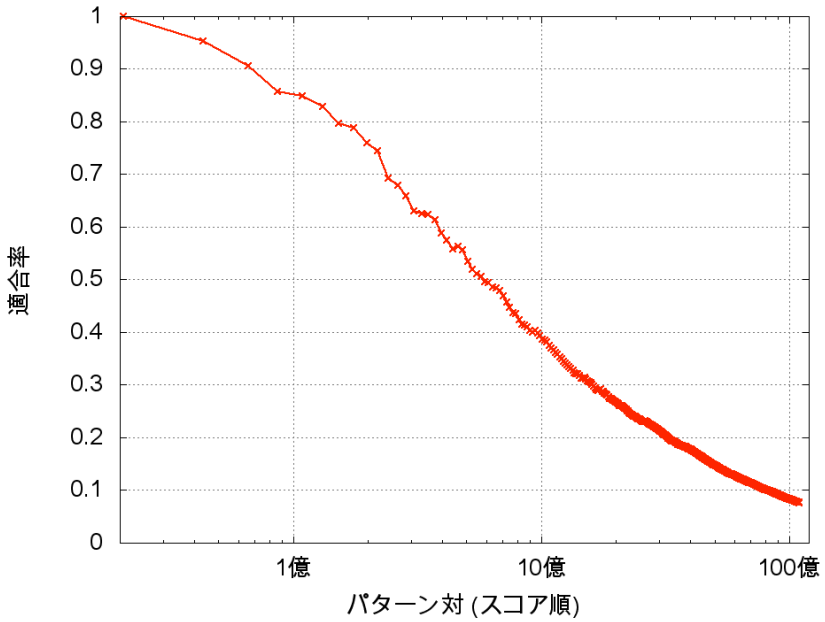


図 21 構文パターン間の含意認識の適合率

上述した条件において，この手法の上位 1 億対（データサンプル数 49）では約 85%の適合率を示し，上位 2.37 億にて約 70%の適合率を保持している．本手法で利用される含意パターンデータベースは, 後述する方法により質問文から得られる可能性のある構文パターンの含意パターンを SVM スコアが高いものに絞って格納しており，回答検

[13] http://alagin.jp/

索に用いる含意パターンの適合率は**図** 21 に示される上位の適合率に相当するものと考えられる．

質問文から得られる構文パターンの言い換えに利用する含意パターンデータベースを構築するため，500 種類の意味クラスの任意のペアのうちで，同じ名詞句対を異なり数で 3 以上共有するパターン対を考える．こうしたパターン対の総数は 108 億ペア存在し，そのすべてに対して，分類器を適用して SVM スコアを求める．ついで，SVM スコアが計算されたパターン対の内，以下の手続きで最終的な含意パターンデータベースを構築する．まず，上述のパターン対に含まれるパターンを「含意されるパターン」P として一つ選択し，SVM スコアが 0 以上のパターンを「含意するパターン」Q としてスコア上位から順に取得する．Q が 500 個を超えた場合は，スコア上位 500 個のみを P と対にして含意パターンデータベースに格納する．この操作を 108 億個のパターン対に含まれるパターン各々を P と仮定して繰り返す．なお，上位 500 個という数値は決定的なものではなく，パラメータのひとつであるが，求める性能と応答速度のトレードオフによって決まる．現在の 500 という数値は，さまざまな質問の言い換えを行い，経験的に決めたものである．

また，パターン P と Q の含意を考えた場合，「Q が P を含意する」と「P が Q を含意する」という含意の方向性があり，当然これらの含意パターン対の SVM スコアは異なる．そこで，通常の「Q が P を含意する」場合のスコアと，逆向きの「P が Q を含意する」場合のスコアが両方向ともに 0 以上のパターン対のみにデータベースへの登録を限定する．これは片方向の論理的含意関係が成立しているものの，あまりに意味的にかけ離れているパターン対で質問文の言い換えを行なうことを防ぐためである．例外として，取得できたペアが 500 個未満の場合には，Q のスコアが 0 以下であって，同じ内容語（動詞，名詞または形容詞）を持つパターンペアをスコアの高いものから順に取得し，登録数が 500 個になるまでデータベースに登録する．

4.4.2 クラス非依存バイナリーパターン獲得

クラス依存バイナリーパターンは，特定の意味クラスの組み合わせを持つ含意表現を取得することができる．一方，特定の意味クラスに依存せず，広い文脈で含意表現として利用できるバイナリーパターン（クラス非依存バイナリーパターン）も回答の抽出には役立つはずである．そのため，名詞句に何らの意味的制約が加えられていないバイナリーパターンを対象として，クラス非依存バイナリーパターンを取得する．多くの意味クラス対で含意関係が得られるパターン対はロバストで一般的な言い換え表現であるという前提のもと，クラス依存バイナリーパターン間の各意味クラス対での SVM スコアを平均したパターン対のデータベースを作成する．この際，1 つの意味クラス対とし

か共起しないパターンは一般性がないと考えて除外する. あるパターンのクラス非依存バイナリーパターンは, クラス依存バイナリーパターン獲得のケースと同様のアルゴリズムで選別する．また，パターン対の含意の方向性についても，データベースへの登録の制限に利用されている.

4.4.3 ユーナリーパターン獲得

マイクロブログから得られるテキストはインフォーマルな書き方で知られている. 特に Twitter の場合では，tweet が 140 文字以内という制限があり，必要最低限の情報しか含まない tweet が多い．そのため，2 つの名詞句の存在を前提とするクラス依存バイナリーパターンやクラス非依存バイナリーパターンがうまく取得できない場合がある.

このような問題に対処するために，4.4.2 で説明したクラス非依存バイナリーパターンを一つの名詞句の存在を前提とするユーナリーパターンに分割する．例えば，「ＸがＹで孤立する」というバイナリーパターンは，構成する係り受け関係「Ｘが孤立する」と「Ｙで孤立する」に分割することができる.

ユーナリーパターンの含意パターンデータベースは，次のように作成する．既に説明したクラス非依存バイナリーパターンの含意パターンデータベースを入力とし, それらのバイナリーパターン対を分割し, 変数毎にユーナリーパターンの候補ペアを生成する. 例えば，（「ＸがＹで孤立する」，「ＹではＸに連絡できない」）というクラス非依存バイナリーパターン対から（「Ｘが孤立する」，「Ｘに連絡できない」）と（「Ｙで孤立する」，「Ｙでは連絡できない」）という 2 つのユーナリーパターン対を含意候補として生成する．このユーナリーパターン対の含意スコアは，クラス非依存バイナリーパターン獲得の際と同様に, その生成元のクラス非依存バイナリーパターン対のスコアの平均とする．ただし，生成元のバイナリーパターン対が 1 つしかないユーナリーパターン対は一般性に欠けていると考えて除外する. さらに, クラス依存バイナリーパターン, 非依存バイナリーパターンと同様に含意の方向性について考慮し，両方向のスコアが 0 以上のパターン対のみをデータベースに登録する.

以上の方法で作成したユーナリーパターン対は, それがもたらされたクラス非依存パターン対のスコアを平均した値をスコアとして持っているが, パターンに含まれる用言相当の表現と変数との関係を考慮していないため，信頼性を欠く場合がある．そこで, 次の 2 つの方法で，ユーナリーパターン対をクリーニングする.

- 活性・不活性極性 [Hashimoto12] を用いて，ユーナリーパターン対を構成する 2 つのパターンの活性・不活性の極性が異なるユーナリーパターン対の場合は削除する.

- ユーナリーパターン対のPとQにおいてパターンを構成する動詞が同一であるが，変数とその動詞を媒介する助詞が異なるユーナリーパターン対は削除する．例えば，「Xが不足する」と「Xに不足する」などのユーナリーパターン対である．ただし，助詞「は」と「が」の組み合わせは許容し削除しない．

ここで，活性・不活性極性とは，Hashimotoらが提案した意味極性であり，助詞と動詞の組，すなわち本論文で言うユーナリーパターンに対して活性，不活性，中立の 3 つの極性が付与されている．活性極性が付与されたユーナリーパターンはそれを埋める名詞の主たる機能，効果，目的，役割，影響が準備あるいは活性化することを意味し，その典型例としては「X を引き起こす」「X を使う」「X を買う」が挙げられる．不活性のユーナリーパターンは逆にそれを埋める名詞の主たる機能，効果，目的，役割，影響が抑制あるいは不活性化されることを意味し，典型例は「X を防ぐ」「X が不足する」「X を破壊する」などが挙げられる．中立のユーナリーパターンは活性，不活性のいずれも付与できない意味的性質を持つものである．

含意関係を持つものとして生成されたユーナリーパターン対には「X が不足する」「X が足りる」のように意味的に逆であり，含意が成立していないものが含まれていた．これは含意パターン認識で使われている分布類似度がこうした意味的差をとらえられないためであると考えられる．一方で，活性・不活性極性に従えば，「X が不足する」は不活性，「X が足りる」は活性であり，それらの差を見ることによって，意味的差異をとらえることができる．活性ユーナリーパターンを 11,276 個，不活性ユーナリーパターンを 2,764 個，中立ユーナリーパターン 7,523 個を人手でアノテーションしており，このデータを用いて，ユーナリーパターン対で極性が異なるものを削除した．

以上のクリーニングによって，当初 9,192,475 個のユーナリーパターン対から 1,819,651 個のパターン対が削除され，最終的に 8,033,759 個のユーナリーパターン対がデータベースに格納された．なお，このうち，活性・不活性極性によるフィルタリングの結果除かれたユーナリーパターン対は 1,158,716 個であった．なお，回答取得に対するユーナリーパターンのクリーニングの効果については，4.8.3 で述べる．

4.5 構文パターンに基づく回答インデックス作成

大量の tweet から高速に回答を取得するためには，得られた tweet に対して回答を取得しやすいように，あらかじめ解析しておく必要がある．そのため，本節では，回答を取得するために作成されるインデックスについて述べる．ユーザから入力された質問文から生成したクエリに基づき，高速に回答を取得するためのインデックスを，本論文では，回答インデックスと呼ぶ．回答インデックスには，構文情報が充分に存在する文か

ら抽出される情報を格納するバイナリー回答インデックスと構文情報が十分にない文から抽出される情報も格納の対象とするユーナリー回答インデックスから成る. 以下に２種類の回答インデックスについて説明する.

4.5.1 バイナリー回答インデックス

回答インデックスの作成の共通の手順として，まず，対象となるテキスト（tweet）を文単位で形態素解析[14]，構文解析処理[15]を行う．次に，構文解析結果における任意の名詞句２つとそれらをつなぐ文節係り受けのパスを構成する表層上の連鎖を取得する．

例えば,「[宮城県で][炊き出しが][行われる]」という結果からは, 係り受けの構造上,「宮城県」と「炊き出し」という名詞句の間に「行われる」という文節を含むパスが存在するため, このパスを構成する 2 つの名詞句それぞれを変数で置き換えたパターンが取得される．この例では，それぞれの変数 X を「宮城県」，Y を「炊き出し」とすると，「X で Y が行われる」というパターンを取得できる．ここで取得されるパターンのことをバイナリーパターンと呼び, バイナリーパターンとそれに含まれる変数に対応する名詞句 2 つの三つ組をパターントリプルと呼ぶ. またパターントリプルを含む tweet 内のすべての名詞句を周辺名詞句として取得する．上記の例では, X で Y が行われる, X=宮城県, Y=炊き出し，の３つ組がパターントリプルとなる．最終的に，バイナリー回答インデックス には，パターンとして「X で Y が行われる」，変数に対応する名詞句としてそれぞれ「宮城県」「炊き出し」，周辺名詞がキーに登録され，その値には変数に対応する名詞句と当該 tweet の ID が格納される．

4.5.2 ユーナリー回答インデックス

ユーナリー回答インデックスは, バイナリー回答インデックスに比べて構文情報が十分でない tweet を対象とするための回答インデックスである．そのため，バイナリー回答インデックスで得られる回答に比べて, このインデックスを用いた回答の信頼性は高くないが，より広範な回答を得るために使用することができる．ユーナリー回答インデックスでは，ユーナリーパターンとその周辺名詞句をキーとして格納する．バイナリーパターンは，構文解析結果において二つの名詞句をつなぐパスから作られたが，ユーナリーパターンは名詞句一つと動詞，名詞，形容詞のいずれかへの係り受け関係から生成される．

---

[14] 形態素解析: MeCab を利用 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[15] 構文解析: J.DepP を使用 http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/ （辞書は JUMAN 体系を使用．）

「宮城県です．透析用器具が足りません．」といったtweetからは任意の２つの名詞句間に係り受けが存在しないため，バイナリーパターンを抽出することはできず，上記のtweetからの情報はバイナリー回答インデックスには全く反映されない．そこで，係り元の名詞句を変数としたユーナリーパターンを抽出することでこのような構文情報が十分でないtweetからも回答を取得できるようにする．この場合は「X（＝透析用器具）が足りません」が抽出され，周辺名詞句である「宮城県」「状況」「透析用器具」とをキーとして，変数に対応する名詞句「透析用器具」とtweetのIDを値としてユーナリー回答インデックスに登録する．

4.5.3 回答インデックスの利用

バイナリー回答インデックスは，パターントリプルを用いて作成したインデックスであり，ユーナリー回答インデックスは，パターントリプルが取得できないtweetにも対応することで，幅広い回答を取得するためのインデックスである．２種類の回答インデックスのキーと値を**表 16** にまとめる．

