メディア教育研究　第７巻　第１号
Journal of Multimedia Education Research 2010, Vol.7, No.1, R1－R8

特集（研究開発速報）

メタデータの自動生成を目的としたシラバス文書の情報抽出

辻　靖彦[1)]・森本　容介[1)]

本研究では，検索システムへ登録するための学習オブジェクトメタデータ（LOM）を自動生成することを目的としている。LOMを生成するためには，「タイトル」，「コンテンツの説明」，「URL」，「専門分野」など，その学習オブジェクトに関連する情報が必要となる。そこで本研究では，それらの情報をWeb上の講義シラバスから自動で抽出するために「科目名」や「講義のねらい」などシラバスの中で頻繁に登場すると考えられる項目名に着目した。インターネットで公開されているHTML形式のシラバスから構造を調べ，項目名と実際の内容である項目値との位置関係を考慮する手法を用いて「授業科目名」と「授業目的・内容」の情報抽出を行った。その結果，「授業科目名」属性では93.8%，「授業目的・内容」属性では87.7%の精度が確認された。

キーワード
シラバス，メタデータ，LOM，情報抽出

1．はじめに

さまざまな教育機関によって開発されているWeb上の教育用コンテンツを効率的に検索するために，メディア教育開発センター（現在：放送大学ICT活用・遠隔教育センター）では学習オブジェクトメタデータ（Learning Object Metadata：以下LOM）を用いた学習コンテンツ検索システムNIME-glad（清水・辻・小河原・高野，2005）及びGLOSS（森本・清水，2009）を開発し，運用を行っている。NIME-gladでは2010年３月の時点で，eラーニング教材を始めとする約17万件弱の学習コンテンツのLOMが登録されており，キーワード検索やカテゴリ検索によりそれらのコンテンツを一度に検索することができる。LOMには各学習コンテンツのタイトル，概要，キーワード，サムネイル，ファイルフォーマット，URL，著作権情報などの情報が含まれている（清水，2004）。LOMは，センターの専門スタッフがWeb上の学習コンテンツを閲覧し，手入力することで作成している。この作業には多くの時間がかかっており，自動化が求められている。

一方，Web上に公開されているHTMLから情報抽出を行う研究がさまざまな手法で行われている（梅原・岩沼・鍋島，2002：山田・池田・廣川，2003）。また，Web上のシラバスから情報抽出を試みている研究もあるが（板井・高須・安達，2003：渡辺・絹川・井田・芳鐘・野澤・喜多，2004：野口・山田・池田・廣川，2004），いずれの研究も，特定の学科のシラバス文書でしか検証を行っていないか，もしくはHTMLの<table>要素の内部にある情報しか抽出できない，といった制限がある。NIME-gladのような検索システムに登録する事を想定して情報抽出を行う際には，「専門分野」及び「HTMLの特定のフォーマット」に依存せず，かつ多数のシラバスから情報抽出を高精度で行えることが要求されると考えられる。

以上の背景を踏まえて本研究では，Web上に公開されている学習コンテンツの一つである電子シラバスから，検索システムに登録するためのLOMを自動的に生成する手法を開発することを最終目的とする。本研究で用いる手法を図１に示す。図１より，本研究では「①Web自動収集エージェント」によりインターネット上のシラバスを自動的に収集し，「②情報抽出エージェント」で収集した各シラバスからLOMに必要な項目情報を抽出し，「③LOM自動生成エージェント」で項目情報を基に所定のXML形式のLOMフォーマットに変換を行うことでLOMの自動生成が実現可能であると考える。①についてはWebクローリング技術及び，「授業科目名」などのシラバス中で用いられる用語の頻度に着目する手法（篠原・地蔵，2006）を用いて実現可能と考えられる。また，③については項目情報を特定のXML形式に変換すればよいので実現可能である。②については，「タイトル」，「概要」，「URL」などLOMの各項目に対応する情報を，様々な大学，学科のなるべく多くのシラバス文書から高精度で自動的に抽出できる事が求められる。そこで本論文では，LOMの項目の「タイトル」として利

1) 放送大学

用可能な「授業科目名」属性及び, LOMの項目の「概要」として利用可能な「授業目的・内容」属性の両者について自動的に情報抽出を可能とする手法を開発することを目的とした。初めに単純な抽出アルゴリズムを用いて情報抽出の予備実験を行い，精度を確認した。続いて抽出に失敗したシラバスの書式を加味した上でアルゴリズムの改良を行い，抽出精度を確認した。

