다국어 학습을 위한 XML 기반 학습시스템의 설계

정휘웅, 윤애선 (hwjeong1@langue.fr.pusan.ac.kr, asyoon@hyowon.cc.pusan.ac.kr)
부산대학교 인지과학협동과정, 부산대학교 불어불문학과

Developing XML based multilingual language education system

Hwi-woong Jeong, Aesun Yoon
Department of Cognitive Science, Department of French, Pusan National University

요약

XML 은 언어정보의 재사용성 및 다른 유형의 정보로 변환이 용이하여 최근 그 사용이 급증하고 있다. 그러나 XML 은 아직까지 일부 분야에 국한되어 이용되고 있으며, 국내에서도 XML 을 실제 활용하여 개발되고 있는 시스템은 극히 미약하다. 본 연구에서는 XML 의 이점을 살려 한글을 포함한 다국어간 언어학습 컨텐트를 쉽게 구성하고 가공할 수 있는 XML 문서 내의 다국어 표현 방법에 대해 연구하였다. 또한 다국어 정보를 웹 환경에서 구현하기 위한 XSL 과 유사한 문서 변환 구조 및 이를 처리할 수 있는 XML 처리기의 구조에 대해서도 소개한다. 본 연구에서 소개하는 문서 변환 구조를 이용할 경우 문자로 표현 가능한 매체를 매개로 하여 다양한 멀티미디어 컨텐트를 쉽게 작성할 수 있다.

1. 서론

XML 은 HTML 과는 달리 사용자의 의도에 따라 정보 구조를 구성할 수 있다. 따라서 XML 은 정보 구성의 기준을 개발자-혹은 제공자- 중심의 정보 구조에서 수요자 중심의 정보 구조로 이동시키고 있다. 그러나 수요자 중심의 정보 구조는 많은 사용자들의 주관적 해석에 따른 구성 및 비효율적인 정보 구성으로 그 효용성이 HTML 에 떨어질 수 있다.

특히 주관적 결정에 따른 XML 태그는 상호 문서간의 의미 불일치를 유발시켜 정보 재사용성 및 가공성 극대화라는 XML 의 근원적 목적을 감소시킨다. 이를 위해 Namespace 제정 및 기업간 문서 교환 포맷의 개념적 규약을 정의하고 있으나, 이 역시 전자상거래 및 일부 분야에 국한되어 있다.

더욱이 이러한 XML 정보는 다양한 형태로 가공될 수 있으나, 아직까지 이러한 문서 가공에 대한 정보 및 연구가 많이 이루어져 있지 않다. 이러한 현실은 언어학습 컨텐트에서 더욱 많이 발견된다. 멀티미디어 정보에 대한 지원 및 비정형적 정보에 의한 컨텐트 제작의 어려움은 표준화 작업을 더욱 어렵게 만들고 있다. 그러나 이 경우 각각의 대상언어에 대한 언어 학습 컨텐트를 개발하는 경우 매번 새로운 문서를 작성해야 한다. 이는 재사용성 및 경제적 측면에서도 매우 비효율적인 개발방식이다. 및 본 연구에서는 이러한 자료 재사용성의 극대화 및 효과적 컨텐트 구성을 위해 XML 에 기반하여 다국어 지원이 가능한 언어학습 컨텐트의 기본 구조를 소개하고, 이를 웹 환경에서 구현하는 방법에 대해 소개하겠다.

2 장에서는 다국어를 XML 을 이용하여 표현하는 방법 및 대안을 제시하겠으며, 3 장에서는 이를 웹 환경에서 효과적으로 나타내는 방법, 4 장에서는 이러한 문서 구조가 시스템 환경에서 구동되기 위해 구현된 시스템 구조를 소개하겠다. 5 장에서는 결론 및 향후 연구에 대해 소

개하겠다.

2. XML 기반 다국어 언어학습정보

언어학습을 위해서는 다양한 문자 표현이 가능해야 한다. 이를 위해 본 연구에서는 유니코드에 기반하여 다국어 정보구조를 구현하였다. 그러나 유니코드의 코드 영역만으로는 해당 문자가 어느 언어에 속하는 단어인지를 알아내기는 다음과 같은 이유에서 매우 어렵다.

첫째, 한글 및 러시아어, 그리스어의 경우 코드영역이 명확하게 분할되어 있으나, 서유럽어 및 동유럽권 언어는 동일한 코드페이지에 존재하는 문자열이 많다. 둘째, 각 단어의 언어를 자동으로 검색하기 위해서는 지정된 단어의 의미를 포함하는 사전을 구성해야 하나, 아직까지 이에 대한 언어 정보가 부족하다. 더욱이 동일한 알파벳으로 구성된 단어의 경우 정확한 언어 분석을 위해 어휘의 전후관계를 검색하여 지정된 언어를 검색해야만 한다.