表 16 回答インデックスの構成

インデックス名	キー	値
バイナリー 回答インデックス	パターン，名詞句X，名詞句Y， 周辺名詞句	回答（＝名詞句X，名詞句Y）， リソースID
ユーナリー 回答インデックス	パターン，周辺名詞句	回答（＝名詞句），リソース ID

回答インデックスを用いて，どのように回答が取得されるかを説明する．バイナリー回答インデックスの場合，「震災後，宮城県で透析用器具が不足しています」というtweetからは，バイナリーパターンとして「XでYが不足しています」，名詞句X及び名詞句Yとして「宮城県」「透析用器具」，周辺名詞句として「震災後」「宮城県」「透析用器具」「不足」がキーに登録され，その値に名詞句X及び名詞句Yとそのtweetの IDが格納される．このようなエントリデータは，「宮城県で何が不足していますか」といった質問の回答を取得する際に使われる．この場合，質問文からは「XでYが不足しています」というパターンと「宮城県」という名詞句Xから成るクエリが生成される．バイナリー回答インデックスの検索の結果，そのエントリに値として登録されている名詞句Yの「透析用器具」が回答として，tweetのIDとともに出力される．

tweetからの得られた一つのエントリーを用いて，複数の質問にも対応することができる．例えば，「どこで透析器具が不足していますか」という質問に対しても回答を取

得できる．この場合，質問文から「X で Y が不足しています」というパターンと「透析用器具」という名詞 Y を持つクエリが生成され，値に登録されている名詞句 X の「宮城県」が回答として，tweet の ID とともに出力される．

一方，ユーナリー回答インデックスの場合，「何が足りませんか」という構文情報が十分得られないシンプルな質問に対して利用することができる．ユーナリー回答インデックスを検索することで，「透析器具が足りません」という tweet から「透析器具」を回答として取得することができる．さらに，ユーナリー回答インデックスは，「宮城県で何が足りませんか」という本来バイナリー回答インデックスを利用する質問にも利用できる．この質問では，「宮城県」は「足りません」という動詞にかかっているが，この宮城県を周辺名詞句として捉え直し，「X が足りません」というユーナリーパターンと「宮城県」という周辺名詞を指定することで，ユーナリー回答インデックスを検索できる．本来であれば，先の tweet の解析時に共参照解析等を行い，「透析器具が足りません」という文には「宮城県で」という表現が省略されていることを認識した上で処理を進めるべきであるが，共参照解析等の精度が高くない現状に鑑み，共参照，省略表現を一括して周辺名詞句として扱うことでユーナリー回答インデックスを利用して柔軟な回答の抽出を狙っていることになる．

なお，いずれのインデックスの作成時においても，retweet が入力として与えられた場合には，同一内容の retweet があるかをチェックし，もし存在すれば 1 つの retweet のみを登録し，これと同一内容の複数ある retweet はインデックスには登録しない．一方ですべての retweet の ID のリストは別途保存しておく．これは retweet の処理による質問応答の処理時間の増加を防ぐための処理である．

4.6 質問応答のためのマイクロブログの地名補完

マイクロブログ，特に Twitter では，その投稿に対する字数の制限等から，ユーザが言いたいことが重要視されており，それ以外の語は tweet から大胆に省略される傾向がある．そのため，回答インデックスを作成する上で重要であるにも関わらず，tweet の各文において省略されがちな地名を補完する処理を行うことが必要である．

Twitter では，スマートフォンなどの GPS 機能を持つ機器からの書き込みの場合，位置情報の開示設定がされていれば，tweet が書き込まれた場所を特定することができる．しかし，多くのユーザは，プライバシー等の問題から該当機能を有効にはしていない場合も多い．そのため，本研究では，マイクロブログの本文の情報から，自然言語処理により，地名や場所名の抽出を行い，必要な地名情報の補完を行う．しかし，地名の処理には次の問題があり，難しい課題となっている．

- 場所の非明示性：マイクロブログへの書き込みには，明示的に県や市の名称が書かれていないことが多い．さらには，tweet に限らず，一般的に，イベントが起きた場所を指す名詞句がイベントを表す動詞等に明示的には係らないことも多く，動詞で表されたイベントと地名を結びつけることは容易ではない．
- 場所の包含性： 場所には包含性がある．例えば，仙台市が宮城県の中にあることを知らなければ，たとえ文中に「仙台市」と記述されていても，「宮城県で」と地域を限った質問には回答できない．
- 場所の曖昧性： 一部の地名は曖昧性を持ち，上記の包含性を扱おうとする場合に，特に問題となる．例えば，「福島」という地名は日本全国に 50 以上もあり，そこから正しい一つを選ぶ必要がある．

これらの問題に完全に対応することは難しいが，次の手続きによって，地名とイベントとを対応させている．まず，現在入手可能なデータから大規模な地名・場所名辞書を自動生成する．さらに，地名等の包含性，曖昧性をヒューリスティックスによって対処し，回答インデックスに地名の情報を取り込む．

4.6.1 地名・場所名辞書

日本郵便が公開している郵便番号データを用いて地名辞書を作成した．郵便番号データからは，「都道府県／市区町村／町域」で表される住所の情報から，用いられる可能性がある地名文字列とその詳細な住所との対応を取り出す．地名文字列は「山元」のように断片的なものである場合が多いが，こうした対応づけを用いて，断片的な文字列から「宮城県亘理郡山元町」のようなより詳細な住所が入手可能となる．さらに，「都道府県／市区町村／町域」という住所の階層性は，先に挙げた場所の包含性に対処するための情報源となる．このようにして，2,486,545 のエントリ（地名文字列-住所の対の数は 5,129,162）を持つ地名辞書を作成した．そのうち，84,633 エントリが曖昧性をもつ地名であった．

Twitter などへの書き込みでは，住所のような地名の他に学校や施設，ランドマーク的名称の正式名称から通称までが幅広く用いられる．そこで，Wikipedia から抽出した上位下位関係 [Yamada09] から，上位語として自治体をとり，「（自治体名）の(*X)」（X は「施設」「学校」など）というパターンにマッチする下位語を取り出して利用する．例えば，「名取市の増田小学校」などである．これは，「学校」などの，郵便番号データには載っていないような場所にもその詳細な住所を対応づけるためである．上位語中の自治体名を，地名辞書で検索して下位語に住所を付与し，255,273 エントリを持つ場所辞書を作成した．

地名辞書および場所辞書は，全自動で作成しているため，それをそのまま文字列マッチによる単純な地名検出手法とともに適用した場合には，ノイズが問題となる場合がある．例えば，「枝野官房長官」の名字と同じ「枝野」が宮城県の地名として使われている場合があるなど，地名には人名と同じものが多くあり周辺の情報から適切に処理される必要がある．また，高頻出な普通名詞をいずれかの辞書のエントリとして含んでおり，誤って地名処理される場合もある．そこで，このような問題となるエントリを人手でフィルタし，最終的に，2,726,944 エントリを持つ地名・場所名辞書を作成した．

4.6.2 地名・場所名特定と拡張

回答インデックスを作成するために形態素解析，構文解析がされた解析結果の各文節に対し，形態素をその単位として最長の名詞句を抽出し，地名・場所名辞書を用いて地名・場所名を特定し，当該名詞句に詳細な住所候補を付与する．その際，名詞句全体がマッチしない場合でも，その範囲内で最左のマッチを選び，できるだけ住所を付与する．１文字の地名・場所名は誤ったマッチである可能性が大きいため対象外とする．なお，地名・場所名の特定に関して，通常の固有表現抽出を用いることが考えられるが，風間らの報告では，その有効性が確認されておらず [風間 12]，4.8.4 節で述べる比較実験においてもその有効性を確認できなかったため，本処理では固有表現抽出の結果を利用していない．

また，情報が無ければ最も広範囲な地域を表す住所，直前に曖昧性解消された住所がある場合には，それと最も整合性のある住所を選ぶルールに基づく曖昧性の解消を行っている．候補のうち，県・郡・市 （郡部の場合は町）部分が tweet 中の文字列と一致すれば，より広い地域レベルで文字列と一致しているものを優先する．例えば，「福島」の場合には，「福島県：福島市」，「大阪府：大阪市：福島区」等数多くの曖昧性があるが，最も広範囲な「福島県」が選択される．

前節でも述べた通り，高速に回答を抽出するため，tweet を取得する毎に構文解析を行い，回答インデックスを作成している．特定した地名・場所名を回答インデックスに反映させるため，地名・場所名を付与した新たな構文解析結果を生成する．具体的には，「イベントの場所は文中で直前に出現した地名・場所」という仮定を置き，直前の地名・場所に助詞「で」を加えたものを，イベントを表す動詞等に係るように付け加えた構文解析結果を生成する．例えば，「気仙沼中学校へ避難しています」という文があった場合，「避難」イベントの場所は，直前の場所である「気仙沼中学校」と認識され，さらに地名・場所辞書により「気仙沼中学校 →宮城県／気仙沼市」を取得し，「宮城県で」，「気仙沼市で」などの助詞「で」で終わる複数の文節が元の構文木に挿入される．こう

してできた構文解析結果を利用することで，補完された場所に関連する質問に対応したインデックスが生成される．これにより，元の文には「宮城県」という表現が含まれていない場合にも「宮城県でどこへ避難していますか」という質問に対し回答を取得できる．

4.7 質問応答処理

本節では，上記で説明した含意パターンデータベースと，回答インデックスを利用した質問応答処理について述べる．この処理では，ユーザが入力した質問文から回答集合を出力するまでの一連の処理で構成される．具体的には，ユーザからの質問を分析する質問解析処理と，その結果を利用して回答インデックスから回答を検索し回答リストを出力する回答検索処理から構成されている．質問文が入力されて回答が出力されるまでの処理を**図 22**に示す．また，質問文をユーザから受け取り，その回答を提示する入出力処理についても本節で説明する．

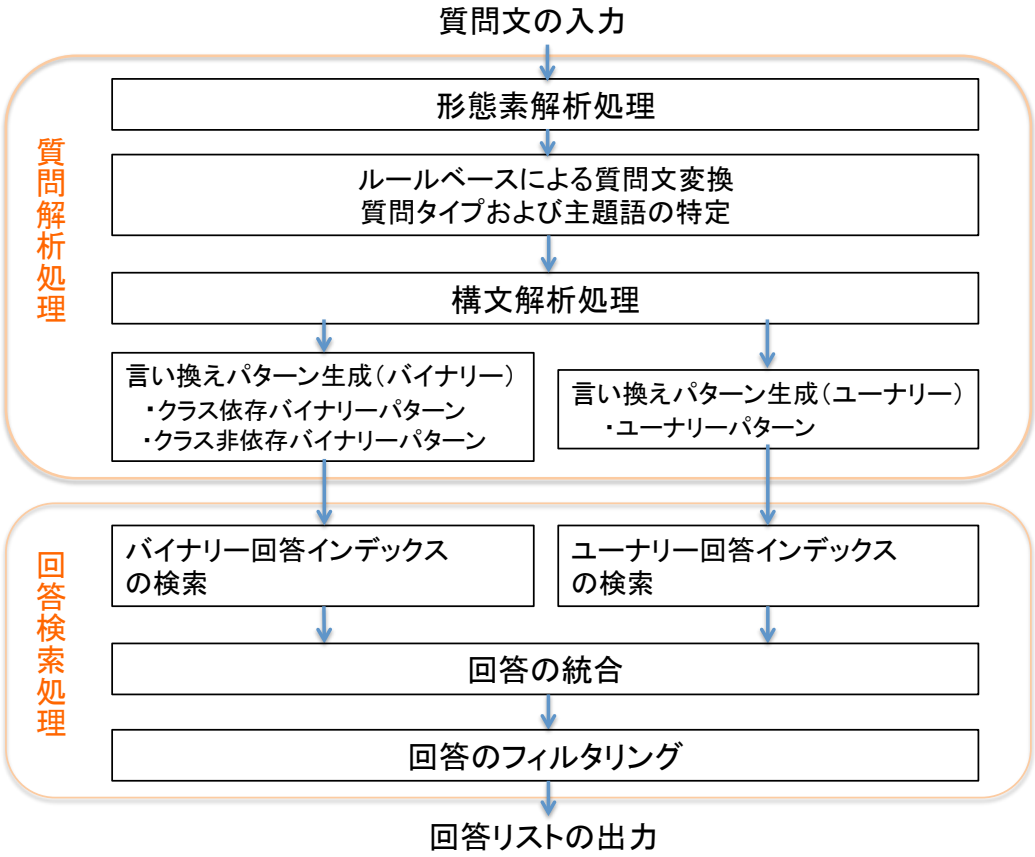


図 22 質問応答処理の流れ

4.7.1 質問解析処理

質問解析処理では，まず自然言語で入力された質問文に対して，形態素解析を行い，その解析結果を利用して格助詞の変更や疑問代名詞の位置の入れ替えなどを規則処理で行う．これは，質問文から得られる構文パターン（質問構文パターン）を複数生成してより多くの含意パターンを獲得し，幅広い回答を取得するための処理である．また，この際，場所を聞いている質問などの質問タイプを疑問詞や主題語から決定する．主題語の特定処理では，疑問代名詞に助詞「は」とともに直接係る名詞がある場合，その名詞を主題語として取得している．例えば，「被災地で不足している食べ物は何ですか」という質問が入力された場合，名詞「食べ物」を主題語として取得する．この主題語は，得られた回答との分布類似度 [Kazama08] により，回答候補を選別するための情報として利用される．例えば，「食べ物」に対して分布類似度が高い上位の名詞には，「お菓子」,「酒」，「魚」,「肉」，「ワイン」，「チョコレート」などの食べ物が含まれている．逆に食べ物と関連性の薄い「タオル」や「電化製品」の分布類似度は非常に低い．このように，主題語と回答候補との分布類似度は，質問の回答として相応しくない回答候補を除外する特徴として利用できる．

