

대화 코퍼스의 구축 및 주석 정보의 구조적 문서화

강창규, 김영일, 김봉완, 이용주
원광대학교 컴퓨터 공학과, SiTEC

Construction of Dialogue Corpus and Structured Documentation of Annotation Information

Chang-qui Kang, Young-il Kim, Bong-wan Kim, Yong-ju Lee
Department of Computer Eng., Wonkwang Univ., SiTEC

요약

음성인식의 연구 대상은 낭독음성에서 대화음성으로 발전해가고 있다. 이를 위해서는 대량의 대화코퍼스가 필요하다. 그러나 아직 충분한 양의 대화코퍼스가 구축되어 있지 못하며 코퍼스의 주석 정보 또한 복잡하고 다양하게 표현하고 있어 효율적인 활용이 어렵다. 따라서 본 논문에서는 대화 영역으로 텔래뱅킹 영역을 설정하고 대화코퍼스를 구축하여 구축된 대화코퍼스의 주석 정보를 XML(Extensible Markup Language)로 표준화할 수 있도록 DTD(Document Type Definition)를 정의하여 문서 구조화하였다.

1. 서론

대화음성의 인식이나 합성 등으로 대표되는 대화정보처리 기술의 개발에 가장 기본적인 연구 자원은 대화코퍼스(Dialogue Corpus)이다. 자연어 처리에서 텍스트 코퍼스와 마찬가지로 대화의 경우에는 통계적 처리 방법이 주류를 이루고 있어 대화코퍼스의 중요성 및 역할은 막중하다. 대화코퍼스는 음성을 기록하여 보존하고 다양한 색인 정보도 가지고 있다. 따라서 지정한 단어 또는 문장을 바로 찾을 수 있고 대화의 주석 정보를 포함한 자료들을 검색해 볼 수 있다[1]. 또한 화자정보(성별, 연령, 사투리, 출생지)도 포함되어 있어 발성자에 따른 여러 대화 현상들도 분석해 볼 수 있다.

이와 같이 다양하고 복잡하게 나타나는 대화코퍼스 주석 정보는 표준화되어 체계적으로 관리되어야 한다. 따라서 대화코퍼스의 주석 정보를 효과적으로 사용하기 위해 XML 언어로 표준화하여 대화코퍼스를 문서 구조화하는 방법을 검토하고자 한다. 구축된 대화코퍼스에서 주석 정보는 XML로 표준화하고 DTD로 작성하여 문서 구조화한다. 대화코퍼스에서 나타나는 주석 정보를 XML 기반으로 기술함으로써 모든 대화코퍼스의 수집 및 대화 시스템을 설계할 때, 특정한 어플리케이션에 종속되지 않고 데이터의 독립성을

보장할 수 있다. 또한 새롭게 나타나는 주석 정보에 확장 가능한 태그를 통해 구조화된 문서에 추가함으로써 대화코퍼스 정보 간에 재사용이 용이해진다.

본 논문에서 2장은 대화코퍼스 구축 절차에 대해 살펴보고 3장에서 대화코퍼스 데이터의 전사(transcription)에 대해 설명한다. 4장에서는 DTD에 의해 정의된 대화코퍼스의 문서 구조화에 대해 살펴보고 마지막으로 5장에서 결론을 맺는다.

2. 대화코퍼스 구축

대화코퍼스는 대화의 모든 정보들이 나타나기 때문에 대화 연구에 필수적이다. 대화코퍼스는 대화의 상호작용에서 나타나는 표현과 여러 유형에 따라 대상 영역을 설정할 수 있다. 본 논문에서는 텔래뱅킹 영역을 설정하여 대화코퍼스를 구축하였다.

