어휘 정보를 이용한 문장완성의 구현

황인정, 이은실, 민홍기
인천대학교 정보통신공학과

Implementation of Sentence Construction using Lexical Information

Einjeong Hwang, Hongki Min
Dept. of Information and Telecommunication Eng., Univ. of Incheon

요 약

본 연구는 어휘 정보를 이용하여 구어체 문장구성을 하였다. 구어체 문장구성의 목적은 언어생활이 불편한 사람들을 위한 통신보조기기에 사용하기 위해서이다. 통신보조기기는 사용자가 원하는 문장을 만들어 음성으로 출력해주는 시스템이다. 그러므로 문장을 구성하기 위해서 어휘 정보를 통신보조기기의 개념에 맞도록 변형하여 도입하였다. 어휘는 도메인별로 발췌하고 분류하였으며, 각 어휘에 대해 시소러스와 하위범주화사전을 만들었다. 어휘정보에 관한 상세한 정보는 문장구성과 재사용 그리고 문맥상 어색한 문장검출을 위해 중요한 자료가 된다.

I. 서 론

통신보조기기에 적용을 목적으로 어휘정보를 이용하여 문장구성을 하였다. 통신보조기기는 일상적인 언어생활이 불편한 사람들을 위한 보조기기이므로, 사용자의 특성과 환경에 따라 입력방법의 다양성, 언어능력에 따른 어휘의 수와 내용이 다를 수 있고, 기기의 크기와 부착형태 등 여러 모델이 존재하지만, 기본적인 형태는 대화에 필요한 문장을 만들어 음성으로 출력하는 기기이다. 이러한 요구에 부응하기 위해서는 각 부분에 대한 연구와 함께 다양한 방법들이 적용되어야 한다. 본 연구에서는 사용자 인터페이스에 맞는 문장구성에 초점을 맞추었다. 문장구성 방법으로는 어휘 정보를 적절하게 변형하여 적용해 보았다.

II. 어휘 정보를 이용한 문장구성 방법

일상생활에 쓰이는 구어체 문장구성을 위한 어휘정보를 수집하기 위해서는 어휘의 발췌와 분류가 필요하다. 어휘정보는 사용자가 원하는 문장을 만들고 사용된 문장의 재사용 그리고 에러처리를 위한 여러 목적으로 사용되어야 하기 때문에 초기 어휘정보 체계의 구축이 중요하다. 어휘 정보는 문장구성을 위한 어휘 관련부분의 정보로 상세히 분석하여 구축된다.

1. 어휘의 발췌와 분류

구어체 문장구성을 위해서는 가장 먼저 어휘발췌를 하여야 하며, 어휘발췌는 사용자의 환경을 고려하여 발췌되어야한다. 사용자의 환경이란, 신체적, 정신적 환경을 말하는 것이다. 신체적 특징에 따라 통신보조기기의 모델이 변하지만, 그 부분은 본 연구에서는 제외하고 문장구성에 초점을 맞추었다. 문장구성을 위한 환경으로는 사용자의 어휘와 자주 이용하는 장소를 들 수 있으며, 여기에 덧붙여 사회적으로 통용되는 연령별 어휘에 대한 자료가 포함된다. 사용자의 어휘능력과 이용 장소만을 사용하여 어휘발췌를 한다면 한정된 어휘사용으로 어휘능력의 향상을 기대할 수 없으므로, 사회적으로 통용되는 어휘자료를 추가하여야 한다. 그러나 아직 사회적으로 통용되는 연령별 어휘가 정립되어있지 않은 관계로 본 연구에서는 어린이를 사용자로 정하여 어휘발췌를 하였으며, 어휘발췌는 교과서와 회화 책, 인터넷을 통한 자료모음과 수화 사이트 그리고 몇 사람에 의해 수집된 어휘와 논문 등을 참고하였다. 장소도메인은 교통, 집, 쇼핑, 병원으로 정하였다. 장소도메인에 따른 어휘를 발취하고, 도메인에 국한하지 않은 중심어휘도 발췌한다. 중심어휘는 인사말, 도움말을 포함하고 있다. 표 1은 중심어휘의 예를 보여준다.

<표 1> 중심어휘의 예

안녕하세요. 도와주세요. 고맙습니다. 미안합니다, 괜찮습니다. 반갑습니다. 예. 아니오.