次に，言い換えられた質問文の構文解析結果から疑問代名詞以外の名詞句一つと疑問代名詞を特定し，その間の係り受け関係パス上にある表現からバイナリーパターンを取得する．例えば，「宮城県で何が不足していますか」という質問が入力された場合，「X（＝宮城県）で Y（＝何）が不足している」という基本的なバイナリーパターンに加え，言い換えられた質問文から「Y が X で不足している」（格要素の入れ替え），「Y は X で不足している」「Y が X では不足している」「X で Y は不足している」「X では Y が不足している」（助詞の変換），「X で不足している Y」（ガ格疑問代名詞の被連体修飾化）などのバイナリーパターンが得られる．この際，クラス依存の言い換えのために，X に対応する名詞句の意味クラスを [Kazama08] の手法により取得する．このようにして得られたバイナリーパターンを用いて，回答インデックスを検索するクエリが生成される．例えば，「宮城県で何が不足していますか」の質問からは，パターンに「X で Y が不足している」，X に対応する名詞句 に「宮城県」を指定したクエリが得られる．ユーナリーパターンについてもバイナリーパターンと同様の処理が行われ，「Y が不足している」，周辺名詞に「宮城県」を指定したクエリが得られる．

疑問代名詞以外に 2 つ以上の名詞句が含まれる場合は，疑問代名詞と名詞句一つとそれをつなぐ文節で表される複数のパターンを抽出する．例えば，「宮城県ではどこで携帯が充電できますか」が入力された場合，「X（＝宮城県）では Y（＝どこ）で充電できる」，「Y（＝どこ）で X（＝携帯）が充電できる」のバイナリーパターンが取得

される．この結果から，パターンに「X ではY で充電できる」，X に対応する名詞句に「宮城県」，周辺名詞句に「携帯」が指定されたクエリと，パターンに「Y でX が充電できる」，X に対応する名詞句に「携帯」，周辺名詞句に「宮城県」が指定されたクエリが生成される．同時に，ユーナリーパターンとして「Y で充電できる」，周辺名詞句に「宮城県」と「携帯」が指定されたクエリも生成される．なお，クエリで指定される周辺名詞句は，質問文に含まれる全名詞句から，パターンや名詞句に含まれる名詞句を除外し作成される．

最終的な回答の取得するために，質問文から直接獲得した質問構文パターンで，含意パターンデータベースを引くことで大量の言い換えパターン（クラス依存バイナリーパターン，クラス非依存バイナリーパターン，ユーナリーパターン）が取得される．具体的には，一つの質問から得られる複数個の質問構文パターンの各々につき，最大で 1,500 の言い換えパターンが生成される．その内訳はそれぞれデータベースに格納されているクラス依存バイナリーパターンが最大で 500 個，クラス非依存バイナリーパターンが最大で 500 個，ユーナリーパターンが最大で 500 個となる．これらのパターンは質問文中に出現する名詞句と組み合わせて回答インデックスの検索に使われる．

4.7.2 回答検索処理

回答検索処理では，質問解析処理で得られた言い換えパターンを利用して，回答インデックスが検索され，回答と回答が抽出された tweet の ID が得られる．回答インデックスを検索する際のキーとしては，質問文中で共起する疑問代名詞以外の名詞句と含意パターン，質問文中の周辺名詞句が利用される．

回答検索では，各々の回答インデックスは本論文の実験では数千万件レベルの大量の tweet をカバーしているため，如何にこの回答インデックスを引く操作を高速化するかが重要になる．そのため，BloomFilter [Bloom70] を利用して，回答インデックスに共起がないパターンと名詞句の組み合わせから成るパターントリプルをメモリー上の操作のみで近似的に検出し，ディスクアクセスを伴う回答インデックスの検索回数を大幅に減らしており，これにより実用的な速度を得ている．

これまでにも述べたとおり，二つの名詞句をつなぐバイナリーパターンと周辺名詞句をキーとするバイナリー回答インデックスは，質問文からパターントリプルが取得できた際に検索される．ユーナリーパターンをキーとするユーナリー回答インデックスは，二つの名詞句をつなぐバイナリーパターンが質問文から抽出されたときも含め，ユーナリーパターンが得られる場合すべてにおいて使用される．さらに，ユーナリー回答インデックスに対して，パターンやその内容語を周辺名詞句として検索することで，パターンに直接係り受けがない回答も取得できる．また，ユーナリーパターンに含まれる内容

語のみをとりだし，それを周辺名詞句として検索することも行う．これは例えば「何が不足しているか」という質問に対して，「不足」のみを周辺名詞句として検索することに相当する．

なお，抽出された回答にはストップワードによるフィルタ，場所名フィルタ，非場所名フィルタが適用される．ストップワードによるフィルタは，あらかじめ用意したストップワードリストに回答が含まれる場合にそれを回答リストから削除するものである．ここで使用しているストップワードリストは含意パターンデータベース構築の際に用いた 6 億ページの Web 文書から形態素態素解析器を使って自動的に認識された名詞句（複合語および単語）のうちで，明らかに解析ミスであり語として認められないものや非常に漠然としており明確な概念を指しているとは言えないもの（例：「皆さん」「その他」），さらには主として機能語的に利用される語（例：「理由」「モノ」）を人手で集めたものである．これは現在 164,064 の名詞句を含んでいる．

場所名フィルタは，疑問代名詞「どこ」を含む質問に関して，前述した地名・場所名辞書にある語を含む回答，前述した単語クラスタリングの結果から場所名をさす語を多く含む 48 クラスに含まれる語を含む回答，あらかじめ用意した正規表現によるパターン 113 個に合致する回答のいずれでもないものを回答リストから削除する．一方で疑問代名詞「何」を含む質問に関しては，非場所名フィルタを適用する．これは場所名フィルタを逆に用いて地名フィルタでは削除される回答のみを最終的な回答リストに含めるフィルタである．

4.7.3 入出力処理

入出力処理は，ユーザからの質問文を質問解析処理へ送信し，回答検索処理から出力される回答リストをユーザに提示するための処理である．入出力処理は，Web ブラウザ上で動作するユーザインタフェースを備えており，一連の操作はブラウザを通してテレビ，パソコン，タブレット，スマートフォンなどから操作することができる．また，入出力処理では，質問応答エンジンから出力される大量の回答の俯瞰的な把握を可能にするために，意味的，地理的，時間的観点から分類した上でそれらの全体像を把握する２種類の表示方法で結果を提示することができる．こうした俯瞰的把握によって，通常考えつかない想定外の事象の発見も可能になり，また，それらへの対処も容易になることが期待できる．表示方法のひとつは，回答結果を単語の意味クラス毎にまとめて表示し，もう一方は，場所を尋ねる質問に適した結果の表示方法として，地図上に回答を表示する．以下で，それぞれについて説明する．

図 23 意味的回答提示

- **意味的回答提示**

意味クラスを利用した回答の提示方法での実行例を**図 23** に示す．この回答提示では，回答が意味クラスごとにまとめられ表示される．意味クラスは，[Kazama08]で提案された手法により計算されたものを用いるが，意味クラスの計算対象外であるような長い名詞句に対しては，文字列の部分マッチを適用するなどして対応する．また回答の文字列をクリックすると，回答を抽出してきた情報源（tweet）へのリンク，もしくは回答を抽出してきた tweet そのもの表示するウィンドウがポップアップし，回答が抽出された tweet の内容を確認できる．同じ回答でも複数の tweet から得られる場合があるため，ひとつの回答にひもづけられる tweet は多いものでは数百から数千に及ぶ．

- **地理的回答提示**

回答を地図上へ表示する提示方法での実行例を**図 24** に示す．この表示方法では，質問の回答となる場所の位置が地図上で表示される．例えば，「宮城県のどこで炊き出しをしていますか」という質問に対して，炊き出しが行われている地点が容易に把握できるようになる．この提示方法において受け取る情報は，意味マップモードの場合と同一である．この提示方法では，地図上に回答を表示するために，次のことを行う．

1. 質問が場所を尋ねる質問（～はどこですか，どこで～できますかなど）の場合，回答は地名・場所名であることから，回答に対応する詳細な記述を地名・場所名辞書から得る．
2. 1.で得られた記述を使って，geocoding[16]を用いて住所やランドマーク名から緯度経度の獲得を行い地図に表示する．
3. 場所を尋ねる質問以外（～は何ですか，～は誰ですか，～はいつですか）の場合，回答の情報抽出源に対し，4.6 節で述べた地名補完処理で取得した地名の詳細な記述を取得する．
4. 3.で得られた記述を使って 2.と同様の処理を行い，回答を地図上に表示する．

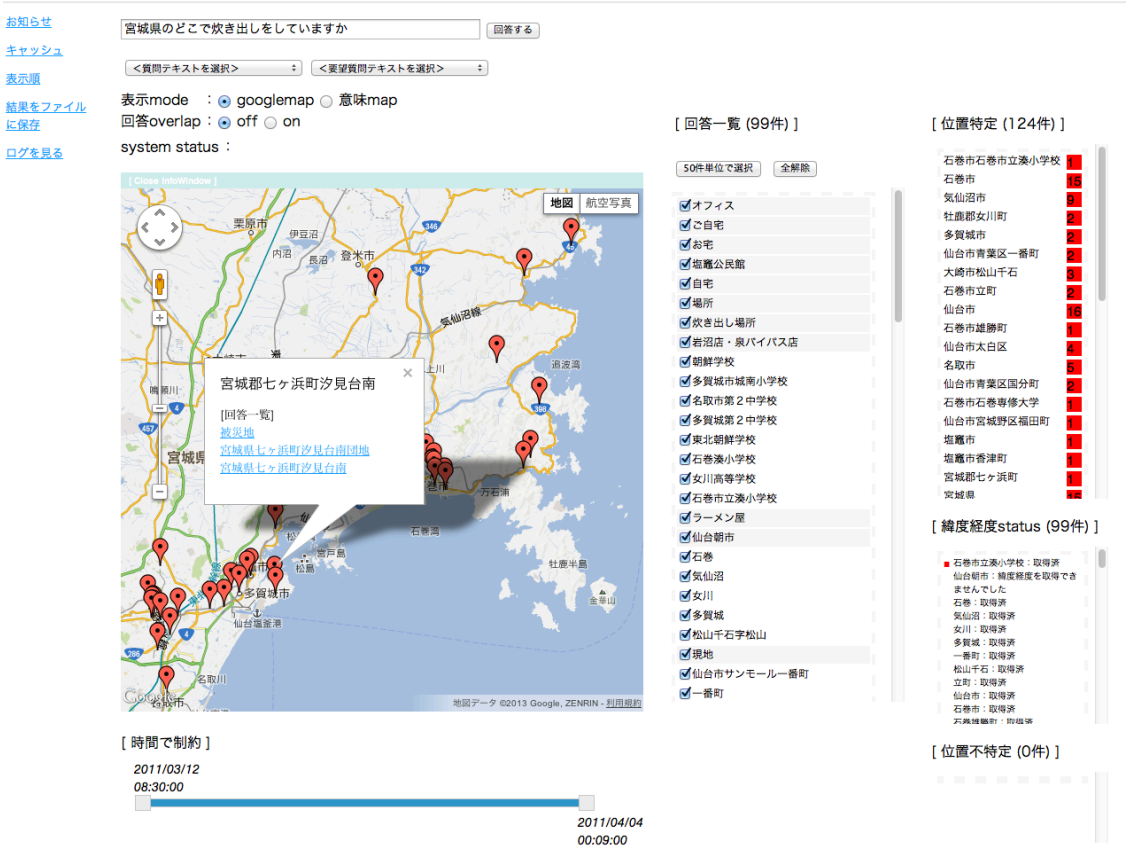


図 24 地理的回答提示

[16] https://developers.google.com/maps/documentation/geocoding/

地理的回答提示においても，意味的回答提示と同様に，地図上に配置されたマーカーをクリックすると，対応する回答とその回答が抽出された tweet へのリンクが表示される．

- **時間的回答提示**

上記で説明した意味的回答提示と地理的回答提示の共通の機能として，情報抽出源のテキストの発信時刻による回答の限定，すなわち時間的回答提示を行うことが可能である．それぞれの提示画面の下部に表示されているタイムスライダーを操作することにより時間帯を指定すると，その時間帯に発信されたテキストから抽出された回答のみが表示される．通常，回答が抽出されたテキストの発信時刻は，一般の Web ページを対象としていた場合は特定が困難であるが，Twitter や 他の SNS（Social Networking Service）においては発信した時刻がその情報に記載されているため，テキストの内容と作成された時刻を容易に結びつけることができる．