2.1 시나리오 설계

대화코퍼스에서 텔래뱅킹 영역의 자연스러운 문장들을 유도하기 위해서 시나리오의 사용은 중요하다. 시나리오는 각 화자가 시스템에서 사용하는 실제 텔래뱅킹 상황의 서술로 정의된다. 본 논문에서는 실제

텔레뱅킹 영역을 참조하여 <표 1>과 같이 도메인 영역을 세부 영역으로 분류하였다. 또한 텔레뱅킹 시스템의 인식오류에 대한 사용자의 반응을 살펴보기 위하여 시스템에서 발생 가능한 오류를 가상으로 재현하기 위해 수집 대상 영역에 포함하였다.

| 서브 도메인 | 내용 |
|--------|--|
| 공과금 납부 | 전화 요금 납부, 전화 요금 납부 내역 조회, 아파트 관리비 납부, 아파트 관리비 납부 내역 조회 |
| 신용카드 | 결제 대금 조회, 사용 내역 조회, 사용 한도 조회, 현금 서비스 이체, 사용 명세서 주소지 변경 |
| 분실신고 | 현금 카드 분실, 직불 카드 분실, 수표 분실, 통장 분실, 인감 분실 |
| 서비스 | 우편 서비스, 팩스 서비스 |

표 1. 시나리오 영역

2.2 녹음환경과 방법

텔레뱅킹 영역의 데이터의 수집 환경은 일상적인 분위기를 만들어낼 수 있는 사무실에서 수행하는 것으로 하였다. 발성 전에 간단한 예행연습으로 발성자의 긴장감을 완화시킨다. 녹음 방법에서 태스크 설명은 화자들이 발성할 일정표를 배포하고 발성방법 및 상황을 설명한다. 시나리오의 기본 상황은 텔레뱅킹을 위한 태스크이다.

| 장비명 | 기능 | 용도 |
|---------------------|-----------------|--------------------------------|
| Teac Tascam M-08 | Mixer | 녹음 채널 구성 |
| Sennheiser HMD 25-1 | Headset | 발성자 착용 헤드셋 |
| Dialogic D/4 PCI | Telephony Board | 전화 음성 녹음을 위한 PC interface card |
| SoundBlaster Live | Sound Card | 마이크 채널 챔플링 |
| 유선전화기 | Sound Card | 마이크 채널 챔플링 |
| Pentium III PC | Computer | 녹음 컨트롤 |

표 2. 사용 장비

2.2 녹음 시스템

텔레뱅킹 녹음시스템은 WOz(Wizard of Oz) 시뮬레이션을 통해 화자의 발화를 녹음하고 Wizard의 버튼을 눌러 TTS(Text to Speech)로 시스템 발화 후 사용자의 자연스러운 발화를 유도해 낸다. 대화 형식의 시나리오를 시스템에 입력하면 시스템 프롬프트는 자동 발화가 되고 그것을 듣고 난 후, 시나리오에 의해 사용자가 발화하면 녹음이 되어 하나의 파일로 저장된다.

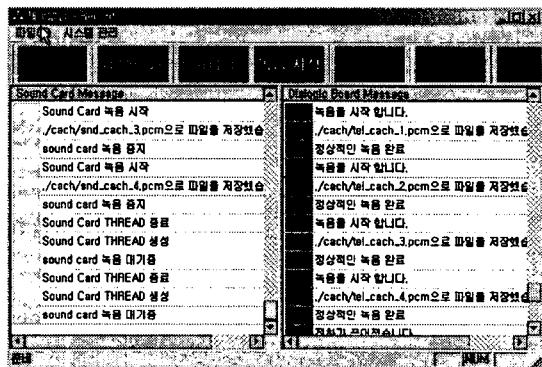


그림 1. 녹음 시스템

3. 대화코퍼스 전사

전사는 발성된 음성을 철자 및 발음표기를 표현하는 것을 말한다. 이를 이용하여 대화의 음향 모델과 언어모델의 훈련에 이용할 수 있다. 음향 모델을 훈련하기 위해서는 발성한대로 표기가 필요하며 언어모델을 위해서는 사투리나 잘못 발성된 단어들에 대해서는 표준 표기로 바꾸어 주는 것이 필요하다. 발성된 표기함을 기본 원칙으로 하고 이에 표준 표기를 덧붙여 준다. 전사된 대화 코퍼스 정보의 예는 다음과 같다.