중심어휘와 장소도메인에 따른 문장을 발췌하고 특성도메인의 어휘도 발췌한다. 특성도메인은 색, 숫자, 시간 날짜 등으로 분류할 수 있다. 어휘 발췌의 목적은 특정 장소나 특정상황에서 사용자가 쉽게 어휘를 선택하고자 하는데 있으며, 어휘의 범위를 줄인다면 문장구성에 대한 신뢰도의 증가를 가져올 수 있다. 발췌된 어휘는 다양한 문장구성과 각 어휘별 연관성에 대한 정보 수집을 위하여 일정하게 분류하였다. 어휘분류는 한정된 어휘를 이용하여 많은 문장을 만들 수 있도록 패턴별 문장구성방법을 선택하였다. 패턴별 분류는 크게 명사부와 동사부로 나눈다. 이 방법은 문법적으로는 다소 논란의 요지가 있으나, 동사에 따라 명사의 종류를 한정할 수 있고, 해당

명사를 바꾸어 많은 문장을 만들 수 있다. 그리고 명사의 범위가 한정되므로, 새로운 어휘가 입력되었을 때도 기존의 문장 패턴을 이용할 수 있는 장점이 있다.

발췌된 문장은 표 2에서처럼 분류하였다. 발췌된 문장은 장소나 특성 도메인이름을 대분류로 하였고, 패턴별 이름을 정하여 소분류 명으로 구성하였다. 소분류 명 아래에 다시 소분류명이 존재할 수 있다.

<표 2> 어휘 분류의 예 (교통도메인)

(장소)에서 내리려면 얼마나 걸려요? (시간)
몇분이나 걸려요? 걸립니다. 갈까요. (시간)
(장소)는 어디에 있습니까? 있습니다. (거리)
(장소)는 어떻게 가야 합니까? 가야 합니다.(거리)
(장소)역까지는 얼마입니까? (요금)
교통 카드값(요금)은 얼마입니까? (요금)
잔돈(요금)이 없습니다.(기타)

표 2에서와 같이 발췌된 문장을 분류한다. 교통도메인에서 명사부의 소 분류명은 장소, 수단, 요금으로 동사부의 소 분류명은 시간, 거리, 요금, 기타로 하였다. 각 도메인은 표 3에서와 같이 시소러스 계층을 만드는데, 분류명의 이름은 분류자의 의도에 따라 달라질 수 있다.

2. 어휘정보의 구성

문장 패턴 분류가 끝나면 도메인별 시소러스를 구축한다. 구어체 문장구성을 위한 시소러스는 기존의 모든 어휘에 대한 시소러스 계층 구조와는 차이가 있다. 일반적인 시소러스는 생물, 무생물로부터 시작하여 동물, 식물, 광물 그리고 다시 계층적인 분류를 하여, 모든 어휘가 특정 계층에 속하도록 하는 구조를 가지고 있다. 그러나 도메인에 의해 발췌된 어휘를 위한 시소러스는 도메인별 문장 발췌로부터 만들어졌으므로, 전체 시소러스의 계층구조가 필요하지 않고, 도메인별 시소러스 구조가 필요하므로 그 범위는 작다. 예를 들면 교통도메인의 경우 발췌된 문장의 분류로부터 시소러스를 생성한다. 문장 분류시 장소, 교통수단, 시간, 거리, 요금, 기타를 소분류명으로 하며, 소분류명인 장소는 지명, 역명, 위치로 나뉘어진다. 시소러스는 세분화하여 나눌 수 있는 계층적 구조를 가지고 있다. 표 3은 교통도메인에서의 시소러스 계층을 나타낸다. 시소러스 계층에서 장소, 수단에 포함된 명사 어휘들은 같은 계층의 동사부와의 연결이 용이하다. 그러므로 이미 연결이 가능한 계층부터는 아래 계층의 어휘들을 포함하여 어휘연결로 문장구성이 가능하다.