この時間的回答提示機能を利用することで，意味的回答提示と地理的回答提示のインタフェースにおいて，特定の期間に発信されたテキストからの回答が欲しい場合や，古くなった情報を非表示にしたい場合に，必要とする期間の回答のみを提示できる．これにより，例えば「宮城県で何が不足していますか」という質問で，災害発生時から一週間に不足していたものとそれ以降に不足しているものの時間的な推移などを調べることも可能である．

4.8 東日本大震災時のマイクロブログデータによる質問応答の性能評価

本節では，ここまで述べた方法を実装した質問応答システムを評価する実験について述べる．実際に運用される場面を想定したシステムの性能を評価することが望ましいが，今回実装したシステムは，多くの処理モジュールから構成され，その複雑性や開発途上にあることを考慮して，システムの基本機能，すなわち質問応答機能に関して評価を行った．

システムが回答を取得するための対象となるマイクロブログのデータとして，2011 年 3 月 9 日から同年 4 月 4 日までの Twitter の約 2 億 2 千万 tweet（(株)ホットリンク提供）を用いた．このうち，実験では，災害に関連する 345 個のキーワードによりフィルタした約 5,400 万の tweet を用いた．この全 tweet から，システムが回答を取得するための回答インデックスとして，約 1 億 2 千万エントリを持つバイナリー回答インデックスと，約 7 億 6 千万エントリを持つユーナリー回答インデックスが生成された．なお，こ

の回答インデックスの作成には，Apache Jakarta Project のもとで開発が進められている Lucene[17]を利用した．

また，提案システムの評価に加え，ユーナリーパターンの作成における活性・不活性極性のクリーニングの有用性を確認する実験，地名補完処理における固有表現抽出の有効性を確認する実験，教師有り学習を用いた回答のランキングの有効性を確認する実験のそれぞれについても本節で報告する．

4.8.1 実験条件

マイクロブログに投稿されている膨大な情報を整理・分析し，全体的な把握を可能とする本手法では，入力された質問に対して対象データにおいて，その回答の頻度が高いなどの目立った回答だけではなく，ロングテール部分に存在する想定外も含めた事実を回答として網羅的に取得する必要がある．そのため，質問応答で得られる回答の再現率が重要な評価指標であると考えられる．本質問応答システムの性能を評価するために，川田らが東日本大震災の tweet データから大規模に作成した質問応答評価セットを利用する [川田 13]．この評価セットは，6 名で作成した質問 300 問の各々について，質問に関連するキーワードで対象とする tweet を全文検索した結果をランダムに 1,000 件を取得し，その結果から人手で回答を抽出することができた 300 問中の 192 問の質問文とその正しい回答のペア（以下，正答と呼び，その数は 17,524 個である）のセットである．

評価セットの正答には質問とは表層的に大きく異なる表現で記載された表現から抽出されたものも多数含まれる．また評価セットでは，質問に対する回答が一意に求まるものではなく，ひとつの質問に対して複数の正解が存在する．また，この評価セットは単に質問と正答，つまり名詞句のペアをデータベース化しただけではなく，正答が抽出された tweet も含んでいる．質問応答の評価実験では，再現率は評価セットに含まれる正答のうちいくつシステムが回答できたかで評価する．当然ながら，評価セットに含まれていないが，正解と判定される回答をシステムが出力することが考えられるが，それを考慮して再現率を計算すると，新たな正解が見つかる度に再現率がかわるため，評価セットに含まれる正答のみ考慮して再現率を求めた．一方，適合率は，システムの回答をランダムサンプルし，正解かどうかを人間が判定して求めた．**表 17** に実験に利用した質問の一部を示す．

[17] http://lucene.apache.org/core/

表 17 実験に利用した質問例

インフラ	どこに給水車が来ますか
	停電した時の注意点はなんですか
	どこでガスが復旧していますか
物資	必要な家電は何ですか
	スポーツドリンク代わりになる物は何ですか
	支援物資の受付窓口はどこですか
生活	どこでお風呂に入れますか
	どこで遺体の火葬をやってもらえますか
	営業しているお店はどこですか
ボランティア	ボランティアに適した服装は何ですか
	介護士のボランティアはどこで募集していますか
	どこで復旧作業が行われていますか
支援活動	どこで募金ができますか
	支援が必要なのはどこですか
	どこで炊き出しをしていますか
情報・交通	どこで携帯電話の電波は入りますか
	どこで道路が寸断していますか
	機能している空港はどこですか
災害状況	津波の高さはどのくらいですか
	震災による経済的損失はいくらですか
	どこで土砂崩れが起きていますか
病気・負傷	はやっている病気は何ですか
	クラッシュ症候群の注意点はなんですか
	不安解消に効くものは何ですか
原発・汚染	放射能が高いのはどこですか
	なにが汚染されていますか
	線量はどのくらいですか
安否確認	どこで安否確認ができますか
	どこで身元の確認ができますか
	救援を求めているのはどこですか
政府・行政	どこに自衛隊がいますか
	自治体の就労支援には何がありますか
	緊急車両が通れる道はどこですか

4.8.2 評価結果

評価では，再現率を計算する際に，システムの回答が正答を部分文字列として含んでいるか，システムの回答が正答に部分文字列として含まれているいずれかの場合を正解とした．その結果，再現率 0.519 (9,099/17,524) が得られた．この部分文字列による照合では，正答かシステムの回答が一文字である場合に，多数の回答にマッチし，評価の精度が問題になる可能性があるが，前述したように提案システムは一文字からなる単語を回答として出力しない．また，評価セットの正答で一文字のものは全部で 106 個あったが，システムの出力でそれらにマッチしたものは 67 個であった．これはシステムの回答の 4%程度に相当する．しかし，これらすべてを回答から除外した場合の再現率は，0.519 (=(9,099-67)/(17,524-106))と変わらず，この影響は小さいと考える．また，192 の質問ごとに再現率を求め，その平均をとると 0.428 であった．これは，もともとの正答数が小さい質問において，再現率が 0 となってしまう場合が多い（192 問中 41 問，そのうち回答数が 0 のものは 32 問）ためであり，このことから，逆に質問の正解が得られた場合の再現率は，この数値よりも大きい場合が多いことを期待できる．適合率に関しては，全回答から質問と回答のペア 250 個をランダムサンプルし 3 名の評価者で正解かどうかを調べ，その多数決により正解を決めた．評価者間の一致度合は Kappa 値 [Fleiss71]が 0.507 であった．回答の評価に際しては，回答が抽出された元の tweet が非常に大量の場合があるが，ランダムに選択した最大 3 個の tweet から正解かどうかを判断した．評価の結果，250 問の適合率は，0.608（152/250）となった．

バイナリーパターンを利用した質問では，「どこで風評被害が起きていますか」という質問の回答では，「Y で X（＝風評被害）が出ている」「X（＝風評被害）が Y で発生している」「Y で起きている X（＝風評被害）」「X（＝風評被害）が Y で起こる」「X（＝風評被害）が Y で起きている」などのパターンにより回答を取得している．また，ユーナリーパターンを利用した質問では，「なにが汚染していますか」という質問で「Y が汚染されてしまう」「Y が汚染される」「Y の汚染」などのほか，「Y から検出される」「Y からは検出される」などのユーナリーパターンが含意パターンデータベースから取得され利用された．これにより「4 号機，正門，ヘリ」などの tweet に「汚染」を含んでいない回答も得ることができている．

再現率を下げている要因の一つとしては，回答がまったく取得できない質問が 32 問あることがある．これらの多くは，質問文を構成する名詞句が tweet において非常に低頻度であり，手掛かりとして役に立たない場合である．例えば，「専門職ボランティア」，「被災者相談窓口」，「被ばく相談」，「被災者就労支援」など，質問作成者がノ格を省略したと考えられる複合名詞や，「津波肺」「クラッシュ症候群」「誤嚥性肺炎」な

どの専門的すぎる用語のため tweet には現れない固有名である．これらは，該当する複合名詞や固有名が回答インデックスに存在しないか登録されていても非常に少数であった．対応策としては，「被災者相談窓口」を「被災者の相談窓口」とするなどの複合語の分割が有効であり，さらにサ変名詞を語尾にもつ「被ばく相談」「就学支援」のように複合名詞が「行う」「できる」「実施する」などに係る場合は，「被ばくを相談する」，「就学を支援する」などのより汎用的な表現に変換することが必要である．専門用語については，同義語や上位下位概念の言語資源を利用して，より汎用的な表現に言い換えることなどが考えられる．今後，複合語の構造解析手法や名詞の言い換え手法などを取り入れ，より幅広い質問にも対応できるようにする予定である．

また，適合率を評価した回答 250 についてより詳細に分析した．これらの回答がどういった処理によって抽出されるかを見るとまず，クラス依存，クラス非依存をふくめて「X が Y で不足している」のように二つの変数を含むパターンによって得られた回答は全体の 6%（15 個）であり，その適合率は 0.933 であった．また，「Y が不足している」のようなユーナリーパターンで抽出された回答は 72%（180 個）を占め，適合率は 0.656 であった. さらにユーナリーパターンの内容語を抽出して得られた回答は 22%（55 個）であり適合率は 0.364 であった．期待されるように制約の強いパターンで取得されている回答は適合率が高いものの，変数を二つ含むパターンの適用例はきわめて少なかった．これは「どこが渋滞していますか」のようなそもそも二つの変数を含むパターンが抽出できない比較的簡単な質問が評価セットに多かったことも理由である．今後「宮城県のどこで渋滞していますか」のようなより複雑な質問を評価セットに加えると，この制約が強いパターンが適用される割合も増加するものと考える．

誤った回答が抽出された要因を見ていくと，もちろん，パターン間の含意の認識誤りも含まれてはいるが，むしろ目立つのは「水は不足していますか」「水が不足したりして」「水は不足していません」などのように単純な肯定文以外の文から「X が（は）不足する」のようなパターンが抽出されている場合である．モダリティ解析を導入することによって，これらの文を除くことで適合率改善ができると予想している．一方で，「水は不足していますか」のような質問や要望，「水が不足していたとしたら」のような仮定も，災害時において非常に有用な情報であり，個別に認識することは重要な課題だと考えている．

また，地名補完処理の誤りによって，パターンやその内容語から離れた位置に出現する場所名が誤って回答として抽出されるケースがあった．これらは今後，省略，共参照解析を導入することで改善していく予定である．

4.8.3 ユーナリーパターン対のクリーニングの効果

ユーナリーパターン間の含意関係のクリーニングが質問応答全体に及ぼす影響について評価を行った．ユーナリーパターン間の含意関係とは，例えば「X が崩落する」「X が崩壊する」の間に成立する含意関係である．このクリーニングにおいては，活性・不活性極性を用いたクリーニング（活性・不活性クリーニング），ならびに同一の動詞を含むユーナリーパターン間で助詞のみが異なるものを削除するクリーニング（助詞クリーニング）の二種類を行った．まず，提案システムの再現率は 0.519，適合率は 0.608 であったが，ユーナリーパターン間の含意関係に対して助詞クリーニングのみ適用し，活性・不活性クリーニングを適用しなかった場合の回答を，提案システムと同様に回答 250 サンプル（評価者 3 名による評価）を抽出し，評価したところ，**表 18** に示すとおり，再現率 0.524，適合率が 0.536 となった．つまり，再現率は 0.005 とわずかに向上したが，適合率が 0.072 と大きく低下したことになる．さらに，活性・不活性クリーニング，助詞クリーニングの両方を適用しなかったときの性能は，再現率が 0.533，適合率が 0.448 となり，やはり再現率がわずかに向上したものの適合率の大幅な低下が見られた．最終的にいっさいクリーニングを行わなかった場合と提案手法を比べると，再現率が 0.014 程度向上するのに対して，適合率は 0.160 と大幅に低下している．まとめると，ユーナリーパターン対のクリーニングは最終的な回答の質において非常に重要であるということが分かった．特に，一見含意関係とは関係の薄い，活性・不活性という意味極性がそのクリーニングにおいて重要な役割を果たすことが確認できた．

表 18 ユーナリーパターン対のクリーニングの効果

	再現率	適合率
提案システム	0.519	0.608
活性・不活性クリーニング未使用	0.524	0.536
活性・不活性及び助詞クリーニング未使用	0.533	0.448

4.8.4 地名補完処理における固有表現抽出の効果

本研究での提案システムは地名補完処理で固有表現抽出処理を使用しなかったが，それは以下の実験結果により，本システムにおいては固有表現抽出の有用性が認められなかったためである．まず IREX 固有表現コーパス [Sekine00] において LOCATION タグ

のみを残し，これを学習データ 1 とした．次に，Twitter API を使用して，実験で用いる tweet とは異なる期間の tweet 22 万 5 千件を取得し，これに対し，災害関連のキーワード 345 個のいずれかを含む 11 万 tweet に対して学習データ 1 から作成した固有表現の認識器（NER）を適用し，LOCATION タグを付与した．この結果のうち 4 万文を人手で修正し，これを学習データ 2 とした．これらの学習データ 1，データ 2 を合わせて学習データとし，CRF++[18]を用いて形態素単位の NER を構成した．素性テンプレートは CRF++パッケージのサンプルとして含まれているものをそのまま利用した．

この NER を評価するために，対象としている 5,400 万の tweet から 1,000 tweet（3,017 文）をランダムサンプルし，構成した NER を適用した．その結果を人手で修正し，評価用テストセットを作成した. この評価用テストセットの形態素数は約 33,000 であり，LOCATION とされる名詞句は，521（866 形態素）存在する．これを用いて構成した NER を評価したところ，適合率は 0.930，再現率は 0.839 であった．

