学位報告

| 報告番号 | ※ 乙 第　　号 |
|---|---|

# 主 論 文 の 要 旨

論文題目　自然言語処理による放送コンテンツ解析手法に関する研究

氏　名　山田一郎

# 論 文 内 容 の 要 旨

本論文は，放送番組で扱われるテキストデータを対象とした情報抽出技術により，番組の説明となるメタデータを自動生成する手法と，メタデータ生成・利用のための知識を獲得する手法に関する研究成果をまとめたものである．

まず，本研究の背景を述べる．近年，放送局では番組を蓄積・管理するシステムが普及し，放送した番組映像などの放送コンテンツを大量に蓄積できる環境が整備されてきた．NHK においても 2007 年 6 月現在，過去に放送された約 61 万番組が NHK アーカイブスに蓄積されている．このような放送コンテンツには有益な情報が多く含まれており，放送コンテンツを効率的かつ効果的に二次利用することが課題となっている．

そこで本研究では，放送コンテンツの内容をシーンごとに詳細に説明したメタデータを効率的に生成することを目的とする．メタデータにより，放送番組中の貴重な映像の存在などの詳細な情報を容易に把握でき，放送コンテンツを効率的かつ効果的に二次利用することが可能となる．メタデータは，大量に放送されている番組から視聴者が好みの番組を選択する際にも有用となる．

次に，本研究の基本課題について述べる．大量の番組に対してメタデータを人手で生成する作業には，大変な労力を要する．実際に，メタデータを利用したサービスとその有効性が明確に示されなければ，放送局におけるメタデータ生成は難しく，現状では，放送番組に対してメタデータは付与されていない．また，スポーツ中継などの生放送番組に対しては，人手を介してもリアルタイムにメタデータを付与することが困難と考えられる．番組に対するメタデータ生成のための研究は，映像解析のアプローチを中心に，これまでに数多く行われてきた．しかし，映像解析による，リアルタイム処理が難しく，精度にも課題が残る．そこで，本論文では，番組に付与されたクローズドキャプションなどのテキストデータを解析して，映像中の内容を特定し，メタデータを自動付与する手法を提案する．この処理では，如何にテキスト情報を解析して，効果的なメタデータを付与することができるかが課題となる．さらに，人間が

持つ一般的な語彙知識をテキスト解析処理に取り入れることにより，解析精度の向上が見込まれる．このような知識の獲得手法も課題となる．

本研究の特徴は，情報系番組，生放送スポーツ番組，ニュース番組といった 3 種類の番組に対して，それぞれの番組に適した統計処理による解析を行い，各番組のクローズドキャプションの特徴をとらえることによりメタデータを自動生成することである．また，各番組に対して付与されたメタデータを利用することにより，テレビ番組のコンテンツを効果的に二次利用するためのアプリケーションの提案も行う．さらに，メタデータ生成の精度向上や効果的な利用に役立つ知識を獲得する手法を提案する．

本論文の構成と内容は以下の通りである．

1 章では，自然言語処理による放送コンテンツ解析手法に関する研究の背景，研究の位置付け，関連研究，特徴について言及し，本論文の大要を説明する．

2 章では，放送コンテンツ中のテキストデータを統計的に解析のために使用する AdaBoost アルゴリズム，GibbsBoost アルゴリズム，Support Vector Machine，最大エントロピー法，そして，EM アルゴリズムについて，その概要を説明する．

3 章では，情報系番組を対象としたメタデータ自動生成手法について述べる．情報系の番組のクローズドキャプションでは，「場所紹介」や「人物紹介」など特定の事柄を表現するために同じような言い回しが多用される．このような言い回しを含む文章区間が抽出できれば，対応する番組映像区間の場所紹介や人物紹介といったメタデータを付与することができる．局所的な部分木しか特徴として利用されない従来法の問題点を改善し，大域的な文章構造の類似性を利用する手法と，さらに，2 種類のサンプリング処理を行うことにより処理時間の問題点を改善した手法を提案する．場所紹介を行うシーンを抽出する実験により，解析結果の調和平均を示す F 値が 0.478 と，従来手法より良好な結果であることを示した．サンプリング処理を行う改善手法でも，精度を維持したまま，処理速度を 45 倍以上の速さに向上できることを確認した．生成したメタデータを利用して放送コンテンツを効果的に二次利用するマルチメディア百科事典についても述べる．