비유창성(disfluency)은 일관성 있는 발화의 구조에서 벗어나는 대화 음성 언어의 특징이며 그 유형에는 간투어, 대용, 반복, 삽입, 삭제, 음성 에러 등이 있다 [2]. 잡음(noise)은 대화 현상에서 빈번하게 발생하게 된다. 사람에 의해 발생될 수도 있고 사람이 아닌 요소에 의해 발생 될 수도 있다.

중첩(overlap)은 상호간의 대화에서 한 참가자 또는 여러 참가자에 의해 동시에 발생하는 중복된 음성이다. 대화에서 흔히 발생하며 이에 대한 구분 기준은 3 가지로 분류된다. 첫째, 수직 정렬 방법은 기술적인 문제가 있으며 순서에 잘 맞지 않는다. 둘째, 분리 정렬 방법은 가장 이상적인 방법이지만 중첩 부분을 분리해서 체크하기 때문에 중첩 부분을 정확히 찾을 수 없다. 셋째, 시간 정렬 방법은 두 사람 이상의 중첩 부분도 표현이 가능하다. 또한 발생 구간도 정확히 확인할 수 있어 본 논문에서는 시간 정렬 방법을 사용하였다.

편집(comment)은 마침표, 물음표, 느낌표, 쉼표로 나타내고 문맥적인 의미를 파악하여 표시한다. 외국어

(foreign)는 일반 사용자들이 사용하는 외국어 및 외래어를 의미한다. 축약어(verbalism)는 실제 대화 음성에서 많이 나타나며 한국어의 경우 구어체에서 많이 쓰인다. 사투리(dialect)는 공통어나 표준어와는 다른 지역에서 사용되는 특유한 변이 형태다. 아래 <그림 1>은 대화코퍼스의 전사 정보를 나타낸다.

| | |
|-------------|-------------|
| <TEI> | <header> |
| <text> | <utterance> |
| <speaker> | <turn> |
| <foreign> | <language> |
| <verbalism> | <origin> |
| <dialect> | <standard> |

Table data:

| .TEI | text | utterance | turn |
| text | utterance | turn | language |
| speaker | turn | language | standard |
| foreign | language | origin | standard |
| verbalism | origin | | |
| dialect | | | |

그림 2. 텔레뱅킹 전사 정보

4. 문서 구조화

본 논문에서는 대화코퍼스에서 대화의 다양한 주석 정보를 표준화하여 체계적으로 표현하기 위해 XML을 사용하였다. 대화코퍼스의 다양한 주석 정보를 의미태그로 선정하여 표기 방안을 표준화 하였다.

4.1 태그 정의

대화코퍼스의 주석 정보를 XML로 표기하기 위해서는 먼저 의미 태그를 선정하고 DTD를 정의해야 한다. 대화코퍼스의 화자 정보와 주석 정보의 의미 태그는 TEI(Text Encoding Initiative)와 LDC(Linguistic Data Consortium)의 기준을 기반으로 작성하였다[3]. 기본적인 설계는 3개의 정보로 나누어 의미태그와 속성을 분류하였다. 우선 대화코퍼스의 화자정보와 텍스트 정보, 주석 정보를 바탕으로 의미태그를 할당했으며 추후 추가되는 정보는 의미태그를 정의하여 추가/삭제할 예정이다.

| 태그 | 속성 | 의미 |
|---------|----------|------|
| speaker | name | 이름 |
| | sex | 성별 |
| | region | 지역 |
| | date | 발생날짜 |
| | language | 언어 |

표 3. 화자 정보

| 태그 | 속성 | 의미 |
|-----------|-----|------------|
| text | | 실질적인 대화 정보 |
| utterance | who | 발화자 |
| turn | | 한 화자의 발화 |

표 4. 텍스트 정보

| 태그 | 속성 | 의미 |
|------------|----------------|-----------|
| disfluency | pause | 간투사 |
| | substitution | 대용 |
| | repetition | 반복 |
| | insertion | 삽입 |
| | deletion | 삭제 |
| noise | cg | 기침소리 |
| | h | 숨쉬는 소리 |
| | lg | 웃음소리 |
| | ls | 입술소리 |
| | sp | 숨소리 |
| overlap | other | 이외의 사람 잡음 |
| | overlap_length | 중복 |
| | P | 마침표 |
| comment | C | 쉼표 |
| | Q | 물음표 |
| | E | 느낌표 |
| | foreign | 외국어 |
| verbalism | origin | 축약되기 전 단어 |
| dialect | standard | 방언 |