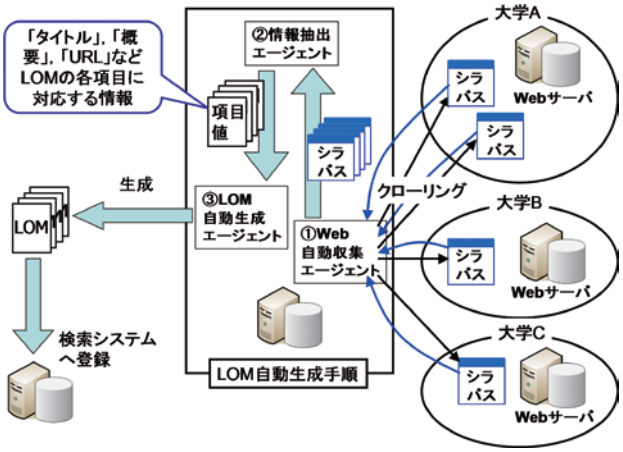

図１　LOM自動生成の流れ

2. 手法

本研究で情報抽出の入力要素として扱うシラバスは，講義ごとに科目名，目的，内容，そして評価方法等が詳細に記されているシラバスを対象としている。科目名しか書かれていないような科目一覧や，履修案内，スケジュールのみ記されている授業計画，簡単な授業内容のWebページは対象外とする。対象となるシラバスから情報抽出を行うために，本研究における基本的な考え方を次に記す。

2.1　情報抽出手法

篠原他（2006）は，対象となるWebページの文章の中に「授業科目名」,「単位」,「開講時期」,「評価方法」など８つの属性を定義し，その属性に含まれる項目名の出現する割合の組み合わせからシラバスかどうかを判別する手法を開発した。

本研究では篠原他（2006）の手法を参考に,「授業科目名」や「本講義のねらい」などの項目名に着目して情報抽出を行う。それらの用語の近くには項目値（実際の授業科目の名前や講義の目的を表す文字情報）が存在する可能性が高く，その位置関係が分かれば抽出可能であると考えられる。その位置関係を把握し抽出手法を確立するために，情報抽出の予備実験を行った。

2.2　予備実験

実験条件を以下に示す。

実験目的

HTML形式のシラバスから「授業科目名」属性と「授業目的・内容」属性の項目名と項目値の位置関係を把握するため。

属性の定義

本論文ではLOMのタイトルに相当する「授業科目名」属性，及びLOMの概要に相当する「授業目的・内容」属性を抽出対象とする。「授業科目名」属性とは篠原他（2006）の手法で用いられた｛授業科目，科目名｝の用語群,「授業目的・内容」属性とは｛授業（の）目標，授業（の）目的，授業のねらい，講義（の）目標，講義（の）目的，講義のねらい，学習目標，到達目標，達成目標，授業内容，授業概要，講義内容，講義概要｝の文字列群を指す。

実験手続き

インターネット上に公開されているac.jpドメインに属する84のWebサイトをクローリングし，篠原他（2006）の手法で機械的にシラバスと判定された12,066件のHTML形式のシラバスに対して，授業科目名と授業目的・内容の抽出を試みた。目視により抽出した情報の精度を確認した。抽出手法を以下に記す。

抽出手法

Step１. チャンク（HTMLタグに挟まれた文字列）から改行以外の空白文字を除去する

Step２. HTMLソースの始めから,「授業科目名」属性又は「授業目的・内容」属性の項目名を含むチャンクを検索する。

Step３. ヒットしたチャンクの次のチャンクを項目値として抽出する。

Step４. HTMLソースの最後のチャンクを検索するまで各チャンクに対してStep１～３を繰り返す。

2.3　予備実験結果

12,066件のシラバスに対して「授業科目名」属性はヒットした14,786事例からランダムに選択した189事例,

表1　「授業科目名」属性の抽出結果

抽出成功（件数）	項目名「科目名」		79
	項目名「授業科目」		19
	小計		98
抽出失敗（件数）	想定外の箇所の検出	「英文授業科目名」	74
		長い文章中	3
		「本授業科目に関する情報」	2
		小計	79
	縦型のテーブル		9
	項目値が改行を挟んで２行に渡る		3
	小計		91
合計（件数）			189
抽出精度（%）			51.9