次に，地名補完処理における処理対象の特定に NER を組み入れた場合と，形態素単位の文字列によって直接辞書引きすることで特定する場合との違いが質問応答性能に与える影響を調べた．なお，実験にはユーナリーパターン対のクリーニングを行う前のシステムを用いた．**表** 19 に示すとおり，NER を用いない場合が再現率 0.533，適合率 0.448 であり，NER を用いた場合には再現率 0.516，適合率 0.392 となり，再現率，適合率ともに低下した．この結果から，あるエンティティが地名・場所名辞書に存在しているにもかかわらず，NER がそれを特定できなかった場合や，逆に NER が地名補完処理での処理対象を特定できても地名・場所名辞書に登録されていない場合などがあり，地名・場所名辞書を直接辞書引きしたほうが，より高い性能を発揮できたと考える．

表 19 地名補完処理における固有表現の認識の効果

	再現率	適合率
NER 未使用	0.533	0.448
NER 使用	0.516	0.392

より具体的に，NER で特定されたものがどれだけ地名・場所名辞書を用いて地名補完処理されたかを見てみると次のようになった．NER はテストセットに 521 あるエンティティのうち，437（再現率 0.839）を正しく特定できているが，このうち，地名補完

[18] http://crfpp.googlecode.com/svn/trunk/doc/index.html

処理の対象となったのは，わずか157個である．この数字が小さい理由は，現在の地名補完処理はシステムの持つ地名・場所名辞書にあるエントリしか処理対象としないからであり，さらにはNERの結果と地名・場所名辞書との食い違いが大きいからである．一方，地名補完処理での処理では，214個の地名・場所名を特定し，地名補完処理がなされた．もちろん，この地名補完処理がなされた地名・場所名には誤ったものも多数含まれている．NERを導入した動機は，NERにより一般名詞や人名等を地名として誤認識することを防げるかもしれないということであった．つまり，地名補完処理の対象認識の適合率の向上をねらったということである．地名・場所名の誤認識がNERにより防がれたケースもあったと推測されるが，そもそも地名補完処理が起動されないデメリットの方が大きく，最終的な質問回答の性能が低下したと考えられる．

今後NERの認識結果を地名・場所名辞書に追加することによって，性能向上は可能かもしれない．そこで障害となるのはエンティティの基準と，地名補完処理において処理対象とするエンティティ，すなわち地名・場所名辞書のエントリの認定基準とが異なっていることである．例えば，複合名詞中，「富士スピードウェイ」の「富士」のみが地名として認識されるといった問題も存在する．また，地名・場所名辞書では地名間の包含関係が情報として含まれているが，NERの結果にはそうしたものは含まれない．これらの問題をどう解決していくかが，今後の課題の一つとなる．

まとめると，風間ら [風間 12] の報告と同様に提案システムでのNERの効果は確認できなかった．この理由は，現状の地名補完処理では，固有表現の特定後に地名・場所名辞書にて詳細な地名情報を取得する必要があり，この辞書の網羅性等が性能に影響するためである．したがって，システムの性能を向上させるためには，NERの認定基準と本タスクで必要とされる地名・場所名の認定基準との擦り合わせ，さらには地名・場所名辞書との整合性をとる自明でない作業が必要となる．

4.8.5 回答のランキングの効果

本研究で提案するシステムでは，ロングテールに存在する回答についても取得し，マイクロブログ全体の情報の俯瞰を目的の一つとしているため，再現率を重視し，今まで述べてきた手法で発見できた回答をすべて出力している．一方で，自明な拡張は回答にランキングメカニズムを導入し，さらなる拡張を図ることである．本来，再現率を重視しつつ，ランキングを導入し，提案手法よりも高い性能を達成するためには，提案手法よりも公汎な回答を出力し，ランキングに基づいて回答の足切りを行うべきであるが，現状はそこまでの実験は行えていない．本節では，提案システムが出力する回答を教師あり学習に基づいてランキングした結果について報告する．

今回行った実験で使ったランキング手法は, 回答とパターンに関する素性をもとに学習したSVMのスコアによりランキングを行うものである．**表 20** に，SVMの学習に利用した素性を示す.

表 20 回答のランキングに使用する素性一覧

素性の種別	素性
パターン	質問構文パターンにより回答が得られたかどうか
パターン	クラス非依存バイナリーパターンにより回答が得られたかどうか
パターン	クラス依存バイナリーパターンにより回答が得られたかどうか
パターン	ユーナリーパターンにより回答が得られたかどうか
パターン	ユーナリーパターンでのキーワード検索により回答が得られたかどうか
パターン	ユーナリーパターンの内容語によるキーワード検索で回答が得られたか
パターン	クラス依存バイナリーパターンのスコア
パターン	クラス非依存バイナリーパターンのスコア
パターン	ユーナリーパターンのスコア
パターン	回答を取得したパターン数
パターン	連体修飾型でないパターンの有無
パターン	回答取得したパターン数と連体修飾型でないパターンとの比率
パターン	質問構文パターンと同じ漢字を含むパターンの有無
回答	複数のパターンから得られた同じ回答の個数
回答	回答の形態素数
回答	回答の文字数
回答	回答の意味クラス
回答	未特定の意味クラスかどうか
回答	部分文字列によるクラス特定を利用したかどうか
回答	バイナリーパターンと2つの名詞句の意味クラスのPMI
回答	質問構文パターンと質問文中の名詞に基づく回答の意味クラスの尤度
回答	疑問代名詞タイプ
回答	回答が疑問代名詞の対応するクラスに属するかどうか
回答	回答と主題語との分布類似度
回答	回答が主題語の下位概念かどうか
回答	回答の末尾が主題語となるかどうか

まず, パターンの属性に基づく素性として, 質問構文パターン, クラス依存パターン, クラス非依存パターン, ユーナリーパターンからのいずれのパターンで回答が得られたか, あるいはユーナリーパターンとその内容語によるキーワード検索を用いたかを示す 2 値の素性を用いる．これに加え，クラス依存パターン，クラス非依存パターン，ユーナリーパターンの各スコアを用いる. ある回答が複数の異なるパターンから得られた場合には，その全パターン数，パターンが回答を連体修飾していないどうか，全パターン数と回答を連体修飾していないパターンとの比率を利用する. また回答を抽出した含意パターンやユーナリーパターンが, 質問構文パターンと共通の漢字を持つかどうかも利用する.

回答の属性に基づく素性では，まず，様々なパターンから得られた同じ回答の個数, その文字数及び形態素数を用いる．次に，回答の意味的な情報として，回答の意味クラス，その意味クラスを特定する際に部分文字列を用いたか，回答のクラスが未特定かどうかの 2 値の素性を用いる．また，回答を獲得したパターントリプルのバイナリーパターンと 2 つの名詞句の意味クラスの PMI（相互情報量 : Point-wise Mutual Information）, 質問構文パターンと質問文中の名詞に基づく回答の意味クラスの尤度 [De Saeger09] を利用する. さらには, 質問文から得られる疑問代名詞と主題語を利用した素性として, 疑問代名詞タイプ，回答が疑問代名詞の対応するクラスに属するかどうか，回答と主題語との分布類似度，回答が主題語の下位概念となるかどうか，回答の末尾に主題語を含むかどうかを用いる.

上記の素性を用いて，線形，多項式（二次），放射基底関数（RBF, 比例定数 1）の各カーネルを用いて SVM の学習を行い，いずれのカーネル関数を用いるべきか検討した．学習データは，災害に関連の深い質問 60 問（これまでの評価で利用した質問とは別である）と，システムが出力した回答のペア合計 5,044 個に対して正解／不正解のラベルを付与したデータである．なお，このデータは提案システムの古いバージョン，つまり, 場所名フィルタやユーナリーパターンの含意パターンデータベースのクリーニングを行っていないシステムの出力を含んでおり, 現在の提案システムでは出力できない回答も含まれている. 10 分割交叉検定の結果, 線形カーネルで F 値 0.642（適合率 0.681, 再現率 0.607），多項式カーネルで F 値 0.631（適合率 0.626，再現率 0.634）, RBF カーネルで F 値 0.529（適合率 0.719，再現率 0.419）が得られた．本システムでは F 値が最も高かった線形カーネルにより学習された分類器の出力するスコアを利用することを検討した.

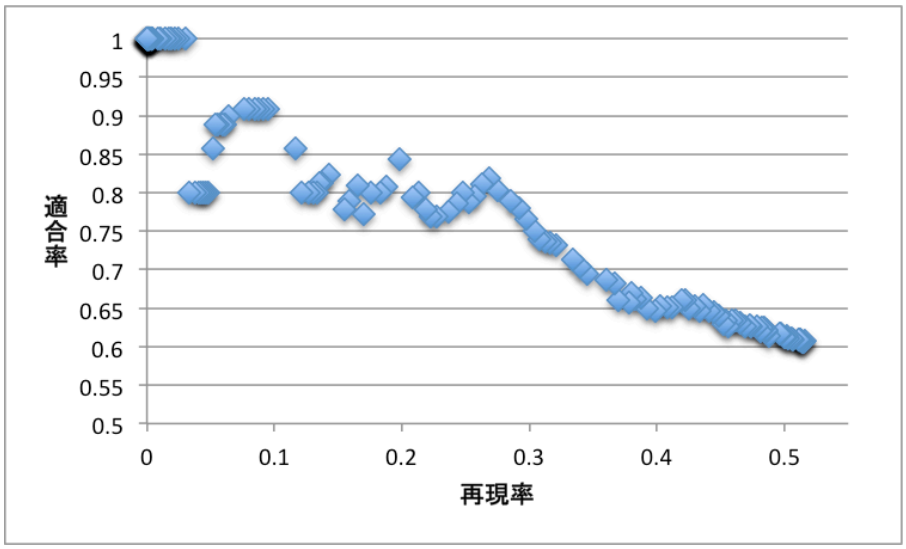


図 25 回答のランキング結果

4.8.2 節の実験にて利用した 250 個の回答サンプル（適合率 0.608）を以上の分類器のスコアでランキングした結果が**図 25** である．グラフの再現率は提案システムの出力すべてを SVM のスコアに閾値を設けて足切りを行い，足切りを生き延びた回答集合を 17,254 件の正解データに照らし合わせて計算されたものである．これによると，再現率が 0.1 前後では適合率が 0.90 前後となっており，極めて高いものとなっている一方，システムの全回答の再現率 0.508 に近い点，例えば，再現率 0.4 前後の点では提案手法の適合率に比して，わずかな適合率の向上（0.05 前後）しか見られず，また，もうすこし離れたデータポイント（例えば，再現率 0.3 前後）までの適合率の改善具合も極めてなだらかである．

この評価はあくまで現状のシステムの出力結果のみをランキングしているため，確定的なことは言えないが，前述したように学習データは現在のシステムが出力できない回答に関するものも含まれていないことも考え合わせると，仮に現在のシステムをより大量の回答を出力するように改変し，ランキングによる足切りをおこなったとしても，例えば，再現率 0.5 前後の部分での適合率向上はきわめて小さなものになる可能性が高いと考えられる．これは再現率を重視するという立場とは相容れないものであり，今回の提案システムにはランキングによるフィルタリングは導入しなかった．

今後システムを改善できる可能性があるとすれば，今後さらに学習データを増やしていくことが重要であるが，現在でも約 5,000 件という少なくない量の学習データを利用していること，また，あるドメインに特化した学習データを大量に作ることは現段階で

は望ましくないと考えられることから，少なくともランキング手法の導入については慎重に検討する必要があると考えている．あるドメインに特化した学習データのアノテーションをクラウドソーシングなどで行い，質問応答の精度を高めるといったシナリオは魅力的に見えるかもしれない．しかし，そうしたシナリオを実現するためには， NER の場合と同様にシステム全体としての最適化の枠組みなどが必要だということかもしれず，これも慎重に検討する必要があると考えている．

4.9 まとめ

本章では，Twitter などのマイクロブログにおいてリアルタイムに書き込まれる大量の書き込みを，既存メディアを補完する情報として捉え，既存メディアでは取り扱うことができない局所的で詳細な情報をマイクロブログから取得する質問応答技術を利用した情報取得手法を検討した．提案手法では，様々な質問に対して回答を提示できるようにするために，構文パターンの言い換えに基づく質問文の拡張を行うことを提案した．また，地名の補完処理により情報が不足している tweet からでも回答を取得を可能とすることや，得られた大量の回答を地理的，時間的，意味的観点から分類する回答提示手段を検討した．提案手法の質問応答機能の有効性を示すため，東日本大震災時の tweet データを利用した性能評価を実施し，大量の回答がある場合にでも再現率を維持しつつ，適合率 0.608 が得られたことを報告した．エラー解析の結果，質問解析における複合語処理の問題や要望，疑問，仮定を含む tweet の特定の必要性が明らかになった．