<표 3> 시소러스의 예 (교통 도메인)

교통 =>장소 -> 지명 ->
-> 역명 ->
-> 위치 ->
수단 ->
시간 ->
거리 ->
요금 ->

어휘의 발췌와 분류로 구성된 시소러스에 의해 말뭉치 사전을 작성한다. "말뭉치 사전은 다양한 문장들의 모음으로 규정되어 있고, 언어 연구를 목적으로 한 말뭉치는 어느 수준 이상의 크기를 지녀야 하며, 분야별로 정리된 특수한 말뭉치와 언어 습득에 따른 말뭉치 또는 일반적인 구성분포를 지니는 말뭉치등의 다양한 형태를 지니고 있다."라고 정의되었다. 본 연구에서의 말뭉치는 유사한 단어들의 모음으로 정하여 유사단어들을 확장하는 방법을 선택하였고, 말뭉치의 크기는 시소러스의 범위를 벗어나지 않도록 수집하였다. 유사단어를 확장하다보면 말뭉치의 크기가 커져 다른 말뭉치와 겹쳐질 수 있기 때문이다. 말뭉치 구축은 원활한 문장구성을 위해서이기 때문에 시소러스 범위를 벗어난 말뭉치는 유용하지 않다. 시소러스에 의한 말뭉치 사전의 구축은 어휘발췌에서 얻지 못했던 어휘들을 얻을 수 있는 장점이 있다. 그리고 시소러스의 계층도 어휘발췌와 분류과정을 통하여 새로운 시소러스를 만들어 나갈 수 있다. 현재 교통도메인에서 장소, 수단 등으로 나누었지만, 발췌된 문장의 분류를 통해 새로운 시소러스계층을 얻을 수 있다. 시소러스와 말뭉치에 의한 어휘발췌는 많은 문장을 만들 수 있는 방법을 제공하며, 후에 통신을 통한 어휘획득이 이루어질 때도 중요한 부분을 차지할 것이다. 새로운 어휘가 입력되면 시소러스에 의해 어휘의 위치가 지정되므로, 원활한 업 데이트가 가능하다. 표4는 시소러스에 의한 말뭉치의 구성을 보여준다.

<표 4> 말뭉치의 예 (교통도메인)

번호 1. 장소
1.1 지명 : 인천, 서울,
1.2 역명 : 제물포, 주안, 서울,...
1.3 위치 : 이곳, 저곳, 그곳, 여기, 저기,
번호 2. 수단
2.1 교통수단 : 지하철, 전철, 버스, 택시. 기차..

시소러스와 말뭉치로 발췌된 어휘 중 동사는 기본형, 의문문, 청유문의 동사활용으로 정리한다. 동사활용으로 정리하는 것은 사용자가 질문과 답변, 청유, 부정으로 문장을 바꾸어 사용할 수 있다. 시소러스와 동사활용의 지식을 기초로 하여 동사의 하위범주화 사전을 구축한다. 하위범주화사전에는 동사의 기본형과 동사활용 그리고 연결 가능한 명사의 시소러스 계층명과 조사 그리고 동사의 시소러스 계층명을 포함하는 구축한다. 조사도 각 명사에 따라 자세히 분석한다. 동사활용은 평서:답변, 의문, 요청, 허용, 부정으로 편의상 이름을 나누었다. 표 5는 문

장구성을 위한 하위범주화 사전 구축의 예를 보여준다.

<표 5> 하위범주화 사전의 예 (교통도메인)

[문형번호1]
[기본형] (시간이) 걸리다.
[시소러스계층명] 시간
[명사시소러스계층명:명사종류] 위치/수단
[조사] 까지, 으로/로
[평서:답변] 걸립니다.
[요청] 걸리게 해주세요.
[의문] 얼마나 걸립니까?
[요청:허가] 걸려도 됩니다.
[평서부정] 걸리지 않습니다.
[요청부정] 걸리지 않게 해 주세요.
[의문부정] 걸리지 않습니까?
[요청부정] 걸리면 안 됩니다.

하위범주화 사전에 따라 어휘의 연결이 이루어진다. 사용자가 명사를 선택하면 그 해당명사의 하위시소러스명(명사종류)이 검색되고, 그에 맞는 조사와 사용자에 의해 선택된 동사가 연결되어 문장이 만들어 진다. 명사는 하나뿐만 아니라 여러개의 선택도 가능하며, 해당 조사가 하위범주화사전에 따라 검색되어 붙여지게 된다. 은/는, 을/를 등의 조사는 명사의 종성을 검색하여 알맞게 붙여지도록 하였다. 하위범주화사전에서 명사종류가 여럿 나타나는 이유는 여러 도메인에서 중복되어 사용되기 때문이다. 조사는 한국어 초급 교재의 기초 어휘 선정에 관한 논문을 참고하여 "까지, 께, 께서, 나, 도, 만부터, 서, 에, 에게, 에서, 와/과, (으)로, 은/는, 을/를, 의, 이/가"를 기본으로 하였고, 그 외 어휘발췌에서 발견된 조사를 더하여 28부분으로 정리하였다. 현재 본 연구의 도메인은 교통, 식당, 쇼핑, 병원, 집, 중심어휘(일상),특성어휘로 나누었다. 발췌된 어휘는 여러 도메인에 중복되어 사용된다. 표 6은 명사이름, 종류, 소도메인명을 포함한 명사데이터베이스이다.