「授業目的・内容」属性についてはヒットした12,745事例からランダムに選択した346事例について，目視により抽出精度を確認した。

表2　「授業目的・内容」属性の抽出結果

抽出成功（件数）	項目名「授業の目的」		80
	「達成目標」		56
	「授業のねらい」		21
	「授業内容」		15
	「講義目的」		12
	「到達目標」		7
	「授業目標」		5
	「授業の目標」		2
	「授業概要」		2
	「講義目標」		1
	小計		201
抽出失敗（件数）	項目値を一部のみ抽出		80
	項目名を誤抽出	長い文章中に出現	33
		授業計画の表中に出現	6
		冒頭の項目名	4
		title要素	1
	Word		19
	項目名：項目値		1
	項目名の中で改行タグ		1
	小計		145
合計（件数）			346
抽出精度（%）			58.1

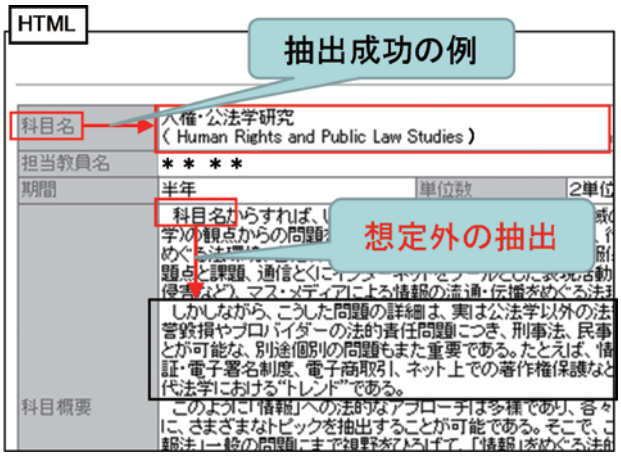

図２　想定外の抽出の例

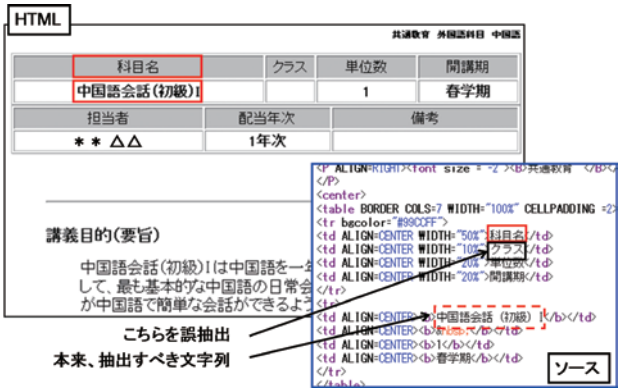

図３　縦型のテーブルの例

2.3.1　「授業科目名」属性の抽出結果

抽出結果を表１に示す。これによると正確な授業科目名が抽出できた精度は51.9%（98/189）に留まった。誤って抽出した91事例について，失敗した理由の内訳を調べると，その大部分が特定のパターンであった。１つ目は想定していない箇所で項目名を検出しているパターンであり，図２にその例を示す。図の上部では本研究で想定している「授業科目名」属性の項目名を検出しているのに対し，図の下部では長い文章の冒頭で「科目名」の文字列を検出している。この文章は文字数が多くかつ助動詞または動詞を含んでおり，そこには実際の授業科目名は存在しないので，ここで検出された「科目名」の文字列は求める「授業科目名」属性の項目名ではないと考えられる。２つ目のパターンは，実際の授業科目名を含むテーブルにおいて項目名と項目値が縦に配列しているために，項目名の次のチャンクが項目値とは異なる場合である。図３の例では，「科目名」の文字列の下にある「中国語会話（初級）I」の文字列を抽出しなければならないが，本アルゴリズムでは次のチャンクである「クラス」の文字列が誤抽出されてしまう。この２つのパターンだけで91事例中88事例に達した。残りの３事例は授業科目名の一部のみ抽出されたパターンであった。精度を高めるにはこの２通りの失敗例を克服できることが必要であると考えられる。