따라서 본 연구에서는 모든 Document Object에 언어를 나타내는 속성값을 지정할 수 있도록 구성하였다. 또한 문서의 언어를 나타내는 기본적 속성값을 나타내는 Document Object를 구성하여, 언어 표현 속성값이 없을 경우 기본 언어가 대체될 수 있도록 구성하였다.

그러나 XML이 Document Object Model의 한 유형이듯, 내부를 구성하는 각 Object는 다른 요소와 구분되는 ID를 포함하고 있어야 한다. 그러나 본 연구에서는 다국어 정보를 하나의 Object에 표현해야 하므로, 각 언어에 해당하는 언어 정보를 구성하였다. 이 경우 2 가지 방법을 고려할 수 있다. 첫째는 언어를 인지할 수 있는 정보를 태그로 나타내는 방법이며, 둘째는 언어를 인지할 수 있는 정보를 속성으로 나타내는 것이다.

언어정보를 태그로 나타내는 경우 각 언어태그별로 구분이 가능하고, 해당 언어 태그에 대한 특수한 속성값을 새롭게 정의할 수 있으므로, 문서 구성에 있어 보다 다양한 옵션을 제공할 수 있다. 그러나 이 경우 상위 태그의 특성이 모든 내부 요소에 동일하게 적용됨을 알리는 것일 뿐만 아니라, 상위 속성을 설명하기 위해 각 언어정보는 하나씩만 존재해야 한다는 단점이 있다.

언어정보를 속성으로 나타내는 경우 태그 구조가 단순화되며, 기술된 정보의 특성을 설명하기가 용이하다. 더욱이 태그 내부의 속성은 조건 분기문 구성이 용이하다. 이러한 문서 구조를 예를 들어 기술하면 다음과 같다.

```
<MONO ID='A'>
        <ENG>I speak French</ENG>
        <FRN>Je parle français</FRN>
        <KOR>나는 불어를 말합니다.</KOR>
</MONO>

<MONOGRP ID='A'>
        <MONO LANG='ENG'>I speak French</MONO>
        <MONO LANG='FRN'>Je parle
                français</MONO>
        <MONO LANG='KOR'>나는 불어를
                말합니다.</MONO>
</MONOGRP>
```

본 연구에서는 문서의 의미 뿐만 아니라 변환 과정에 구현되는 문법 구조를 단순화 하기 위해 후자의 형태를 취하였다. 이에 기반한 문서 구조는 표 1과 같다.

표 1 언어학습 컨텐트 구조 정의

태그명	의미 및 속성
MONOGRP	(ID, TYPE='EXAM|SCR|QSTN') 일반 문장. 내부에 저장된 모든 MONO 태그는 동일한 의미를 가지는 것으로 가정한다. 예문, 일반 서술형 문장, 질문으로 그 유형을 정의한다.
MONO	(LANG) 언어 를 지정하여 각 언어에 해당하는 문장을 지정한다.
TALKGRP	(ID) 대화 문장. MONOGRP 태그와 같이 내부에 저장된 모든 TALK 태그는 동일한 의미를 가지는 것으로 가정한다.
TALK	(WHO, LANG) 화자(話者)를 지정하며, 각 언어에 해당하는 대화내용을 지정한다.
DESC	(LANG) 내부 설명문을 의미한다. LANG 태그는 내부 설명문이 어떠한 언어로 구성되었는가를 의미한다.
DOC	(DEFLANG)문서 전체를 구성하는 영역으로서 DEFLANG 에서 기준 언어를 구성한다. 이후 태그에서 LANG 태그가 누락되는 경우 DEFLANG의 언어를 기준으로 컨텐트 내용을 해석한다.
TABLE	(COLSPAN, ROWSPAN, CLASS)테이블을 구성한다. 하위 태그는 TD, TR, TH 등 HTML 포맷과 동일하다. CLASS 는 레이아웃 구조를 간편하게 하기 위해 CSS의 클래스를 지정한다.

태그명	의미 및 속성
QGRP	(ID, TYPE='OBJ|SBJ', ANS) 질문 구조를 구성한다. 질문 형태는 TYPE에서 정의하며 객관식인지 주관식인지를 정의한다. 정답과 문제 ID를 정의한다.

3. Web 환경내 구현

XML 기반 정보는 Microsoft Internet Explorer 5.0 이상에서만 계층적 구조로 단순하게 표현된다. 그러나 XML 만으로는 그 자체가 단순한 정보를 제시하는 형태를 취하므로, 다양한 매체 및 언어 학습 효과를 향상시키기 위한 다양한 미디어 및 시각적 효과, 선별적 정보 제공 등 다양한 사용자 요구를 만족시키지 못한다.