<표 6> 명사 데이터베이스의 예

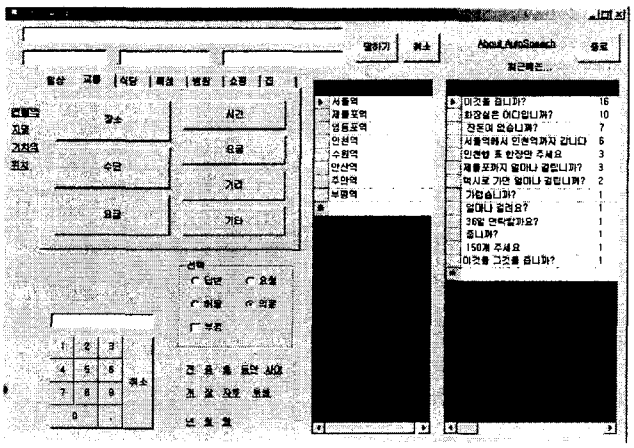
[명사번호 1]
[명사] 밝은색
[명사종류] 물건
[소도메인명] 색깔

3. 재사용과 에러 처리

본 연구에서는 출력된 문장을 저장하여 재사용 하고, 의미적으로 어색한 문장은 저장되지 않도록 구성하였다. 사용된 문장의 재사용은 필요한 어휘를 다시 선택하여 문장을 만들지 않아도 되므로 편리하다. 문장의 재사용방법은 사용자가 한번 사용한 문장을 저장하는데, 동사부분에 카운터를 두어 자주 사용되는 문장은 저장 데이터베이스에서 윗부분을 차지하도록 구성하였다. 그러므로 사용된 문장은 저장할 때 데이터베이스를 검색하여 일치하는 동사가 없는 경우 새로 저장되고, 사용된 동사가 있는 경우에는 카운터를 하나씩 늘려주도록 하였다. 그리고 저장된 문장에서 명사를 바꾸어서도 재사용이 가능하도록 하였다. 문장구성에서 에러처리는 의미적으로 어색한 문장을 저장하지 않는 방향으로 진행하였다. 어색한 문장의 제거는 사용자가 선택한 어휘에 대해 출력을 한번 해주는 대신, 출력된 문장 중 적절하지 않은 것은 저장하지 않아, 재사용이 되지 하도록 하였다. 의미적으로 적절하지 않은 문장의 검색은 하위시소러스 사전에서 이미 구축된 정보를 이용한다. 동사에 연결될 수 있는 명사종류를 정해놓았기 때문에 사용자가 명사, 동사를 선택했을 때 하위시소러스 사전을 검색하여 의미적으로 어색한 문장을 판별하여 저장 여부를 결정한다.