2.3.2　「授業目的・内容」属性の抽出結果

抽出結果を表２に示す。この表より，正確に授業目的・内容を抽出できた精度は58.1％（201/346）であった。抽出できなかった事例では，
，<p>要素などで授業目的や授業内容が改行，改段落されているなどの理由で実際の内容の一部分のみ抽出されたパターンが多く，23.1％（80事例）を占めている。その他は「授業科目名」属性と同様に，想定していない箇所での用語の検出が12.7％（44事例）確認できた。その他の失敗した事例には，Wordで作成したHTMLのためにHTMLタグが複雑になっているパターン，検出した項目名の次に「：」の文字が来てその後に項目値が続くパターンなどがいずれも少数であるが確認できた。また，「授業科目名」属性の場合と異なり，項目名及び項目値を含むテーブルが縦型であるために抽出できないというパターンは確認できなかった。

2.3.3　抽出結果のまとめ

上記の抽出結果より，「授業科目名」属性と「授業目的・内容」属性の項目値の抽出精度を上げるには，失敗パターンを考慮し，クリアできる手法を検討する必要がある。失敗パターンは各属性で異なる傾向が確認できたので，抽出アルゴリズムもそれぞれ別の手法を用いる必要があると考えられる。例として，「授業科目名」属性においては

- 項目名の誤抽出，特に長い文章や助動詞または動詞を含む文章中における誤抽出を減らす事
- 項目名と項目値が縦に並んでいるテーブルに対応できる事

を考慮する必要があると考えられる。「授業目的・内容」属性においては

- 項目名の誤抽出，特に長い文章中における誤抽出を減らす事
- 項目値の一部のみ抽出しているパターンをすべて抽出できるようにする事

を考慮する必要があると考えられる。次節でこの旨を考慮し，抽出アルゴリズムを改良する。

2.4　抽出手法の改良

前節の結果を踏まえて，「授業科目名」属性及び「授業目的・内容」属性の項目値抽出アルゴリズムを改良した。「授業科目名」属性の項目値抽出の流れを以下に示す。入力はHTML形式のシラバス文書である。

「授業科目名」の抽出手法（改）

Step１． チャンクから改行以外の空白文字を除去。

Step２． 項目名をヒットしたチャンクに対して項目名の前後の文字列を調査。

Step２－１． 前後に文字が存在しない，又は「・」，「■」，「□」，「◆」，「◇」，「：」などの特定の記号・改行文字しか存在しない場合はStep３.へ。そうでなければStep２－２.へ。

Step２－２．「授業科目名」属性の項目名の前の文字列を見て，「この」や「英文」など特定の文字を含む場合は項目値を抽出せずに終了。

Step２－３．「項目名：項目値」の形であれば項目値を抽出して完了。

Step２－４． 項目名を除く前後の文字列の長さの合計が10以上であれば項目名とみなさずに終了。また，項目名を含む全文字列を形態素解析し，形態素の中に助動詞又は動詞を含む場合も項目名とみなさずに終了。そうでない場合はStep３.へ。

Step３．「授業科目名」属性の項目名及び，その次のチャンクを項目値として暫定的に抽出。

Step３－１． 項目名及び次のチャンクが<table>要素内にあるか調査。

Step３－１－１． <table>要素の中ではない場合は次のチャンクを項目値とみなして抽出完了。

Step３－１－２． 同じ<table>要素内の同一の<td>～</td>要素（以下「セル」と呼ぶ）内にある場合も次のチャンクの文字列を項目値とみなして抽出完了。

Step３－１－３． 同じ<table>要素内の別のセルにある場合，その<table>要素の行数及び列数を調査。

Step３－１－３－１．（行数＝１）or（行数!＝２and 列数＝２）の場合，項目名及び項目値が横に並んでいるテーブルとみなし，次のチャンクを項目値として抽出し完了。

Step３－１－３－２．（列数＝１）or（列数!＝２and 行数＝２）の場合，項目名及び項目値が縦に並んでいるテーブルとみなし，項目名の真下のセル内のチャンクを項目値として抽出し完了。

Step３－１－３－３． Step３－１－３－１.とStep３－１－３－２.のどちらにも当てはまらない場合，後述の縦・横テーブル判定手法により，テーブルが縦型か横型かを判断。縦型であれば項目名の真下のセル内のチャンク，横型であれば次のチャンクを項目値とみなして抽出完了。