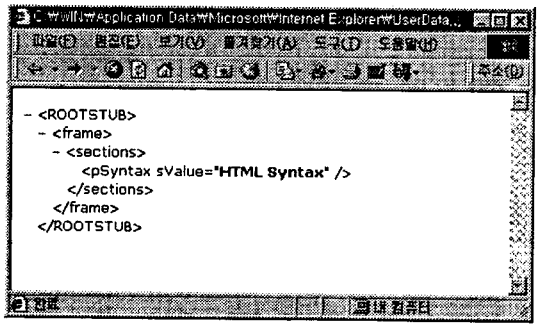

그림 1 웹 환경에서의 XML 정보

따라서 웹 환경에서 다양한 언어정보 및 멀티미디어 정보를 제공하기 위해서는 XSL(Extensible Style Language)과 같은 정보 변환 언어와 CSS(Cascading Style Sheet)와 같은 레이아웃 전문 언어, 기계와 인간간의 대화성을 높이기 위한 Java와 같은 언어 및 기술의 지원이 필수적이다.

그러나 본 연구에서는 궁극적 목표가 범용적 언어정보 구현을 하는데 그 목적이 있으므로 CSS와 Java의 기능은 도입하되, XSL의 부분적 기능을 구현한 변환 XML 문서 구조를 설계하였다. 이 문서 구조가 XML 다국어 언어학습 컨텐트와 변환되는 시스템은 다음 장에서 소개한다.

XSL이 템플릿 및 분기문 구조에 의해 구성되듯이, 본 연구에서 제시하는 변환 구조 역시 템플릿 환경에 기반하여 구성된다. XSL은 XML 태그의 Attribute에 의해 구성되나, 본 연구에서 제시되는 문서 구조는 하나의 셋 내부의 특수문자 구분을 기준으로 시스템이 구성된다. 변환 과정을 위하여 기본적인 태그셋이 주어지며, 변환 태그셋은 1개의 태그에 국한된 변환 루틴과, 패턴에 기반한 변환 루틴으로 구성된다. 모든 문서의 기준 문자열은 Unicode에 기반하여 구성된다.

표 2 변환 정보 구성 XML 정보

태그명	의미 및 속성
TAG	하나의 태그만이 존재하며 내부적으로 하나의 CONV 태그를 가진다.
CONV	변환되는 정보를 구성한다. 대괄호-[ ]-는 변환과정이 완료된 후 HTML의 태그 구분문자-< >-로 변환되며, 중괄호 문자-{ }-는 분기 및 XML의 요소 지정에 사용된다.
PATTERN	두 개 이상의 태그가 구성되는 경우 동시에 정보를 처리하기 위해 구성된다. 태그간의 구분은 콤마(,)로 한다.

다음은 표 2에서 제시한 변환 정보에 기반하여 구성한 변환 테이블의 예이다.

```
<TAG>MEAN
<CONV>
      [p class=MEANING]
      {if exist MEAN:ES}[font
      class=ES]영동:{MEAN:ES}[/font]{/if}
      {if exist MEAN:SY}[font
      class=SY]동:{MEAN:SY}[/font]{/if}
      {if exist MEAN:AN}[font
      class=AN]반:{MEAN:AN}[/font]{/if}
      {if exist MEAN:AB}[font
      class=AB]{MEAN:AB}[/font]{/if}
      {MEAN}
      [/p]
</CONV>
</TAG>
<TAG>SEP
<CONV>
      [p]([I]{SEP:CLAS}[/i])
      [font class=PR]、
      &#91;{SEP:PRON}&#93;[/font]
      [ol]{SEP}[/ol]
      [/p]
</CONV>
</TAG>
<PATTERN>EXAM,MEAN
<CONV>
      [p][b]{EXAM}[/b]: {MEAN}[/p]
</CONV>
</PATTERN>
```

CONV 태그의 내부 정보는 별도의 변환 루틴을 가지고 구동되며, 대괄호는 일반 HTML의 TAG 구분 문자로 구성되나, 중괄호로 구분되는 영역은 개별적인 의미와 함께 별도의 처리 루틴을 가진다.

