텍스트 및 전자사전 관리시스템의 설계

이재성*, 최병진**, 이운재*, 최기선*
*한국과학기술원 전산학과
**한국과학기술원 인공지능연구센터

The Design of Text and Dictionary Management System

Jae Sung Lee*, Byung Jin Choi**, Woon Jae Lee* and Key Sun Choi*
*Department of Computer Science, KAIST
**Center for Artificial Intelligence Research, KAIST

요약

자연언어처리 프로그램에서 어휘지식을 제공하는 전자사전은 그 중요성에 비해 작성 및 관리에 불편함이 많다. 본 논문에서는 전자사전의 작성 및 관리를 편리하게 할 수 있도록, 다양한 형태의 어휘 자료로부터 필요한 정보를 추출, 변형하고, 편집할 수 있는 텍스트 및 사전 관리시스템(TDMS: Text and Dictionary Management System)의 설계에 관하여 소개한다. TDMS 에서는 SGML(Standard General Markup Language)의 일부를 사용하여, 표준사전 표기언어(SDML: Standard Dictionary Markup Language)를 정의하고, 이를 이용하여 다양한 형태의 사전 형식을 기술하고 있다. 또, 표준사전 표기언어로 표현된 사전이나 텍스트는 각종 응용프로그램에 독립적인 형태로 존재하여, 정보의 표준화와 교환을 용이하게 한다.

1. 서론

자연언어처리 프로그램들은 대개 많은 양의 문법 정보, 의미 정보, 용례 등을 필요로 한다. 이러한 정보는 전자사전을 통해 제공이 되며, 제공되는 정보의 양과 질에 따라 프로그램의 성능도 많은 영향을 받는다.

그러나 전자 사전을 작성하는 일은 매우 노동 집약적이면서도 전문지식을 필요로 하며, 이러한 사전은 일단 완성이 되었더라도, 계속 새로운 단어를 첨가해야 하고, 경우에 따라서는 새로운 필드나, 정의 등이 추가됨으로써, 계속 새로운 형태의 사전을 만들어야 하는 경우가 많다.

효율적으로 사전을 구축하기 위해서는 사전 구축 과정 중에도 이미 입력한 단어나 내용

을 쉽게 확인하고 참고할 수 있어야 하며, 다른 사전의 내용을 참조하거나, 다른 사전의 내용을 복사할 수 있으면 좋을 것이다. 특히 이미 만들어진 여러 종류의 사전들과 텍스트 코퍼스 등에서 필요한 정보를 자동으로 추출할 수 있다면, 새로운 사전의 생성이나 사전의 갱신들을 효과적으로 할 수 있을 것이다[EDR 93]. 요즈음 컴퓨터에 입력된 종이 사전에서 사전 정보를 추출해 사용하는 사례들도 발표되고 있다[Alshawi 89, Amsler 88].

텍스트 및 사전관리 시스템(TDMS)은 다양한 형태의 어휘 자료로부터 필요한 정보를 추출, 변형하고, 편집할 수 있는 환경을 제공해 준다. 이를 위해서는 다양한 형태의 어휘 자료나 사전 등을 기술할 수 있는 방법과 이를 운용할 수 있는 방법이 있어야 하는데, TDMS 에서는 SGML(Standard General Markup Language)을 이용한 표준 사전 표기언어(SDML: Standard Dictionary Markup Language)를 정의하여 사용하고 있다.

2. SGML 시스템

SGML(Standard General Markup Language)은 마크업 언어의 구문을 정의하는 메타언어로 볼 수 있다. 이 SGML 을 이용하여 각 문서의 구조를 정의하면, 이 구조에 맞추어 실제의 문서들이 마크업된다. 이러한 문서구조의 정의를 DTD(Document Type Definition)로 부르며, 실제 마크업된 문서를 DI(Document Instance)로 부른다[Bryan 88, Goldfarb 90, 표준협 93].