ここで，Step３－１－３－３.で出てきた縦・横テーブル判定手法を以下に示す。

縦・横テーブル判定手法

Step１．「授業科目名」属性の項目名を含むセルを中心として，同じ行と列のセル内のチャンクを抽出。

Step２． 各チャンクが以下に定義する属性用語群を含むかどうか調べ，同列のセル内に含むチャンクが多い場合は横型，同行のセル内に含むチャンクが多い場合は縦型と判定。

尚，この属性用語群は篠原他（2006）のシラバスの判定に用いている用語群をそのまま引用した。以下に属性用語群を示す。

属性用語群＝｛単位，学期，開講期，開講時期，前期，後期，通年，担当教員，担当教官，担当者，教科書，参考書，テキスト，参考文献，使用教材，授業（の）目標，授業（の）目的，授業のねらい，講義（の）目標，講義

（の）目的，講義のねらい，学習目標，到達目標，達成目標，授業内容，授業計画，授業概要，講義内容，講義概要，講義計画，評価（の）方法，成績（の）評価|

改良手法による抽出例と抽出の流れを図４に示す。始めに，HTMLの各要素以外の文字列を１つずつ探索する。そして図４の１より，「授業科目名」属性の項目名を含む文字列に対して，余分な空白を削除する。次に，項目名の前後に文字列が存在しないことを確認した上で，項目名とその次のチャンク文字列を暫定的に項目値として抽出する。図４の２よりこの例では，暫定の項目名として「科目名」,項目値として「単位数」を抽出する。続いて，この項目名及び項目値が<table>要素中にあるかどうかを判定する（図４中の３）。これらは<table>要素中にあるので，行数と列数をカウントしてテーブルが縦型か横型かを判定する。その結果，行数＝２かつ列数＝７であるので，これは縦型のテーブルと自動判定され，「科目名」の真下のセル中の文字列「○○○学（総論）」を項目値として抽出する（図４の４）。

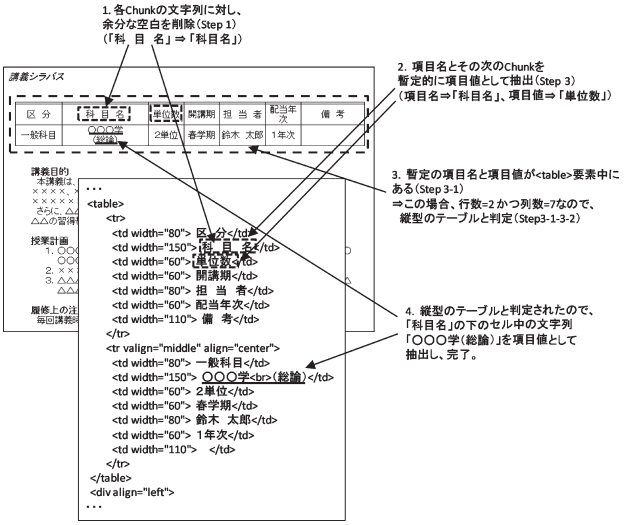

図４　改良手法による「授業科目名」の抽出例

次に「授業目的・内容」属性の項目値抽出における改良手法を以下に示す。

「授業目的・内容」の抽出手法（改）

Step１． 各Chunkから改行以外の空白文字を除去。

Step２． 項目名をヒットしたチャンクに対して，項目名の前後の文字列を調べる。

Step２－１． 項目名が<title>要素内に含まれている場合は想定外の抽出とみなして終了。

Step２－２． 前後に文字が存在しない，又は「・」,「■」,「□」,「◆」,「◇」,「：」,「[」,「]」,「【」,「】」などの特定の記号・改行文字しか存在しない場合はStep３.へ。そうでなければStep２－３.へ。

Step２－３． 「項目名：項目値」の形であれば項目値を抽出して完了。

Step２－４． 項目名を除く前後の文字列の長さの合計が10以上であれば抽出せずに終了。また，項目名を含む全文字列を形態素解析し，形態素の中に助動詞又は動詞を含む場合も抽出せずに終了。そうでない場合はStep３.へ。