本研究の成果をハイブリッドキャストなどの放送と通信の連携の仕組みに応用することで，放送コンテンツで紹介されている情報が不足しているときや，その内容に疑問があったときなどに，受信機上で自然言語での質問を行うとマイクロブログの局所的で膨大な情報から必要な情報を取得できるようになる．また災害時などの場合には，放送コンテンツで広範囲な情報を把握しつつ，マイクロブログに放送で伝えられている情報以外の情報や局所的なローカル情報を把握することができる．必要な情報を必要な人に届けるためには，マイクロブログなどの通信コンテンツの膨大な情報を俯瞰して閲覧できる手段が必要である．また，たとえ情報を必要としている人が少数であっても，回答を提示できるようにすべきである．こうした点に鑑み，マイクロブログ上の全体像を把握するとともに，ロングテール部分に存在する回答を欲しているような質問にも応えることができる情報の取りこぼしのないアクセス手段の構築を今後もさらに進めていく予定である．

第５章 結論

本論文では，放送と通信の連携により登場する新しいメディア（ハイブリッドメディア）において，放送コンテンツを基点に関連のある通信上のコンテンツを取得するためのメディアアクセス技術を提案した．ハイブリッドメディアの普及にあたって，過去の番組を蓄積した番組アーカイブやTwitterなどのマイクロブログなどの通信上のコンテンツに簡単な操作でアクセスできる技術への期待は大きい．これまで，通信上の新しいサービスを自由に使いこなしているのは，コンピュータや情報アクセスに関心のある一部のユーザだけであったが，放送と通信が連携するハイブリッドメディアにおいては，誰もがこれらのサービスをシームレスに利用できるバリアフリーなアクセス手段が必要である．本研究では，これらの通信上のコンテンツへのアクセスする際の現状の課題を確認し，自然言語処理技術を用いて，これらの課題を解決するための効果的なメディアアクセス手法を検討した．さらに，提案手法を用いた実験システムを構築し，その有効性を示した．各章での成果をまとめると，以下の通りである．

第１章では本研究の背景として，これまでの放送メディアの進展について整理し，近年出現した番組アーカイブなどの蓄積メディアやTwitterを初めとするマイクロブログの現状についてまとめた．これらの情報に効率的にアクセスできる手段を提供することの社会的なニーズについて述べた．

第２章では，番組アーカイブにおいて，番組の内容を自然言語で記述している番組概要を用いて，複合語や拡張固有表現を考慮したBM25に基づく関連度による番組検索手法を提案し，その有効性を実験により示した．この研究の成果は，EnVisionと呼ばれる言語情報と映像情報を利用した番組検索用のライブラリに採用され，現在，オンデマンドサービスやアーカイブの検索，映像素材の検索などに幅広く利用されている．また，Wikipediaなどの誰もが利用可能なオープンデータの更新履歴を利用した番組検索結果のリランキングについても検討し，番組アーカイブ以外の外部の情報による番組検索結果の改善を行うことの重要性を示した．

第３章では，エンティティ間の関係を考慮して内容をより反映した番組検索を実現するため，番組概要から人物表現間の関係を示す情報である関係グラフを取得する手法を提案した．評価実験により，番組概要からの関係グラフ取得の性能について検証し，そ

の有効性を示した．また，関係グラフを利用することにより構文情報から直接取得できない人物表現間の関係を取得できことを明らかにし，エンティティ間の関係を考慮した番組検索への応用の可能性を示した．さらに，関係グラフを利用した番組の登場人物相関図を生成するシステムを構築し，番組内容の可視化や特定シーンのアクセス手段などのユーザインタフェースへの応用の可能性を示した．

第４章では，マイクロブログをテレビなどの既存メディアが伝えきない情報を補完する手段としてとらえ，マイクロブログが伝える膨大な情報を効果的に取得し提示する情報アクセス手段について検討した．大規模な含意パターンデータベースを用いて質問文を様々な表現に言い換えることにより，多くの人々が種々の表現で記述したマイクロブログの書き込みから回答となる情報を網羅的に取得する手法を開発した．また回答の意味的分類，地理的分類，時間的分類を行い，マイクロブログの空間でどのような情報が出回っているのかを俯瞰的に提示することを可能とした．東日本大震災時の実際のtweetデータを対象とした評価実験の結果，大量のマイクロブログに対しても高い再現率で回答を取得できることを示した．これにより，記述されている書き込みが少数のもの，つまり通常の検索エンジンでは無視されがちなデータのロングテール部分からも情報を取得できる可能性を示した．

本研究の成果は，放送と通信の連携に伴い登場した新しいテレビ視聴環境における課題を確認し，その課題解決のために大規模な番組アーカイブや膨大な数のマイクロブログから必要な番組や情報を取得するメディアアクセス技術を提案するとともに，その有効性を検証し実用化への道筋を示したことにある．これにより，放送と通信の連携・融合時代の放送メディアへのアクセス方法の確立，そしてハイブリッドメディアの普及に向けて大きな貢献を行えたと考えている．

本論文では，番組概要を用いた番組検索に関連する技術，マイクロブログを対象とした質問応答による情報取得に関する技術を取り扱ったが，両者ともに小規模のテキストを対象としており，固有表現や名詞の意味クラス，構文解析結果を利用した名詞間の関係抽出，共参照解析など共通する技術に基づいている．そのため，例えば，番組検索技術の応用として，マイクロブログの膨大な情報から番組概要に関連する書き込みを取得するようなアプリケーションの開発が考えられる．また放送で送信される番組の関連情報を質問応答の対象に加えることにより，質問応答の回答を番組概要から取得することが考えられる．ただし，マイクロブログの書き込みと放送コンテンツ制作のプロが作成する番組概要ではその文章の性質が大きく異なるため，それぞれの手法を適用する際には各データに対するチューニングを行うなどの注意が必要である．また本研究で取り上げた番組アーカイブやマイクロブログは通信で扱われるごく一部のコンテンツであるため，今後，インターネット上に溢れるテキスト，動画，音声などのあらゆる通信コン

テンツを，放送コンテンツとひも付けていくことが，放送と通信の連携・融合をより一層進展させていくための課題と言える．

2013年，放送と通信の連携・融合を推進する新しい放送方式であるハイブリッドキャストの技術仕様が公開され，すでに新方式に対応した受信機が発売されている．また試験的ではあるがいくつかの放送と通信を連携したアプリケーションが稼働し始めている．また，通信コンテンツに重点をおいたスマートテレビも様々なメーカーから発売されて普及しはじめており，これまで以上にハイブリッドメディアで利用できる通信コンテンツも充実してくるであろう．これにより，放送から得られる番組の情報と通信上のコンテンツとを密接に連携させた多様なアプリケーションが受信機上で利用できるようになる．このような放送と通信の連携・融合環境下では，これまでの放送メディアでは予想もできなかった新しいサービスが出現してくるであろう．本研究の成果が放送と通信の連携・融合時代の新しいメディアの発展の一助となれば幸いである．

謝辞

本論文は，様々な方々のお力添えのもとに完成しました．

はじめに，本論文の主査である総合研究大学院大学／国立情報学研究所の山田誠二教授におかれましては，論文をまとめるにあたり，全般にわたるご指導とご鞭撻を賜りました．先生のご指導により本論文を完成させることができました．厚く御礼申し上げます．本研究の主任指導教官の東京大学／国立情報学研究所の相澤彰子教授におかれましては，ご多忙のなか，研究の方向性を含めた多岐に渡る種々の相談に乗っていただきました．先生のご指導なくしては研究を遂行し，本論文を完成させることはできなかったと感じております．心より御礼申し上げます．

本論文をまとめるにあたり，論文審査委員の先生から多大なご指導とご鞭撻を賜りました．東京大学／国立情報学研究所の佐藤真一教授，総合研究大学院大学／国立情報学研究所の相原健郎准教授，宮尾祐介准教授におかれましては，博士論文の中間発表，予備審査，本審査において多くの建設的なご指導をいただき，本論文をより良いものへと導いていただきました．深く御礼申し上げます．

本研究において番組検索に関する研究の機会を与えてくださり，ご指導，ご鞭撻を賜った日本放送協会の多くの皆様に心より感謝いたします．特に，久保田啓一技師長（前放送技術研究所所長），放送技術研究所の藤沢秀一所長のご理解とご鞭撻は，本研究を遂行し，本論文をまとめるうえで大きな支えとなりました．厚く御礼申し上げます．伊藤崇之氏，八木伸行氏，柴田正啓氏，藤井真人氏，田中英輝氏におかれましては，研究へのご指導をいただくとともに，様々なサポートをいただきました．厚く御礼申し上げます．住吉英樹氏，加藤直人氏，山田一郎氏，佐野雅規氏，小早川健氏，熊野正氏，宮崎勝氏，後藤功雄氏，高橋正樹氏，望月菊佳氏，河合吉彦氏，古宮弘智氏，美野秀弥氏をはじめとした放送技術研究所の皆様には，日頃の議論，ご指導いただきましたことを深く感謝いたします．

本研究において関係抽出に関する研究のご指導いただきましたニューヨーク大学の関根聡准教授に心より感謝いたします．ニューヨーク大学滞在中に先生から御教示いただいた情報抽出に関する知見は，本論文をまとめるうえで大きな力となりました．心から感謝いたします．

本研究において質問応答に関する研究のご指導，ご鞭撻を賜った情報通信研究機構の多くの皆様に心より感謝いたします．特に，情報分析研究室の鳥澤健太郎室長，情報配信基盤研究室の大竹清敬室長のご理解とご指導は，本論文をまとめるうえで大きな支えとなりました．厚く御礼申し上げます．Stijn De Saeger 氏，風間淳一氏，橋本力氏，呉

鍾勲氏，川田拓也氏，István Varga 氏，王軼謳氏，田仲正弘氏，Julien Kloetzer 氏，佐野大樹氏，Pham Quang Nhat Minh 氏，水野淳太氏をはじめとした情報通信研究機構の情報分析研究室および情報配信基盤研究室の皆様には，日頃のご討論，ご指導を賜りましたこと，深く感謝いたします．

最後に，これまで暖かく応援してくれた家族に感謝します．

参考文献

[青島 13] 青島傳隼, 坂本翼, 横山昌平, 福田直樹, 石川博, “文脈的なつながりを考慮したツイート群の効果的な抽出・提示手法の実現,” 情報処理学会論文誌, 6(2), pp.61-84 (2013)

[Ali04] Kamal Ali, Wijnand van Stam, Making Show Recommendations Using a Distributed Collaborative Filtering Architecture. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2004), pp.394-401 (2004).

[Ando03] Akio Ando, Toru Imai, Akio Kobayashi, Shinich Homma, Jun Goto, Nobumasa Seiyama, Takeshi Misima, Takeshi Kobayakawa, Shoei Sato, Kazuo Onoe, Hiroyuki Segi, Atushi Imai, Atushi Matsui, Akira Nakamura, Hideki Tanaka, Eiichi Miyasaka and Haruo Isono: “Simultaneous Subtitling System for Broadcast News Programs with a Speech Recognizer”, IEICE Transactions on Information and Systems, Vol.E86-D, No.1, pp.15-25 (2003).

[Aramaki06] Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe, “Patient Status Classification by using Rule based Sentence Extraction and BM25-kNN based Classifier,” In i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data 2006 (2006).

[ARIB13] ARIB, “デジタル放送に使用する番組配列情報”, ARIB STD-B10 5.2 版 (2013).

[馬場 07] 馬場こずえ, 藤井敦, “小説テキストを対象とした人物情報の抽出と体系化,” 言語処理学会第 14 回年次大会, pp.574-577 (2007).

[Beeferman00] Doug Beeferman and Adam Berger, “Agglomerative Clustering of Search Engine Query Log,” The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2000) (2000).

[Bloom70] Burton H. Bloom, “Space/time Trade-offs in Hash Coding with Allowable Errors,” Communications of the ACM, 13-7, pp.422-426 (1970).

[Buscaldi06] Davide Buscaldi and Paolo Rosso, “Mining Knowledge from Wikipedia for the Question Answering Task,” In Proceedings of the International Conference on Language Resources and Evaluation (LREC2006), pp.727-730 (2006).

[Cheng10] Zhiyuan Cheng, James Caverlee, and Kyumin Lee, “You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users,” In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM2010), pp.759-768 (2010).

[Crammer01] Koby Crammer and Yoram Singer, “On the Algorithmic Implementation of Multi-class Kernel-based Vector Machines, ” Journal of Machine Learning Research, pp.265-292 (2001) .

[De Saeger09] Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda and Masaki Murata, “Large Scale Relation Acquisition using Class Dependent Patterns,” In Proceedings of the IEEE International Conference on Data Mining (ICDM2009) , pp.764-769 (2009).

[Deerwester88] Scott Deerwester, “Improving Information Retrieval with Latent Semantic Indexing,” In Proceedings of the 51st Annual Meeting of the American Society for Information Science, Vol.25, pp36–40 (1988).

[江原 07] 江原学, “放送映像アーカイビング,” 映像情報メディア学会誌, 61-11, pp1567-1570 (2007).

[Eirinaki03] Magdalini Eirinaki and Michalis Vazirgiannis, “Web Mining for Web Personalization,” ACM Transactions on Internet Technology, Vol. 3, No. 1, pp.1–27 (2003)

[Ferrucci10] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer and Chris Welty, “Building Watson: An Overview of the DeepQA Project,” AI Magazine, 31 (3), pp.59-79. (2010).

[Fleiss71] Joseph Fleiss, “Measuring Nominal Scale Agreement among Many Raters,” Psychological Bulletin, 76 (5), pp.378-382 (1971).

