F-024

番組字幕を利用したマルチメディア健康百科事典構築に関する検討

A Study on Multimedia Health Encyclopedia using Closed Caption Data

宮崎 勝† 山田 一郎† 三浦 菊佳† 住吉 英樹† 八木 伸行†
Masaru Miyazaki Ichiro Yamada Kikuka Miura Hideki Sumiyoshi Nobuyuki Yagi

1．はじめに

高齢化社会が進むに従い，メディアに対して医療，健康に関する正しい情報を要望する声が高まっている．NHK では，視聴者の健康増進を目的とした健康情報番組「きょうの健康」を放映しており，番組 Web サイトにおいても，放送した番組に関する要約情報や，キーワードによる治療法検索・専門医検索というサービスを提供している[1]．しかし，現状では入力キーワードに関連した放送済み番組リストを表示するのみであり，ユーザは再放送を待つか，番組概要を読んで情報を得なければならないのが実情である．医療や健康の分野に関してはその性質上，「欲しい時に欲しい情報がすぐ手に入る」ことが重要である．そこで，健康番組の字幕データを利用して，病気・症状などに対する原因や特徴といった関連情報を提示する機能を持った，マルチメディア健康百科事典の試作について報告する．

2．番組字幕データからの節関係獲得

「ある症状が発症した場合，どんな病気の疑いがあるのか」，あるいは「ある病気にはどのような特徴があるのか」といった疑問に答えるためには，病気や症状の間に成り立つ様々な関係を記述し，利用する手法が必要である．医療分野に関しては，オントロジー構築に関する研究が盛んであり，医学辞書などを用いた用語間関係抽出手法などが提案されているが[2]，番組字幕のような話し言葉を対象としてはいない．そこで我々は，健康情報番組に付加されている字幕データを対象とし，同一文中に出現する 2 つの節に対して推定した意味カテゴリーを利用することにより，節間の関係を推定する手法を提案している[3]．この手法により，病名や症状に対応する語句，それぞれの語句が属するクラス名，およびその語句間に成立する属性といった情報を自動抽出することができる．この情報に，字幕のタイムコードや番組映像の URL を付加し，意味構造化して知識ベースに蓄積することにより，ユーザの要求に対して情報を柔軟に提供するサービスの実現が可能となる．

3．健康知識ベースの構築

抽出された節間の情報から，健康知識ベースを構築する手法について述べる．

まず，「きょうの健康」1 番組分の字幕データを人手で分析し，「病気」「症状」といったクラスやその間に成立する「ーを起こす」「ーという特徴を持つ」といったプロパティを OWL[4]言語を用いて記述することにより，小規模な健康オントロジーを構築した．構築したオントロジーをスキーマとし，字幕データから自動抽出した節間の関係をインスタンス化することで，知識ベースを構築することができる．図 1 に，知識ベース内での節関係構造の例を示す．この例は，「動脈硬化になる」という節と「細い血管が詰まる」という節をそれぞれ Disease （病気）クラスのインスタンス(subject123)および Symptom （症状）クラスのインスタンス(object123)で表現し，その間を「ーを起こす」という関係を示す cause プロパティで結ぶことにより，用語間の関係を構造化している．各インスタンスには「動脈硬化になる」といった元の文章中の表現を comment プロパティ値として，「動脈硬化」という標準形を label プロパティ値として付加した．また，節関係自体も CausalRelation クラスのインスタンス(causalRelation123)で表現し，番組映像ファイル URL およびタイムコードをプロパティ値として付加した．このような構造化を行うことにより，様々な関係を用いた検索や，得られた関係から該当する映像を再生することが可能となる．「きょうの健康」約 1,600 番組の字幕データから抽出した節関係を本手法によって構造化，蓄積することにより，健康知識ベースを構築した．

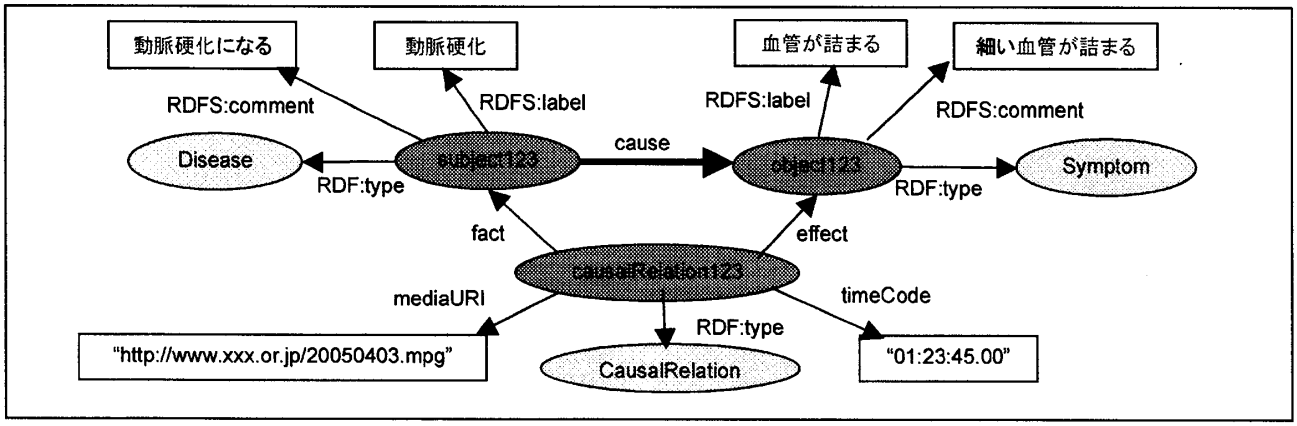

図 1 ：知識ベース内での節関係構造

†NHK 放送技術研究所

4．マルチメディア健康百科事典の試作

Web アプリケーションとして試作したマルチメディア健康百科事典システムのインターフェースを図 2 に示す．ユーザがリストから病名あるいは症状を選択すると，原因や特徴といった関連情報を提示する．ユーザが「糖尿病」という病名を選択した場合，システムは健康オントロジーにアクセスし，病気(Disease)クラスが持つ属性を検索，原因(causedBy)，特徴(hasFeature)，症状(hasSymptom)といった属性リストを得る．得られた属性リストを用いて，「糖尿病」インスタンスが持つ属性値を，知識ベース問い合わせ言語 SPARQL[5]を利用して検索する．得られた属性値を用いて，「糖尿病の原因は肥満」「糖尿病の症状は高血圧」といった情報を提示することができる．さらに情報をクリックすることにより，該当する番組映像を視聴することができる．