Step３． 項目名を含むチャンクを$Chunk_0$，n=1としてその前後のHTMLタグ群をそれぞれTag_{-1}とTag_n，Tag_nの隣から順番に$Chunk_n$，Tag_{n+1}，$Chunk_{n+1}$とする。

Step３－１． 【Tag_{-1}とTag_2】又は【Tag_1とTag_2】の間でブロック要素の<*>～</*>の関係が成立するならば$Chunk_1$を項目値として抽出し完了。

Step３－２． Tag_2が<td>を含まず，かつ</td>を含むならば$Chunk_1$を項目値として抽出し完了。

Step３－３． $Chunk_2$がStep２－２.の記号群を含む場合，$Chunk_1$を項目値として抽出し完了。

Step３－４． $Chunk_2$が属性用語群中のいずれかの属性用語を含み，かつlength（$Chunk_2$）≦length（$Chunk_0$）+10，かつ$Chunk_2$が助動詞及び動詞を含まなければ$Chunk_1$を項目値として抽出し完了。

Step３－５． n=2として，Step３－１.と同様に【Tag_{-1}とTag_n】又は【Tag_1とTag_{n+1}】の間でブロック要素の<*>～</*>の関係が成立するならば$Chunk_1$～$Chunk_n$を項目値として抽出し完了。

Step３－６. Step３－２と同様に，Tag_{n+1}が<td>を含まず，かつ</td>を含むならば$Chunk_1$～$Chunk_n$ を項目値として抽出し完了。

Step３－７． Step３－３.と同様に，$Chunk_{n+1}$がStep２－２.の記号群を含む場合，$Chunk_1$～$Chunk_n$を項目値として抽出し完了。

Step３－８． Step３－４.と同様に，$Chunk_{n+1}$が属性用語群中のいずれかの属性用語を含み，かつlength（$Chunk_{n+1}$）≦length（$Chunk_0$）+10，かつ$Chunk_{n+1}$が助動詞及び動詞を含まなければ$Chunk_1$～$Chunk_n$を項目値として抽出し完了。

Step３－９． n=n+1として，Step３－６.～Step３－８.を繰り返す。但し，n=5まで行い，終了しない又は次のChunkが見つからなくなった場合は$Chunk_1$を抽出して完了とする。

改良手法による抽出例と抽出の流れを図５に示す。始めに，各Chunkの文字列に対して空白を削除し，「授業目的・概要」属性の項目名があればその前後の文字列を

調べる（図５の１）。次に，項目名を含む$Chunk_0$の前後のHTML要素群をそれぞれTag_{-1}，Tag_1とし，続いてHTML要素以外の文字列部分を$Chunk_1$，$Chunk_2$，$Chunk_3$…と定義する（図５の２）。次に，【Tag_{-1}とTag_2】及び【Tag_1とTag_2】の間にブロック要素が存在するかを確認するが，図５の例ではTag_1が「
」，Tag_2が「
」であることからブロック要素は存在しないことが分かる。さらに，Tag_2には</td>要素も含まず，$Chunk_2$には「・」や「■」などのStep２－２の記号群は含まない（図５の４）。次に$Chunk_2$がStep３－４の条件を満たすか調べると，$Chunk_2$は属性用語を含まず，さらに$Chunk_0$＋10文字以下の文字列長であるので，この条件も満たさないことが分かる（図５の５）。続いて，図５の３～５と同様に【Tag_{-1}とTag_2】又は【Tag_1とTag_3】の間にブロック要素が存在するかを調べるが，Tag_3が「

」であることからこちらも存在しない。同様にTag_3には</td>要素も含まず，$Chunk_3$にはStep２－２.の記号群は含まれない。続いて図５の６より，Chunk３に属性用語が含まれるかを調べると，「授業計画」の用語があり，かつ$Chunk_0$+10文字以下であることからStep３－８.の条件を満たしている。以上の流れから，$Chunk_1$と$Chunk_2$の文字列が項目値として抽出される。