[藤沢 13] 藤沢寛，出葉義治，武智秀，北里直久，藤吉靖浩，会津宏幸，小澤由佳，松村欣司，川上皓平，原田聡，柳内啓司，坂本典哉，廣野二郎, “ハイブリッドテレビサービスの標準化動向 ～新しい放送通信連携サービスを実現するハイブリッドキャストについて～”, 映像情報メディア学会誌，67(5), pp355-360 (2013).

[Gao12] Wei Gao, Zhongyu Wei and Kam-Fai Wong, “Microblog Search and Filtering with Time Sensitive Feedback and Thresholding based on BM25,” In Proceedings of the 21st Text Retrieval Conference (TREC2012) (2012).

[Goto04] Jun Goto, Kazuteru Komine, Masaru Miyazaki, Yeun-Bae Kim and Noriyoshi Uratani, “A Spoken Dialogue Interface for TV Operations Based on Data Collected by Using WOZ Method,” IEICE Transactions on Information and Systems, Vol.E87-D, No.6, pp.1397-1404 (2004).

[Goto06] Jun Goto, Masaru Miyazaki, Takeshi Kobayakawa, Nobuyuki Hiruma and Noriyoshi Uratani, “A TV Agent System that Integrates Knowledge and Answers Users’ Questions,” In Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI2006), pp.300-302 (2006).

[後藤 12] 後藤淳, 八木伸行, 柴田正啓, “番組登場人物抽出装置及び番組登場人物抽出プログラム,” 特許第 4934090 号 (2012).

[Grishman96] Ralph Grishman and Beth Sundheim, “Message Understanding Conference-6: A Brief History,” In Proceedings of the 16th Conference on Computational Linguistics (COLING’96), pp.466- 471 (1996) .

[Gurini12] Davide Feltoni Gurini and Fabio Gasparetti, “TREC Microblog 2012 Track: Real-Time Algorithm for Microblog Ranking Systems,” In Proceedings of the 21st Text Retrieval Conference (TREC2012) (2012).

[Harris54] Zellg S. Harris, “Distributional Structure,” Word, 10 (23), pp.142-146 (1954).

[Hasegawa04] Takaaki Hasegawa, Satoshi Sekine and Ralph Grishman, “Discovering Relations among Named Entities from Large Corpora,” In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004), pp.415-422 (2004).

[橋本 08] 橋本泰一, 乾孝司, 村上浩司, “拡張固有表現タグ付きコーパスの構築,” 自然言語処理研究会報告, 2008(113), pp.113-120 (2008)

[Hashimoto12] Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh and Jun’ichi Kazama, “Excitatory or Inhibitory: A New Semantic Orientation Extracts Contradiction and Causality from the Web,” In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL2012), pp.619-630. (2012).

[橋元 13] 橋元良明, “調査から見た被災地におけるメディアの役割, ” マス・コミュニケーション研究 (82), pp.19-34 (2013).

[飯田 05] 飯田龍, 乾健太郎, 松本裕治, 関根聡, “最尤先行詞候補を用いた日本語名詞句同一指示解析,” 情報処理学会論文誌, Vol.46, No.3, pp.831-844 (2005).

[Iida09] Ryu Iida, Kentaro Inui and Yuji Matsumoto, “Capturing Salience with a Trainable Cache Model for Zero-anaphora Resolution,” In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP2009), pp.647-655 (2009).

[Iida11] Ryu Iida and Massimo Poesio, “A Cross-Lingual ILP Solution to Zero Anaphora Resolution,” In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL2011-HLT), pp.804-813 (2011).

[Jarvelin02] Kalervo Jarvelin and Jaana Kekalainen, “Cumulated gain-based evaluation of IR techniques,” ACM Transactions on Information Systems, 20(4), pp.422–446 (2002).

[神代 08] 神代大輔, 高村大也, 奥村学, “物語テキストにおけるキャラクタ関係図自動構築,” 言語処理学会第 14 回年次大会講演論文集, pp.380-383 (2008).

[川田 13] 川田拓也，大竹清敬，後藤淳，鳥澤健太郎, “災害対応質問応答システム構築

に向けた質問・回答コーパスの構築,” 言語処理学会第 19 回年次大会発表論文集, pp.480-483 (2013).

[Kawahara06] Daisuke Kawahara and Sadao Kurohashi, “Case Frame Compilation from the Web using High-Performance Computing,” In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006), pp.1344-1347 (2006).

[Kazama08] Junichi Kazama and Kentaro Torisawa, “Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations,” In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL2008-HLT), pp.407-415 (2008).

[風間12] 風間淳一, Stijn De Saeger, 鳥澤健太郎, 後藤淳, Istvan Varga, “災害時情報への質問応答システムの適用の試み,” 言語処理学会第18回年次大会講演論文集, pp.903-906 (2012).

[Kim94] Yeun-Bae Kim and Terimasa Ehara, “A Method of Partitioning of Long Japanese Sentences with Subject Resolution in J/E Machine Translation,” In Proceedings of International Conference on Computer Processing of Oriental Languages, pp.467-473 (1994).

[北口 13] 北口沙也香, 宮西大樹, 関和広, 上原邦昭, “マイクロブログ文書の選択による対話的な災害情報検索システム,” 言語処理学会第 19 回年次大会講演論文集, pp.240-243 (2013).

[Kloetzer12] Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Motoki Sano, Jun Goto, Chikara Hashimoto and Jong Hoon Oh, “Supervised Recognition of Entailment between Patterns,” 言語処理学会第18回年次大会講演論文集, pp.431-434 (2012).

[Lafferty01] John Lafferty, Andrew McCallum and Fernando Pereira, “Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data,” In Proceedings of the 18th International Conference on Machine Learning (ICML2001), pp.282-289 (2001).

[Liu13] Xiaohua Liu, Furu Wei, Shaodian Zhang and Ming Zhou, “Named Entity Recognition for Tweets,” ACM Transactions on Intelligent Systems and Technology, 4 (1), pp.1-15 (2013).

[Matsuo06] Yutaka Matsuo, Junichiro Mori and Masahiro Hamasaki, “POLYPHO-NET: An Advanced Social Network Extraction System from the Web,” In Proceedings of the 15th International Conference on World Wide Web, pp.397-406 (2006).

[Mika05] Peter Mika, “ Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks,” Journal of Web Semantics,” vol.3, no.2, pp.211-223 (2005).

[Miura08] Kikuka Miura, Ichiro Yamada, Hideki Sumiyoshi and Nobuyuki Yagi, “Identification of Names and Actions of Principal Objects in TV Program Segments Using Closed Captions,” Semantic Computing 2(2), pp.191-206 (2008).

[Miyanishi12] Taiki Miyanishi, Kazuhiro Seki and Kuniaki Uehara, “TREC 2012 Microblog Track Experiments at Kobe University,” In Proceedings of the 21st Text Retrieval Conference (TREC2012) (2012).

[Miyazaki08] Masaru Miyazaki, Masahiro Shibata and Nobuyuki Yagi, “Baseball Digest Production System using Inductive Logic Programming,” In Proceedings of the 10th IEEE International Symposium on Multimedia, pp.503–508 (2008).

[Mizoguchi07] Yumiko Mizoguchi, Toshiaki Nakamoto, Kazuma Asakawa, Shinichi Nagano, Masumi Inaba and Takahiro Kawamura, “TV Navigation Agent for Measuring Semantic Similarity between Documents,” In Proceedings of 3rd International Workshop on Agents and Web Services in Distributed Environments (2007).

[森 08] 森辰則, 福本淳一, 加藤恒昭, 桝井文人, 佐々木裕, Chen Hsin-Hsi, Chen Kuang-hua, Lin Chuan-Jie, 三田村照子, Nyberg Eric, 神門典子, “NTCIR における質問応答技術の評価と今後の展望,” 情報処理学会研究報告, 自然言語処理研究会報告, 2008(4), pp.43-50 (2008).

[村上 07] 村上知子, “AV 機器利用者に対する放送コンテンツの推薦,” 情報処理学会誌, 48-9, pp.984-988 (2007).

[Negri09] Matteo Negri and Milen Kouylekov, “Question Answering over Structured Data: an Entailment-based Approach to Question Analysis,” International Conference on Recent Advances in Natural Languege Processing, pp.305-311 (2009).

[Neubig11] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara and Koji Murakami, “Safety Information Mining -What Can NLP Do in a Disaster,” In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP2011), pp.965-973 (2011).

[Pariser11] Eli Priser, “Filter Bubble:What the Internet is Hiding from You,” Viking (2011).

[Ritter11] Alan Ritter, Sam Clark, Mausam and Oren Etzioni, “Named Entity Recognition in Tweets: An Experimental Study,” In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP2011), pp.1524-1534 (2011).

[Robertson99] Stephen E. Robertson and Steve Walker, “Okapi/Keenbow at TREC-8, ” In Proceedings of the 8rd Text Retrieval Conference (TREC1999) (1999).

[澤井 10] 澤井里枝, 有安香子, 藤沢寛, 金次保明, “SNS を利用した協調フィルタリングによる番組推薦手法,” 情報処理学会研究報告, Vol.2010-DBS-151(43), pp.1-8 (2010).

[Sekine00] Satoshi Sekine and Hitoshi Isahara, “IREX: IR and IE Evaluation Project in Japanese,” In Proceedings of the 2rd International Conference on Language Resources and Evaluation (LREC2000), pp.1475-1480 (2000).

[Sekine08] Satoshi Sekine, “Extended Named Entity Ontology with Attribute Information,” In Proceedings of the International Conference on Language Resources and Evaluation (LREC2008) (2008).

[柴田 07] 柴田知秀, 黒橋禎夫, “言語情報と映像情報を統合した隠れマルコフモデルに基づくトピック推定,” 情報処理学会論文誌, Vol.48, No.6, pp.2129-2139 (2007).

[Shinzato11] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara and Sadao Kurohashi, “TSUBAKI: An Open Search Engine Infrastructure for Developing Information Access Methodology,” Journal of Information Processing, Vol.52, No.12, pp.216-227 (2011).

[住吉 03] 住吉英樹, 山田一郎, 村崎康博, 金淵培, 八木伸行, 柴田正啓, “新しい教育放送サービスのための情報検索システム,” 映像情報メディア学会論文誌, Vol.57, No.2, pp.253-261 (2003).

[隆 01] 隆朋也, 渡辺尚, 樽口秀昭, “履歴情報を用いた TV 番組選択支援エージェント,” 情報処理学会論文誌, 42(12), pp.3130-3143 (2001).

[内元 00] 内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均, “最大エントロピーモデルと書き換え規則に基づく固有表現抽出,” 自然言語処理, 7(2), pp.63-90 (2000).

[Walker94] Marilyn Walker, Masayo Iida and Sharon Cote, “Japanese Discourse and the Process of Centering,” Computational Linguistics, Vol. 20, No. 2, pp.193-233 (1994).

[Wang08] Wang Xiaowei，Jiang Longbin and Ma Jialin， Jiangyan，“Use of NER Information for Improved Topic Tracking,” In Proceeding of 8th International Conference on Intelligent Systems Design and Applications (ISDA2008), vol.3, pp165-170 (2008).

[山田 02] 山田寛康, 工藤拓, 松本裕治, “Support Vector Machine を用いた日本語固有表現抽出,” 情報処理学会論文誌, Vol.43, No.1, pp.44-53 (2002).

[山田 06] 山田一郎, 佐野雅規, 住吉英樹, 柴田正啓, 八木伸行, “アナウンサーと解説者のコメントを利用したサッカー番組セグメントメタデータ自動生成,” 電子情報通信学会論文誌, Vol.J89-D, No.10, pp.2328-2337 (2006).

[山田 07] 山田一郎, 三浦菊佳, 河合吉彦, 住吉英樹, 八木伸行, 奥村学, 徳永健伸, “大域的な文章構造の類似性を利用したクローズドキャプション中の定型的な文章区間の抽出, ” 電子情報通信学会論文誌, Vol.J90-D, No.9, pp.2624-2633 (2007).

[Yamada09] Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond and Asuka Sumida, “Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures,” In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP2009), pp.1172-1181 (2009).

[山口 10] 山口瑶子, 瀬々潤, “Web 閲覧履歴を用いた TV 番組推薦システム,” 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), A3-2 (2010).

[Yamangil08] Elif Yamangil and Rani Nelken, “Mining Wikipedia Revision Histories for Improving Sentence Compression,” In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL08-HLT), pp.137-140 (2008).

[吉村 12] 吉村健, “しゃべってコンシェルと言語処理,” 情報処理学会研究報告, Vol.2012-SLP-93 (4), pp.1-6 (2012).

研究業績

本論文を構成する論文

【学術論文】

1. 後藤淳, 大竹清敬, Stijn De Saeger, 橋本力, Julien Kloetzer, 川田拓也, 鳥澤健太郎, 質問応答に基づく対災害情報分析システム, 自然言語処理, Vol.20, No.3, pp.367-404 (2013).

【国際会議】（査読有）

2. Jun Goto, Hideki Sumiyoshi, Masaru Miyazaki, Hideki Tanaka and Akiko Aizawa, “Relevant TV Program Retrieval using Broadcast Summaries,” In Proceedings of the 15th Iinternational Conference on Intelligent Uuser Interfaces (IUI2010), pp.411-412 (2010).
3. Jun Goto, Nobuyuki Yagi, Akiko Aizawa and Satoshi Sekine, “Method for Automatically Generating Networks of Personal Relationships from Story Summaries,” Common Sense and Intelligent User Interfaces 2009 (CSIUI2009) (2009).