SGML 의 특징은 문서의 논리적 구조를 정의하기만 하고 문서의 물리적 구조에 대한 것은 분리하여 정의할 수 있도록 한 것이다. 이러한 특징은 그 문서를 처리하는 응용프로그램이 어떠한 형태가 되더라도, 공통적인 문서 형태에서 필요한 부분만을 뽑아 내어 처리하는 것이 가능하도록 한다. 예를 들어, 출판과정에서 SGML 을 이용할 경우, 하나의 DTD 로 전체의 문서를 마크업해 두면, 편집, 교정, 구성, 삽화 등의 각 응용프로그램이 필요한 부분만을 다룰 수 있도록 할 수 있어서 정보의 처리를 보다 원활하게 한다[Travis 95].

TEI(Text Encoding Initiative) 프로젝트는 모든 문헌을 SGML 로 표현하기 위한 세계적인 프로젝트로서 그 표준형식을 단계적으로 발표하고 있다[Ide 95, McQueen 94]. 이 프로젝트에서는 현재 운문, 산문, 드라마, 구어체 코퍼스, 일반사전, 그리고 용어 데이터 베이스 등의 분야를 중점적으로 연구하여 표준 DTD 를 만들어 내고 있다. 그중 일반사전 분야에서는 주로 기존에 출판된 사전을 그 인쇄형식까지 모두 표현해 낼 수 있는 방법을 제안하고 있다.

컴퓨터를 이용한 자연언어 처리 분야에서는 이러한 기계 가독형 일반사전을 사용할 수도 있겠지만, 보다 효율적으로 표현된 전자사전을 작성하여 사용할 수 있다. 즉 자연언어 처리용 전자사전의 경우, 종이에 인쇄되는 형태를 고려하거나, 불규칙적인 사전 형식을 표현할 필요가 적으므로, 보다 단순한 형태로 전자사전의 구조를 표현할 수 있다[Ide 95,Boguraev 89, Boguraev 94].

3. 텍스트 및 전자사전 관리시스템의 구조

TDMS 시스템 구조

TDMS 시스템은 크게 3 개 요소로 구성된다. 표준 사전 표기 언어(SDML)로 그 사전 형식이 정의되어진 표준 사전(SD), 표준 사전을 생성, 편집, 검색 등을 할 수 있는 사전편집기(SDE)와, 표준 사전이 아닌 기존의 사전들을 표준 포맷으로 바꾸어 주고 또, 표준 포맷을 원하는 특수 포맷으로 바꾸어 주는 변환 프로그램(SD Encoder/SD Decoder)들로 구성된다.

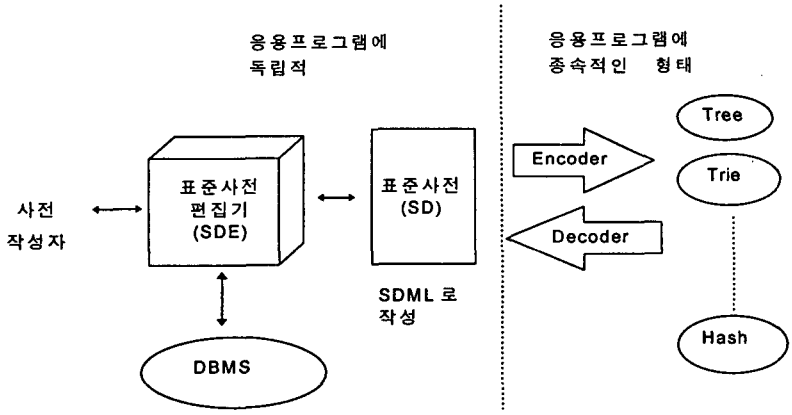
표준사전(SD)

SD 는 SDML(Standard Dictionary Markup Language)로 작성된 사전이며, 각 사전마다, SDML 이 허용하는 범위 내에서 다양한 형태를 가질 수 있다. SD 는 SDML 에서 정의된 구조로 만들어 지고, 각 구조에는 SDML 에 정의된 태그가 부착된다. 사전의 각 내용들은 모두 텍스트 형식으로 표시되며, 따라서 특별한 프로그램 없이도 바로 내용을 확인해 볼 수 있다. SD 의 앞 부분에는 사전의 구조를 나타내는 부분이 있어서 그 앞부분의 파싱을 통해 뒤에 오는 사전의 형태를 올바르게 파악할 수 있다.