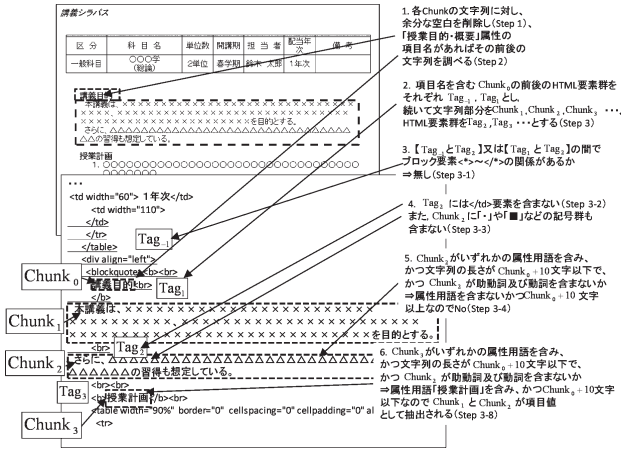

図５　改良手法による「授業目的・概要」の抽出例

2.5　改良アルゴリズムによる抽出実験

前節の改良したアルゴリズムを用いて，2.2節と同様のac.jpの84ドメインの12,066件のシラバスを対象に再実験を行った。その結果，「授業科目名」属性については10,513件，「授業目的・内容」属性は7,159件の抽出結果が得られた。そのすべての結果事例について，目視により精度を確認した。

2.5.1　「授業科目名」属性の抽出結果

抽出結果を表３に示す。10,513件の抽出結果に対して，9,863件の授業科目名が正確に抽出できた。抽出精度は93.8%（9,863/10,513）であった。正確に抽出できた内訳は，項目名が「科目名」であるのが9,212件，項目名が「授業科目」であるのが651件であった。抽出に失敗した650件に対して原因を目視で確認した所，シラバスがMicrosoftのWordにより作成されたHTML形式であるのが半数以上の343件を占めていた。WordでHTMLファイルを作成するとHTMLタグが複雑な構造となるために抽出が困難であった原因が考えられる。他の失敗例としては，本抽出で想定していない「科目名副題」や「本授業科目に関する情報」など，授業科目名の項目名以外の文章にヒットしてしまう事例が268件確認できた。尚，HTMLのテーブル構造の縦・横を誤判定する事例は20件だけであった。2.2節の実験結果における縦型テーブルの割合と比較すると，こちらは大幅に改善できたと考えられる。

2.5.2　「授業目的・内容」属性の抽出結果

抽出結果を表４に示す。7,159件に抽出結果に対し，6,281件の授業目的・内容が抽出できた。抽出精度は87.7%（6,281/7,159）であった。抽出成功した項目名は「達成目標」が3,261件で最も多く，それに続いて「講義概要」1,013件，「講義目的」617件，「授業の目的」404件，「講義内容」244件となった。抽出が失敗した例としては，授業計画の表の中に「授業内容」などの項目名にヒットして誤抽出してしまう事例が最も多く，317件あった。他の例としては，<p>要素，<font>要素，<dd>要素などで実際の授業目的が区切られているために一部のみしか抽出できない事例が多く見られた（265件）。他には「授業科目名」属性と同様に，Wordで作成されたシラバスであるために抽出に失敗している事例が199件確認された。

表3　改良手法による「授業科目名」属性の抽出結果

抽出成功（件数）	項目名「科目名」	9,212
	項目名「授業科目」	651
	合計	9,863
抽出失敗（件数）	Wordで作成したため，タグが複雑	343
	想定外の位置でのヒット	268
	縦型テーブルの解釈に失敗	20
	「項目名：項目値」以外の形式で項目名と項目値が同一文字列中に存在	11
	<title>要素でヒット	5
	項目名を含む文字列が途中で 要素等で改行	3
	合計	650
合計（件数）		10,513
抽出精度（%）		93.8

表4　改良手法による
「授業目的・内容」属性の抽出結果

抽出成功（件数）	項目名「達成目標」	3,261
	「講義概要」	1,013
	「講義目的」	617
	「授業の目的」	404
	「授業のねらい」	266
	「講義内容」	244
	「到達目標」	134
	「授業内容」	106
	「授業概要」	78
	「授業の目標」	67
	「講義の目的」	62
	「学習目標」	19
	「講義のねらい」	5
	「講義目標」	4
	「授業目的」	1
	合計	6,281
抽出失敗（件数）	授業計画の表中の項目名にヒット	317
	<p>要素等で区切られていて一部のみ抽出	265
	Word	199
	項目名を含む文字列が途中で 要素等で改行	46
	想定外のヒット	32
	項目名に対応する項目値が存在しない	10
	項目値が<table>要素で記述	4
	ページ先頭の見出しにヒット	4
	合計	878
合計（件数）		7,159
抽出精度（%）		87.7