【全国大会，研究会】

4. 後藤淳, 宮崎勝, 田中英輝, 相澤彰子, “更新履歴による注目度を利用した番組検索結果のリランキング,” 第10回情報科学技術フォーラム（FIT2011）, D-025(2011).
5. 後藤淳, 宮崎勝, 田中英輝, 相澤彰子, “Wikipediaの更新履歴を利用した番組検索”, 映像情報メディア学会年次大会, 3-4 (2010).
6. 後藤淳, 住吉英樹, 宮崎勝, 田中英輝, 相澤彰子, “閲覧中のWebコンテンツを起点とした関連番組検索”, 映像情報メディア学会冬季大会, 2-1 (2009).
7. 後藤淳, 住吉英樹, 宮崎勝, 柴田正啓, 相澤彰子, “視聴中の番組を起点とした関連番組検索”, 第8回情報科学技術フォーラム（FIT2009）, D-006 (2009).
8. 後藤淳, 八木伸行, 相澤彰子, 関根聡, “照応解析を利用した放送番組からの登場人物の相関図生成,” 人工知能学会第22回全国大会, 2G2-04 (2008).
9. 後藤淳, 松井淳, 八木伸行, 相澤彰子, 関根聡, “マルチモーダル情報を用いた放送番組からの人物相関図生成”, 電子情報通信学会全国大会, D-5-7 (2008).

研究業績リスト

学術論文

1. 後藤淳, 大竹清敬, Stijn De Saeger, 橋本力, Julien Kloetzer, 川田拓也, 鳥澤健太郎, 質問応答に基づく対災害情報分析システム, 自然言語処理 Vol.20, No.3, pp.367-404 (2013).
2. Jun Goto, Kazuteru Komine, Masaru Miyazaki, Yeun-Bae Kim and Noriyoshi Uratani: "A Spoken Dialogue Interface for TV Operations Based on Data Collected by Using WOZ Method," IEICE Transactions on Information and Systems, Vol.E87-D, No.6, pp.1397-1404 (2004).
3. Akio Ando, Toru Imai, Akio Kobayashi, Shinich Homma, Jun Goto, Nobumasa Seiyama, Takeshi Misima, Takeshi Kobayakawa, Shoei Sato, Kazuo Onoe, Hiroyuki Segi, Atushi Imai, Atushi Matsui, Akira Nakamura, Hideki Tanaka, Eiichi Miyasaka and Haruo Isono, "Simultaneous Subtitling System for Broadcast News Programs with a Speech Recognizer", IEICE Transactions on Information and Systems, Vol.E86-D, No.1, pp.15-25 (2003).
4. 安藤彰男, 今井亨, 小林彰夫, 本間真一, 後藤淳, 清山信正, 三島剛, 小早川健, 佐藤庄衛, 尾上和穂, 世木寛之, 今井篤, 松井淳, 中村章, 田中英輝, 都木徹, 宮坂栄一, 磯野春雄, "音声認識を利用した放送用ニュース字幕制作システム", 電子情報通信学会論文誌, Vol.E84-D-II, no.6, pp877-887 (2002).

国際会議（査読有）

1. Kiyonori Ohtake, Jun Goto, Stijn De Saeger, Kentaro Torisawa, Junta Mizuno and Kentaro Inui, "NICT Disaster Information Analysis System," In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), pp.29-32 (2013).
2. Jun Goto, Hideki Sumiyoshi, Masaru Miyazaki, Hideki Tanaka and Akiko Aizawa, "Relevant TV Program Retrieval using Broadcast Summaries," In Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI 2010), pp.411-412 (2010).
3. Hideki Sumiyoshi, Masanori Sano, Jun Goto, Takahiro Mochizuki, Masaru Miyazaki, Mahito Fujii, Masahiro Shibata, and Nobuyuki Yagi, "CurioView: TV Recommendations

Related to Content Being Viewed," IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, mm2010-21 (2010).

4. Jun Goto, Nobuyuki Yagi, Akiko Aizawa and Satoshi Sekine, "Method for Automatically Generating Networks of Personal Relationships from Story Summaries," Common Sense and Intelligent User Interfaces 2009 (CSIUI2009) (2009).
5. Masaru Miyazaki, Takeshi Kobayakawa, Jun Goto, Nobuyuki Hiruma and Nobuyuki Yagi: "OWL Metadata Framework for a Baseball Q&A System," In Proceedings of the 5th International Semantic Web Conference (ISWC2006) (2006).
6. Jun Goto, Masaru Miyazaki, Takeshi Kobayakawa, Nobuyuki Hiruma and Noriyoshi Uratani, "A TV Agent System that Integrates Knowledge and Answers Users' Questions," In Proceedings of the 11th international conference on Intelligent user interfaces (IUI2006), pp.300-302 (2006).
7. Key-Sun Choi, Jae-Ho Kim, Masaru Miyazaki, Jun Goto and Yeun-Bae Kim, "Question-Answering Based on Virtually Integrated Lexical," The 6th International Workshop on Information Retrieval with Asian Languages, pp.168-175 (2003).
8. Jun Goto, Kazuteru Komine, Yeun-Bae Kim and Noriyoshi Uratani, "A Television Control System based on Spoken Natural Language Dialogue," In Proceedings of Human-Computer Interaction INTERACT 2003, pp.765-768 (2003).
9. Jun Goto, Yeun-Bae Kim, Masaru Miyazaki, Kazuteru Komine and Noriyoshi Uratani, "A Spoken Dialogue Interface for TV Operations Based on Data Collected by Using WOZ Method," In Proceedings of 41st Annual Meeting of the Association for Computational Linguistics (ACL2003), Vol.2, pp.101-104 (2003).

研究会，全国大会等

1. Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Motoki Sano, Jun Goto, Chikara Hashimoto, Jong Hoon Oh, "Supervised Recognition of Entailment Between Patterns," 言語処理学会第 18 回年次大会講演論文集, pp.431-434 (2012).
2. 住吉英樹, 古宮弘智, 望月貴裕, 後藤淳, 佐野雅規, 藤井真人, "関連映像検索システム EN-Vision 実験サイト," 映像情報メディア学会冬季大会, 3-5-1 (2011).
3. 後藤淳, 宮崎勝, 田中英輝, 相澤彰子, "更新履歴による注目度を利用した番組検索結果のリランキング," 第 10 回情報科学技術フォーラム（FIT2011）, D-025 (2011).
4. 古宮弘智, 住吉英樹, 後藤淳, 佐野雅規, 藤井真人, "推薦番組の選択行動のパターン化に関する検討", 第 10 回情報科学技術フォーラム（FIT2011）, D-001 (2011).

5. 後藤淳, 宮崎勝, 田中英輝, 相澤彰子, “Wikipedia の更新履歴を利用した番組検索,” 映像情報メディア学会年次大会, 3-4 (2010).
6. 後藤淳, 住吉英樹, 宮崎勝, 田中英輝, 相澤彰子, “閲覧中の Web コンテンツを起点とした関連番組検索,” 映像情報メディア学会冬季大会, 2-1 (2009).
7. 後藤淳, 住吉英樹, 宮崎勝, 柴田正啓, 相澤彰子, “視聴中の番組を起点とした関連番組検索,” 第 8 回情報科学技術フォーラム（FIT2009）, D-006 (2009).
8. 宮崎勝, 住吉英樹, 後藤淳, 藤井真人, 柴田正啓, “スポーツニュースの言語情報を利用したプロ野球映像推薦システムの試作,” 第 8 回情報科学技術フォーラム（FIT2009）, F-008 (2009).
9. 佐野雅規, 住吉英樹, 後藤淳, 望月貴裕, 宮崎勝, 三浦菊佳, 河合吉彦, 高橋正樹, 三須俊枝, 松井淳, クリピングデル サイモン, 藤井真人, 柴田正啓, 八木伸行, “番組を推薦するテレビ CurioView,” 第 8 回情報科学技術フォーラム（FIT2009）, K-049 (2009).
10. 後藤淳, 宮崎勝, 柴田正啓, 相澤彰子, “TV Searchbar: Web からの放送コンテンツの参照,” 電子情報通信学会総合大会, D-41-13 (2009).
11. 住吉英樹, 柴田正啓, 藤井真人, 後藤淳, 山田一郎, 望月貴裕, 松井淳, 三須俊枝, 宮崎勝, 高橋正樹, 河合吉彦, 三浦菊佳, 八木伸行, “CurioView：情報検索を活用した新しい視聴スタイルの提案,” 映像情報メディア学会年次大会, 7－5 (2008).
12. 後藤淳, 八木伸行, 相澤彰子, 関根聡, “照応解析を利用した放送番組からの登場人物の相関図生成,” 人工知能学会第22回全国大会, 2G2-04 (2008).
13. 後藤淳, 松井淳, 八木伸行, 相澤彰子, 関根聡, “マルチモーダル情報を用いた放送番組からの人物相関図生成,” 電子情報通信学会総合大会, D-5-7 (2008).
14. 後藤淳, 八木伸行, 関根聡, “番組検索のための登場人物の関係抽出,” 第 6 回情報科学技術フォーラム（FIT2007）, E-049 (2007).
15. 後藤淳, 宮崎勝, 小早川健, 比留間伸行, 浦谷則好, “知識を統合しユーザの疑問に答える TV エージェント,” 第 4 回情報科学技術フォーラム（FIT2005）, K-74 (2005).
16. 後藤淳, 宮崎勝, 比留間伸行, 浦谷則好 “番組に関するユーザの疑問に答える TV エージェントシステム,” 電子情報通信学会総合大会, A-15-22 (2005).
17. 小峯一晃, 澤畠康仁, 後藤淳, 小早川健, 浦谷則好, “視線情報を利用した番組選択インタフェースの開発,” 情報処理学会研究報告, vol.2004, no.115, p.107-114 (2004).
18. 宮崎勝, 後藤淳, 小峯一晃, 浦谷則好, “番組情報獲得システムにおけるラッパエージェント構築法,” 第 4 回情報科学技術フォーラム（FIT2004）, B-024 (2004).
19. 小峯一晃, 澤畠康仁, 後藤淳, 浦谷則好, “視線情報を利用したテレビ用ユーザインタ

フェースの開発,” 第 3 回情報科学技術フォーラム（FIT2004）, K-074 (2004).
20. 小峯一晃, 森田寿哉, 後藤淳, 浦谷則好, “音声インタフェースによる番組選択操作時の発話内容分析,” ヒューマンインタフェースシンポジウム, pp.631-634 (2002).
21. 後藤淳, 小峯一晃, 森田寿哉, 金淵培, 浦谷則好 “音声対話を用いたテレビ操作インタフェース実験システム,” 電子情報通信学会総合大会, A-14-10, pp.288 (2002).
22. 後藤淳, 小峯一晃, 森田寿哉, 金淵培, 浦谷則好, “テレビ操作のための音声対話インタフェースの試作,” 言語処理学会第 8 回年次大会講演論文集, C3-9 (2002).
23. 森田寿哉, 小峯一晃, 石山邦彦, 後藤淳, 比留間伸行, 浦谷則好, “プロトコル分析を用いたデータ放送コンテンツのユーザーインターフェース評価, ” 電子情報通信学会技術研究報告, HIP2001-12, pp.35-42 (2001).
24. 後藤淳, 清山信正, 三島剛, 今井亨, 都木徹, 安藤彰男, 磯野春雄, “ニュース字幕放送における音声認識の修正支援システム,” 電子情報通信学会ソサイエティ大会, A-15- 8 (2000).
25. 後藤淳, 今井亨, 清山信正, 今井篤, 都木徹, 安藤彰男, 磯野春雄, “ニュース音声認識結果のリアルタイム修正装置, ” 電子情報通信学会総合大会, A-15-15 (2000).
26. 後藤淳, 比留間伸行, 伊藤崇之, 磯野春雄, “プロトコル分析を用いたテレビの見出し画面構成の評価,” 映像情報メディア学会年次大会, 8-2 (1999).

登録特許

1. 特許第 4157418 号　データ閲覧支援装置，データ閲覧方法及びデータ閲覧プログラム
2. 特許第 4934090 号　番組登場人物抽出装置及び番組登場人物抽出プログラム
3. 特許第 4909200 号　人間関係グラフ生成装置及びコンテンツ検索装置，並びに，人間関係グラフ生成プログラム及びコンテンツ検索プログラム
4. 特許第5080368号　映像コンテンツ検索装置及びコンピュータプログラム
5. 特許第5335500号　コンテンツ検索装置及びコンピュータプログラム
6. 特許第5367499号　シーン検索装置及びプログラム
7. 特許第5415369号　番組検索装置および番組検索プログラム
8. 特許第5478146号　番組検索装置および番組検索プログラム

受賞

1. 電子情報通信学会学術奨励賞, “ニュース音声認識結果のリアルタイム修正装置” (2000).
2. 日本放送協会会長賞，“ニュース字幕放送システムの開発と実用化推進” (2001).
3. 電子情報通信学会論文賞, “音声認識を利用した放送用ニュース字幕制作システム” (2002).
4. 日本放送協会放送技術研究所所長賞，“テレビ視聴エージェントの開発” (2006).
5. 映像情報メディア学会「技術振興賞」開発賞，“EN-Vision の開発” (2012).