이런 사전의 구조는 사실상 SGML 의 문서 구조 정의(DTD: Document Type Definition)로 기술된 것이다. 따라서 각각의 사전마다, SDML 규칙에 근거한 사전 구조를 나타내는 SGML DTD 의 일부가 포함되어 있으며 이러한 각각의 형식을 표준사전 형식(SDF: Standard Dictionary Format)으로 부른다.

표준사전 편집기(SDE)

표준사전 편집기는 표준사전(SD)의 내용 및 구조를 편집하고 수정할 뿐만 아니라, 내부의 내용을 검색 또는 브라우징(browsing) 할 수 있다. 한 화면에는 기본적으로 한 표제어 단위로 표시되며, 표시량이 많을 경우, 스크롤하여 표시한다.



TDMS 시스템 구조

실제 사전의 경우, 표제어의 종류에 따라 가변적인 내부 구조를 가질 수 있으므로, 그 표시를 가변적으로 할 필요가 있다. 이를 위해서 표준사전(SD)의 구조를 파악하여 각 표제어 표시시에 동적으로 화면을 구성하여야 한다. 이와 관련된 기본 기능으로는 다음과 같은 것이 있다.

- SDML로 만들어진 DTD를 해석하여 내부 구조를 정할 수 있는 기능
- 표준 사전을 인식하여 하나의 내부구조에 각 데이터를 연결시키는 기능
- 내부구조를 선형의 텍스트 형식(linear form)으로 바꾸어 표준 사전으로 만드는 기능

사전을 변경하는 경우, 크게 내용변경과 구조변경으로 나눌 수 있다. 내용변경은 구조의 변화가 없으므로 각 필드의 데이터만 수정하거나, 이미 정의된 구조내에서 필요한 필드의 추가나 삭제만을 하면 된다. 그러나 구조 변경의 경우에는 편집기에서의 기본 처리 구조가 바뀌게 되므로 그에 관련된 데이터에 대한 처리가 문제가 된다. 따라서 구조적 변형에 따른 각 데이터의 이동 관계 등도 정의되어야 한다. 현재 이러한 부분은 SGML의 DSSSL(Document Style Semantics and Specification Language)프로젝트에서 각 내부구조의 변형방법에 대한 연구가 진행되어 있으므로 이를 참조하여 이용할 계획이다[ISO 94].

사전변환프로그램
(SD Encoder/SD Decoder)

기존의 사전형태는 사용되는 곳의 필요에 따라 여러 가지 구조의 이진화일, 또는 텍스트화일 등으로 저장되어 있다. 이러한 형태의 사전을 표준사전 형태로 바꾸어 주는 것이 Decoder이며, 반대로 필요한 응용프로그램에게 맞도록 표준사전을 변형시켜 주는 프로그램이 Encoder이다.

응용프로그램에서 사용되는 사전형태는 너무 다양하기 때문에 TDMS의 초기 시스템에서는 주로 많이 쓰이는 몇 가지 형태의 사전만을 표준으로 지원해 준다. 지원되지 않는 형태는 각자 SD Decoder를 개발하여 표준사전 형식으로 변형시켜주면, 표준사전 편집기를 사용할 수 있다. 또한 Decoder와 Encoder개발을 쉽게 할 수 있도록 기본 형태의 모듈을 제공하여 개발자가 그 사전에 특수한 기능만을 추가하면 표준사전 Decoder/Encoder가 완성될 수 있도록 한다.