5．まとめ

番組字幕から得られた節関係を利用して知識ベースを構築することにより，ユーザの要求に会わせた情報提示，番組視聴機能を実現した．現状，提供する情報の精度はまだ高いとは言えないものの，番組のポータルサイトなどでより高度なサービスを実現できる可能性を示した．

また，システムの試作を通して，下記の課題を明らかにすることができた．

・ 知識の完全性

字幕データは話し言葉であるため，得られた節が知識として有効なものでない場合が多い．例えば「糖尿病」の症状として「体の変化」という情報が得られたとしても，体の「どの部位の」「どのような」変化なのかといった情報が得られていないと，情報として利用することができない．有益な知識を構成するために必須な属性を獲得する手段が必要である．

・ 関係の信頼度

「糖尿病の原因は肥満」という節関係が得られたとしても，実際に肥満の人すべてが糖尿病になるわけではない．逆に，糖尿病の原因は肥満だけではない．各関係に信頼度の重みを導入するなどする必要がある．

・ 番組映像視聴時の区間

「糖尿病には高血圧という症状がある」という情報提示され，それ関連した番組映像を視聴する場合，字幕に付加されたタイムコードデータ(字幕が表示された時刻)のみの利用では，スタートポイントが適当でないなど，正しく番組シーンを視聴することができないことが多い．得られた知識に対するシーンを正しく再生するためには，字幕のタイムコードだけではなく，意味的区間を示すメタデータが必要となる．

今後，これらの課題を解決する手段についても検討を続けていく．また，「高血圧が動脈硬化を起こす」「動脈硬化が脳卒中を起こす」という関係が獲得できていれば，因果関係の推移性を利用して「なぜ高血圧になると脳卒中を起こすのか？」という Why 型の質問にも答えることができる．今後，複数の関係知識を組み合わせたより高度な推論機能の実現を目指していく予定である．

文 献

[1] “NHK 健康ホームページ”, http://www.nhk.or.jp/kenko/

[2] 荒牧英治，今井健，梶野正幸，美代賢吾，大江和彦：“医学辞書を用いた用語間関係の自動抽出手法と用語の自動分類手法に関する検討”，医療情報学，25(6)，pp463-474(2005)

[3] 山田一郎，宮崎勝，三浦菊佳，住吉英樹，八木伸行：“同一文中に出現する複数の節間における因果関係抽出の検討”，第 6 回情報科学技術フォーラム一般講演論文集，5E-3 (2007)

[4] http://www.w3.org/2004/OWL/

[5] http://www.w3.org/TR/rdf-sparql-query/

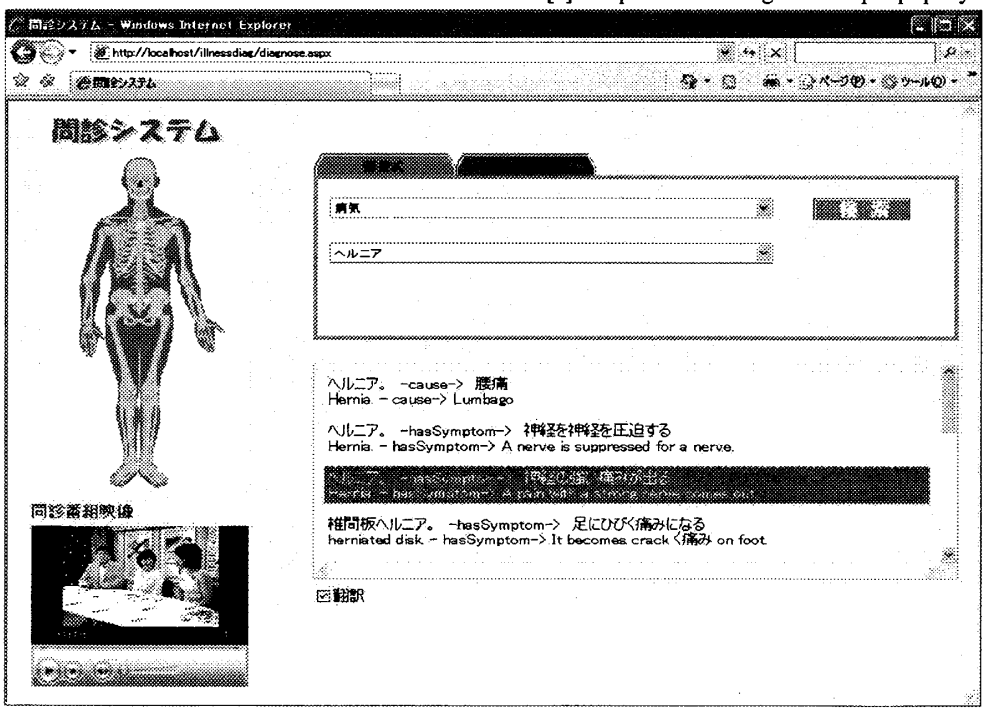

図２：マルチメディア健康百科事典