표 3 변환 관련 명령어 구성표

명령	의미 및 속성
{태그}	해당 태그내부에 있는 문자정보를 출력한다.
{태그:속성}	태그 내부의 속성값을 실제 출력값으로 이용한다. 가령 대화체에서 WHO 라는 속성을 이러한 형태로 구성하면, 화자의 내용이 함께 나타난다.
{CHILD}	현재 태그 이후로 존재하는 자식노드들을 함께 처리한다.
{if exist 태그:속성} {/if}	만약 특정한 속성값이 존재하는 경우 문서를 처리한다.
{if value 태그:속성 = '값'} {/if}	만약 특정 태그의 값이 지정된 값을 가지는 경우 내부의 값을 HTML 값으로 출력한다.

이러한 변환문서 구조는 HTML 문서의 특성에 따라 별도로 구성되나, 문서의 레이아웃 구성은 CSS에 의해 구성된다. CSS는 W3C에 의해 이미 정해진 형태로 구성된다. 그러나 변환 테이블 내부에서 지정되는 자주 사용되는 태그인 P, BODY, FONT, DIV 와 같은 태그에 대한 기본적인 레이아웃 구성과 CLASS에 의한 레이아웃 구성으로 구분된다.

```
BODY
{
    BACKGROUND-COLOR: floralwhite;
    FONT-FAMILY: verdana, 굴림;
    FONT-SIZE: 10pt;
    MARGIN: 2px 0px
}
.AB
{
    BACKGROUND-COLOR: #6666ff;
    COLOR: #99ffcc
}
```

표준 HTML에 대한 변환 형태는 글꼴 및 전반적인 정보를 포함하며, 클래스 정보는 상위 문서정보 저장 영역에서 계승되는 부분을 제외한 변경된 영역만을 나타낸다. 문서 변환 과정은 PATTERN(많은 숫자 → 적은 숫자) → TAG 방향으로 진행된다. 적은 수의 태그가 먼저 변환됨으로서 발생 가능한 문서 구조의 불일치를 해소하였으며, 변환이 완료된 문서는 BODY 태그 내부에 존재하는 형태로 구성된다. 이러한 문서 구조는 기존에 템플릿 형태로 제작된 CSS 문서와 외부 연결 혹은 내부 템플릿 형태로 제공되는 Java 스크립트와 연동되어 HTML 형태로 실제 사용자 환경에 제시된다. 이러한 변환 과정을 도식화하면 그림 2와 같다.

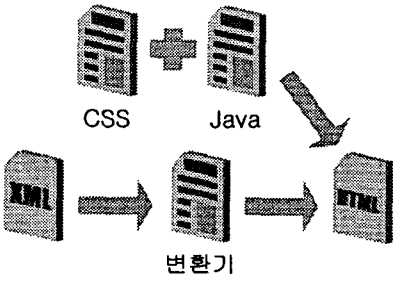

그림 2 XML 정보의 변환과정

4. 시스템 구성

언어학습이 이루어지기 위해서는 XML 기반 언어학습 정보가 학습자들에게 쉽게 전달될 수 있어야 한다. 이를 위해 XML 정보는 한 번 이상의 정보 가공 과정을 거친다. 문서 정보는 데이터베이스로부터 읽혀진 형태로 하나가 존재하며, 변환 과정에서 대치되는 영역만 새롭게 바뀐다. 패턴에 존재하지 않는 영역은 XML 태그 형태로 남겨진다. 이 경우 웹 환경에서 파싱 작업이 진행되지 않으므로, 사용자에게는 해당 정보가 제공되지 않는다.

XML 정보와 변환 정보는 모두 XML 로 구성되어 있어, 실질적인 문서변환 과정에는 두 개의 XML 정보 처리 클래스가 사용된다. 변환 정보에서 Pattern 정보를 추출하여 XML 정보의 태그의 개수를 줄여나간다. 패턴 과정이 완료되면 단편적 태그를 변환정보 XML 클래스에서 검색하여 다시 문서 내부에 일치하는 영역을 변경시킨다. 변환이 완료된 문서는 중괄호 영역이 모두 사용되며, 최종적으로 대괄호로 표현되어 있던 영역을 HTML

태그 형태로 변경시켜 실질적인 XML 정보로 제시한다. 이러한 과정을 도식화 할 경우 그림 3과 같다.

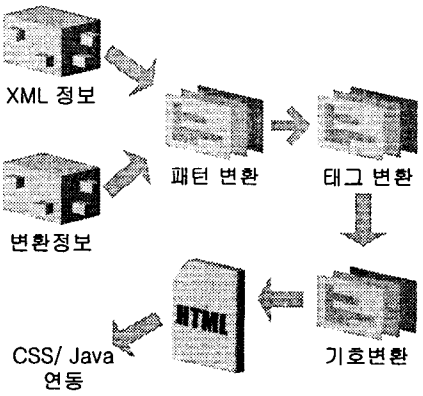

그림 3 정보 변환 구조

그러나 개별적인 XML 정보는 문서 정보 혹은 데이터베이스 형태로 저장되어 있어야 한다. 컨텐트 정보는 XML 형태 이상의 구조를 가지고 있어야 하며, 이는 단계적 언어학습을 위한 기본 구조이기도 하다. 그러나 본 앞장에서 제시한 구조가 제대로 구현되기 위해서는 그림 4와 같이 시스템 기반을 이루는 속성 정보가 데이터베이스에 함께 저장되어 있어야 한다.