4. 표준 사전 형식 및 표기 언어(SDML)

SDML 정의

표준사전 형식은 TDMS의 중심에서 모든 정보를 포함한다. 이러한 정보는 표준사전 편집기(SDE)에서 필요한 편집 및 관리에 필요한 정보(관리정보)와 응용프로그램 쪽의 사전에서 필요한 정보(사전정보)로 구분될 수 있다. 관리정보의 예로는 사전의 각 엔트리를 표현하기 위한 형식 표현이라든지, 하이퍼 텍스트를 위한 포인터 태그 등을 들

수 있다. 또 사전 정보로는 사전의 형식과 내용을 SGML 의 기능으로 정의하는 것이다. 일반적인 사전은 표제어를 인덱스로 하여 엔트리가 반복되어 나타난다. 이 엔트리는 표제어의 성격에 따라서 엔트리 내부의 구조와 내용들이 변한다. 또, 엔트리 내의 필드에서 특정한 부분을 참조하기도 한다. 자연언어 처리용 프로그램에서도 이러한 기본 형식이 유지된다. SDML 은 사전형태를 표현하는데 필요한 SGML 의 일부 기능만을 제한적으로 사용하도록 한 것이다.

SDML 의 구조

SDML 은 크게 헤더, 정의부분, 엔트리그룹으로 구성되어 있다. 헤더는 이 사전의 이름, 버전 등의 내용을 포함하고 있고, 정의부분(front)에서는 사전 내부에서 쓸 각 속성들이 가질 수 있는 값을 규정하고 있다. 엔트리 그룹은 실제 사전의 내용이 포함되어 있으며, 표제어를 포함하는 각 엔트리들의 반복으로 구성된다.

```
<sd>
  <sdHeader> [헤더 정보] </sdHeader>
  <group>
    <entry>
      <wname> [표제어] </wname>
      <body> [표준 사전 요소] </body>
    </entry>
    <entry>
      <wname> [표제어] </wname>
      <body> [표준 사전 요소] </body>
    </entry>
    [ entry 의 반복 ]
  </group>
</sd>
```

표준 사전 요소

이 부분은 <body>에 포함되는 요소로서 기본 설정값처럼 단순한 하나의 필드로 구성될 수도 있고, 복잡한 트리구조를 가질 수도 있다. 이를 위해서 사용자는 필요한 사전의 구조를 정의해서 <body> 의 구조를 변경하여 사용한다. 일반적으로 새로운 구조의 정의는 SGML 형식과 동일하게 정의하지만, SDML 에 제한적으로 사용해야 하는 규칙이 있다.

1. SGML content model 중 사용 가능한 기본 요소는 , | * + ?로 한다. &는 사전 형식을 결정하는데 무의미하므로 제외한다.
2. 이미 정의된 태그가 있으면, 그 태그 이름을 그대로 사용하여 호환성을 가질 수 있도록 한다.

<body>는 초기에 다음과 같이 정의되어 있다.

```
<!ELEMENT body  - - (#PCDATA)* >
```

따라서 새로운 구조로 변형하고자 할 경우, 이 element 만을 재정의(overwrite)한다. 다음은 <body>구조를 <품사>와 <대역어>를 포함하는 새로운 구조로 변형시키는 예이다.

```
<!DOCTYPE   sd SYSTEM "sdml.dtd"   [
<!ELEMENT body  - o (품사, 대역어) >
<!ELEMENT 품사  - o (#PCDATA) >
```

<!ELEMENT 대역어 - o (#PCDATA) >
]>
<sd>
<!-- 위에 정의된 구조(SDF)에 맞는 sd 의 내용 -->
:
</sd>

5. 맺음말

이상으로 전자사전의 표준 형식과 그 표준 형식을 중심으로 한 텍스트 및 전자사전 관리시스템(TDMS)의 구조를 설명하였다. TDMS 의 구조는 표준사전 형식을 통해 이미 작성된 사전의 내용이나, 작성 중인 사전의 정보 공유, 관리, 유지를 편리하게 해준다.