표 5. 전사 정보

4.2 DTD 정의

대화코퍼스 주석 정보를 체계적으로 표현하기 위해 DTD를 설계하였다. 대화코퍼스 DTD 전체적인 구조는 header DTD와 text DTD로 분류하여 구조화하였다[4]. header DTD에는 화자 정보 및 대화 코퍼스 제목 정보를 나타내며 text 정보에는 front, body, back 정보로 구성된다. <그림 3>은 DTD의 전체 구성을 나타내며 <그림 4>은 의미 태그를 DTD로 정의하여 XML 편집기로 표현하였다. <그림 5>는 대화코퍼스 주석 정보를 XML 문서로 작성한 것이다.

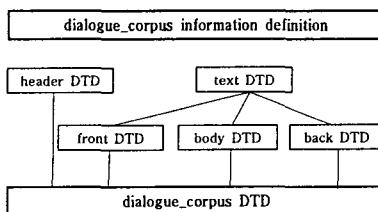


그림 3. DTD 전체 구성도

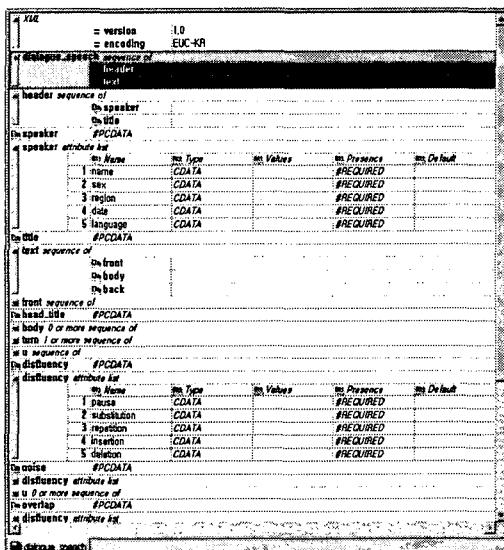


그림 4. DTD 설계



그림 5. XML 문서

5. 결론

자연스러운 대화 연구에는 다양한 언어정보 및 대화코퍼스가 필요하다. 대화의 언어모델과 음향모델의 성능 평가에 주석 정보는 큰 영향을 끼친다. 본 논문에서는 이러한 대화코퍼스의 주석 정보를 XML 언어로 표준화 하여 문서 구조화하였다.

또한 설계된 DTD기준으로 대화코퍼스 정보를 효율적으로 관리할 수 있는 문서를 생성하였다. 이러한 문서의 표준화는 공통적으로 나타나는 대화코퍼스의 정보 검색 및 연구에 매우 유익한 정보의 추출이 가능하다. 본 논문에서는 대화코퍼스의 각종 주석 정보에 대한 상세한 기술이 모두 정의되어 있지 않지만 이를 수용할 수 있는 기본 구조 설계에 초점을 두었다.

향후 연구과제로는 각종 주석 정보의 상세한 기술 기준의 표준화뿐만 아니라 이를 효율적으로 이용할 수 있도록 충분한 검토와 표준화에 대한 노력이 필요하다.

[참고문헌]

- [1] Akira Kurematsu, Yousuke, Shionoya "Identification of Utterance Intention in Japanese Spontaneous Spoken Dialogue by Use of Prosody and Keyword Information", pp.98-101, ICSLP, 2000.
- [2] 남승석, "대화 음성 시스템을 위한 효과적인 DB 구축", 서강대학교 컴퓨터학과 석사 논문, 2002.
- [3] Dafydd Gibbon, Inge Mertins, Roger K. Moore, Handbook of Multimodal and Spoken Dialogue System, pp.1-25, 2000.
- [4] Francois, "Generalized SGML repositories: Requirements and Modeling", Computer Standards & Interfaces, pp.11-24, 1996.