3. まとめ

本稿では，LOMを自動生成することを目的に，Web上の講義シラバスを対象に「授業科目名」属性，「授業目的・内容」属性の項目値の抽出を試みた。その結果，「授業科目名」属性では93.8%，「授業目的・内容」属性では87.7%の精度が確認された。

今後の課題としては，以下が考えられる。

- 再現率の評価
- 「授業科目名」，「授業目的・内容」以外の種類の項目値の情報抽出
- さらに大規模な講義シラバスへの情報抽出

謝辞

本研究の一部は科学研究費補助金（課題番号：17650267）の助成による。

引用文献

清水康敬（2004）．高等教育におけるe-Learningの支援と教育コンテンツの共有．メディア教育研究，Vol. 1, No. 1, pp. 1-9

清水康敬，辻靖彦，小河原正久，高野雄二，“LOM検索システムによる学習ゲートウェイNIME-gladの開発と運用”，JSiSE2005 30周年記念全国大会講演論文集，pp. 377-378, Aug. 2005

森本容介，清水康敬，“学習コンテンツのメタデータ流通基盤と検索アプリケーションの提案”，電子情報通信学会技術研究報告，Vol. 108, No. 406, pp. 13-16, 2009-3

篠原正典，地蔵真作（2006）．Web上の高等教育用コンテンツの自動収集と抽出. 教育システム情報学会誌，Vol. 23, No. 3, pp. 115-124

梅原雅之，岩沼宏治，鍋島英知（2002）．事例に基づくシリーズ型HTML文書の意味論理構造の自動認識，人工知能学会論文誌，Vol. 17, No. 6E, pp. 690-698

山田泰寛，池田大輔，廣川佐千男（2003）．半構造化文書に対する木構造と文字列を組合せたラッパーの自動生成法，情報処理学会研究報告．情報学基礎研究会報，IPSJ SIG Notes, Vol. 2003, No.98 pp. 115-122

板井久美，高須淳宏，安達淳（2003）．HTMLからの情報抽出と統合．NII Journal, No. 6, pp. 9-19

渡辺将尚，絹川博之，井田正明，芳鐘冬樹，野澤孝之，喜多一（2004）．シラバスHTML文書からの情報抽出．情報処理学会 第66回全国大会，講演論文集Vol. 4, pp. 487-488

渡辺将尚，絹川博之，井田正明，芳鐘冬樹，野澤孝之，喜多一（2004）．Web上のシラバス情報の収集とXML変換．第３回情報科学技術フォーラム（FIT2004），pp. 121-122

野口龍太郎，山田泰寛，池田大輔，廣川佐千男（2004）．頻度情報を用いたWeb文書群からのテンプレート抽出，DEWS 2004, 6-B-01

辻（つじ）　靖彦（やすひこ）
2004年東京工業大学大学院社会理工学研究科博士課程修了。博士（工学）。信州大学高等教育システムセンター特別研究員，メディア教育開発センター准教授等を経て，2009年４月より放送大学ICT活用・遠隔教育センター准教授。専門は教育工学。

森本（もりもと）　容介（ようすけ）
2005年東京工業大学大学院社会理工学研究科博士課程修了。博士（工学）。メディア教育開発センター助手等を経て，2010年４月より放送大学ICT活用・遠隔教育センター准教授。専門は教育工学。

メディア教育研究　第７巻　第１号
Journal of Multimedia Education Research 2010, Vol.7, No.1, R1－R8

Information Extraction from Course Syllabi for Automatic Metadata Generation

Yasuhiko Tsuji[1)] , Yosuke Morimoto[1)]

Our purpose of this research is to make Learning Object Metadata（LOM）automatically for information retrieval system. In order to make LOM, various items of information about educational resources, for examples “title”, “content explanation”, “URL”, and “classification”, are required. In this paper, we tried to extract automatically item values of “subject name” and “aim of lecture” from course syllabi by focusing on these item names and by checking up positional relation between them in HTML course syllabus. As a result, a precision rate 93.8% of “subject name” and a precision rate 87.7% of “aim of lecture” were obtained.

Keywords
Syllabus, Metadata, LOM, Information Extraction

[1)] The Open University of Japan