현재 전자사전의 구조나 그 구조로 정해진 필드의 내용에 대한 것은 매우 일반적이거나 제한이 없도록 했다. 그러나 그 구조와 내용에 대한 표준화가 진행되면, 그 표준 구조 및 내용을 포함할 계획이다[과기원 96a, 과기원 96b, 최병진 96].

현재 TDMS 의 기본 기능을 갖춘 1 차 버전이 완성되어 테스트 중이며, SGML 의 기능을 최대한 사용할 수 있는 구조 검색 [Macleod 90], 하이퍼 텍스트[Ide 95], 다른 구조의 사전 병합[ISO 94] 등에 대한 확장이 계획되고 있다. 또한 텍스트는 현재 한 화일 단위를 한 엔트리로 취급하여, 사전과 동일한 방법으로 관리하도록 하고 있지만 [과기처 96], 텍스트 내부에서 사용하는 대부분의 마크업은 아직 처리하지 않고 있다. 이에 대한 처리도 계속 연구되어야 한다.

6. 참고문헌

[Alshawi 89] Alshawi, H. (1989), "Processing Dictionary Definitions with Phrasal Pattern Hierarchies", in Boguraev, B. and E. Briscoe(eds.), 153-170.

[Amsler 88] Amsler, Robert A. and W. Tompa(1988), "An SGML-based Standard for English Monolingual Dictionaries, in Proceedings of the 4th Annual Conference of the UW Centre for the New Oxford English Dictionary, Waterloo, Ontario: 61-80.

[Boguraev 89] Boguraev, B. and E, Briscoe(1989), Computational Lexicography for Natural Language Processing, Longman Limited, Harlow and London .

[Boguraev 94] Boguraev, B. (1994) Machine-readable dictionaries and computational linguistics research, in A. Zampolli, N. Calzolari, and M. Palmer(eds.) pp119-149.

[Bryan 88] Bryan, M.(1988), SGML: An Author's Guide to the Standard Generalized Markup Language, Addison-Wesley.

[EDR 93] EDR(1993), EDR Electronic Dictionary Technical Guide, Japan Electronic Dictionary Research Institute.

[Goldfarb 90] Charles F. Goldfarb, " The SGML Handbook," Clarendon Press, Oxford, 1990.

[Ide 95] Nancy Ide and Jean Veronis, "Text Encoding Initiative Background and Context," Kluwer Academic Publishers, 1995.

[ISO 94] "Information technology - Text and office systems - Document Style Semantics and Specification Language(DSSSL) - Draft", ISO/IEC DIS 10179.2, 1994.

[Macleod 90] Ian A. Macleod, "Storage and Retrieval of Structured Documents," Information Processing & Management Vol 26, No 2. pp 197-208, 1990.

[McQueen 94] Sperberg-McQueen and Lou Burnard, "Guidelines for Electronic Text Encoding and Interchange(TEI P3)," Vol I and Vol II, ACH, ACL, ALLC, April 8, 1994.

[Travis 95] Travis, B. E. and Waldt, D. C. "The SGML Implementation Guide," Springer-Verlag Berlin Heidelberg, pp3-20, 1995.

[Zampolli 94] Zampolli, A., Calzolari, N., and M. Palmer(1994), Linguistica Computationale, Vol. IX.X Current Issues in Computational Linguistics: in Honor of Don Walker, Kluwer Academic Publishers, Dordrecht.

[과기원 96a] "문서구조 표현을 위한 표준화에 관한 연구," 제 1 차년도 최종보고서, 한국과학기술원, 1996.

[과기원 96b] "제 1 회 우리말 정보처리 규격 심포지움," 한국과학기술원 인공지능연구센터, 1996 년 7 월 11 일.

[과기처 96] "통합 국어정보베이스," 제 2 차년도 최종보고서, 과학기술처, 1996.

[최병진 96] 최병진, 이재성, 이운재, 최기선, "표준화를 위한 일반사전의 논리구조," 한글 및 한국어 정보처리, 1996.

[표준협 93] "문서 기술 언어 SGML," KS C 5913-1993, 한국표준협회, 1993.