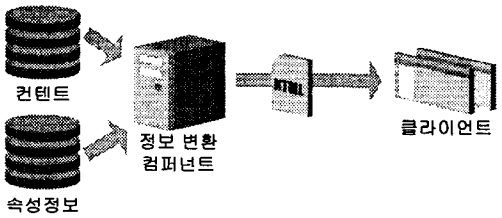

그림 4 XML 기반 다국어 학습 시스템 구조

컨텐트 정보는 속성정보와 함께 정보변환 컴퍼넌트를 거쳐 새로운 HTML 정보로 변환된다. HTML 정보는 웹 서버에서 연계 해 주는 CSS 및 Java Applet 과 연동되어 클라이언트로 전송된다. 전체적인 개념적 구조는 3-tier 구조를 가지나, 두 개의 데이터베이스 정보를 변환하는 컴퍼넌트의 구조에 의해 n-tier 구조 형태로 전체적인 시스템이 구성된다. 클라이언트 환경에서는 요구되는 특정 컨텐트의 정보를 서버로 전송하여 서버에서 가공된 정보를 다시 전송받는다.

5. 결론

본 연구결과 재사용이 가능한 한국어 및 다국어 컨텐트를 구성하고 이를 HTML 형태로 변환하여 실질적 언어학습 컨텐트로 활용할 수 있는 방안에 관해 연구하였다. 현재 본 문서 구조를 이용하여 기존의 학습 컨텐트를 새롭게 구성하고 있다. 그러나 본 연구는 다음과 같은 부분에 있어 추가적인 연구가 요구된다.

첫째, 다양한 컨텐트 유형을 단순화 한데 따른 속성의 다양화 및 특성의 지원이 요구된다. 둘째, Java 코드의 라이브러리화 및 API 화 하는 과정이 요구된다. XML 에서 변환되어 나타나는 문서 정보는 CSS 환경과 같은 표준화된 환경에서 쉽게 구현되고 관리될 수 있으나, Java 와 같은 프로그래밍이 요구되는 루틴은 언어학습 컨텐트에 적합하도록 기능적 선별 및 이를 라이브러리화 하는 과정이 요구된다. 셋째, 본 연구는 문서 구조에 관련된 것으로서, 실질적인 컨텐트 전체 구성에 대한 심도 있는 연구가 이루어지지 못했다. 효율적인 데이터베이스 구조로 컨텐트 전체 내용을 구조화 하고, 표준화할 수 있는 방안에 대한 연구가 요구된다. 넷째, 문제 정보는 새로운 데이터베이스 저장 및 관리가 요구되는 부분이다. 이에 대한 연구 및 데이터베이스 구조 설계, 특정요소 추출에 대한 연구가 요구된다.

참고문헌

[1] Aho, A.V./Sethi, R./ Ullman, J.D. (1993),*Compilers Principles, Techniques, and Tools(한국어판)*, 성안당, 서울.

[2] Date, C.J. (1990), *An introduction to database systems. Vol. I, fifth edition*, Addison Wesley, New York.

[3] Lauren, Simon, LT(1998), XML Primer, IDG Books, New York.

[4] Unicode Consortium(1996), The Unicode Standard 2.0, Addison Wesley, New York.

[5] Pardi, William J.(1999), XML in Action, Microsoft Press,

Redmond.

[6] 김홍규 외.(1998), 「21 세기 세종계획 국어 기초자료 구축」,문화관광부,서울.

[7] 한국전산원(1995),「국가기간전산망 표준화 연구중 - 한국형 SGML 문법지향 편집 시스템 지침(안)-」

[8] 한국전산원 (1995),「국가기간전산망 표준화 연구중 SGML 응용을 위한 지침서」

[9] 윤애선(1998), "그룹웨어 개념에 기반한 온라인 언어 교육", 「초고속정보통신 응용기술사업 결과발표 및 활성화를 위한 포럼 발표논문집」, pp.142~169.

[10] W3C Consortium(www.w3c.org/XML/, /CSS/)

[11] Unicode Consortium(www.unicode.org)

[12] Microsoft MSDN (msdn.Microsoft.com)