4 章では，生放送スポーツ中継番組のアナウンスコメントを解析することにより，メタデータを自動生成する手法について述べる．生放送スポーツ中継番組のアナウンスコメントには，実際に発生したイベントに対する説明（試合記述文）と，発生したイベントとは直接関係しない補足的な説明（解説文）が存在する．試合記述文は，対応する映像に対するメタデータとして有益な情報となる．サッカー中継番組を対象として，2 種類のコメントを統計的に分類した結果を利用することにより，サッカーの試合で発生するイベントを構成する区間を抽出する手法を提案する．さらに，各区間のイベント名とその主関与者となるイベント動作主を抽出し，サッカー中継番組のメタデータを自動生成する手法につい述べる．コメントの分類実験では，その F 値が試合記述文 0.872，解説文 0.919 と良好な結果が得られ，イベント抽出実験では F 値 0.765 と，従来手法に比べて良好な結果であることを確認した．提案する手法は，他のスポーツ中継番組にも有効と考えられる．また，自動生成されたメタデータを利用することにより，サッカー番組をカスタマイズ視聴できるアプリケーションを紹介する．

5 章では，ニュース番組を対象としたメタデータ自動生成手法について述べる．ニュースは社会の情勢や最新の流行など，豊富な情報が含まれているため，二次利用の有用性が高いと考えられる．放送局では大量のニュース記事データを電子化して蓄積するようになり，これらの効率的な管理，活用が急務となっている．そこで本章では，ニューステキストを解析して，管理するための手法について説明する．同じ話題に属するニュース記事集合に

は共通する出来事があり，この出来事はニュース記事集合を特徴付ける重要な項目と考えられる．係り受け関係を持つ 2 つの文節の自立語と係り元の文節の付属語となる助詞の 3 項組の定型性を評価して各話題に共通する出来事を抽出することにより，ニュースの話題の要約する手法を提案する．要約実験では，ニュースの出現数だけでは読み取れないような話題の重要な項目が抽出できていることを確認した．要約した結果は，ニュースのメタデータとして利用できる．

6 章では，大量のテキストデータから知識を獲得する手法について述べる．最初に大量のテキストデータを解析する際に問題となる未知語処理について言及する．次に，従来研究ではほとんど行われていない語の典型的な機能・目的や起源などの語彙知識を自動獲得する手法について述べる．語彙知識を獲得する実験では，提案手法による結果は人手により生成した正解データとの相関が高いことを示した．これらの処理で得られる未知語の上位語推定結果や単語間の関係は，メタデータを自動生成する処理の精度向上のための知識として利用できる．また，番組で難しい用語が使われる場合，語彙を説明するフレーズが出現する．そこで，用語とその説明を抽出し，用語とその説明間の意味関係を 10 種類に分類する手法について言及する．用語とその説明の意味関係分類実験では，適合率 0.814，再現率 0.760 と良好な結果が得られた．語彙知識の一つとして，物事の「原因－結果」の関係を示す因果関係知識がある．この因果関係知識は，人間の思考において重要な役割を果たし，大量に因果関係知識を蓄積できれば，「何故」といった質問に対する答えを推論により導きだすことが可能となる．健康に関する番組を対象として，専門的な知識となる因果関係知識の獲得手法についても言及する．因果関係節がある名詞ペアを抽出する実験では適合率 0.738 と，良好な結果が得られた．節間の因果関係など 4 つの関係に分類するする実験では，再現率は低いが，適合率は 0.810 と一定の弁別能力があることを示した．用語の説明や因果関係知識は，放送された番組を二次利用する効果的なアプリケーションに有用と考えられる．

7 章では，本論文における研究成果として得られた知見を総括する．本研究成果により，番組の種別に応じたテキスト解析手法による効率的なメタデータ生成が可能となる．放送コンテンツを利用した効果的なアプリケーション実現への可能性についても言及する．