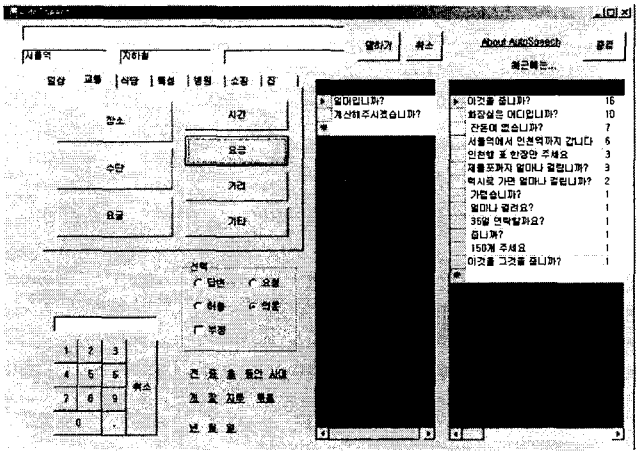
III. 문장구성 구현결과

구축된 어휘정보를 이용하여 문장구성을 한다. 문장구성의 구현은 통신보조기기의 일반적인 형태를 취하였고, 하드웨어 부분은 제외하였다. 전체 구성 요소로는 선택된 어휘를 보여주는 표시부분과 문장의 삭제와 음성출력을 위한 제어부분 그리고 도메인별 어휘부분으로 나누었다. 도메인별 어휘는 시소러스 계층이름으로 나뉘어져 배열하였다. 동사활용은 선택 부분의 버튼을 이용한다. 표시판은 사용자가 선택된 어휘가 상하의 표시부분 중 아래에 배치되고, 말하기를 눌렀을 때 조사가 연결된 완전한 문장이 위쪽에 나타난다. 그림 1은 도메인을 선택하고, 명사 선택을 위한 과정을 보여준다. 장소 시소러스 명을 선택하면 하부 시소러스가 보이고, 전철역을 선택하면 해당되는 명사들이 보인다.

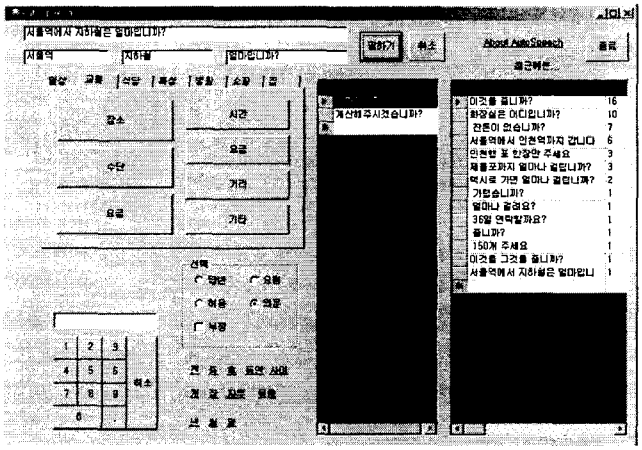


(그림 1) 문장구성 구현 (명사선택)

그림 2는 두개의 명사를 선택한 후 마지막으로 동사선택을 하기위한 그림이며, 그림 3은 선택된 어휘에 말하기 버튼을 이용하여 문장을 만든 것이다.

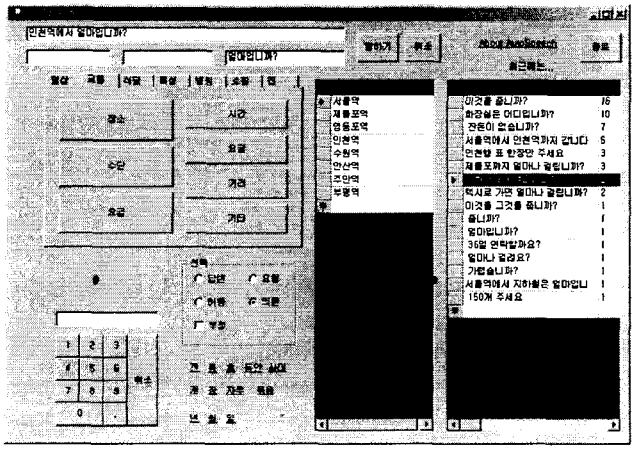


(그림 2) 문장구성 구현 (동사선택)

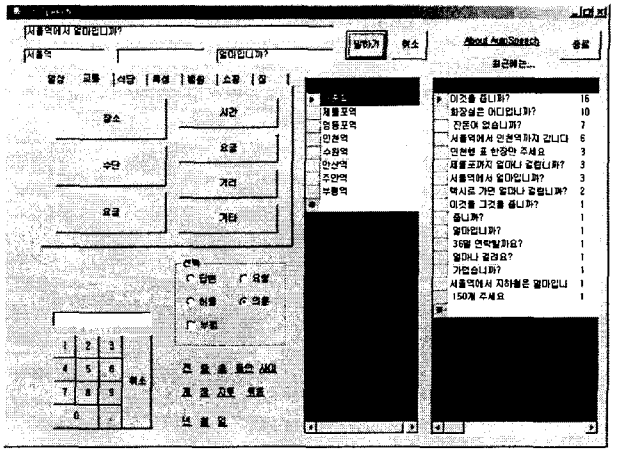


(그림 3) 문장구성 구현

말하기 버튼을 눌렀을 경우에는 문장으로 출력된 후 저장되는데, 최근목록에 해당 동사가 없다면 새로 저장되고, 해당 동사가 저장 데이터베이스에 있다면 카운터가 하나 올라간다. 음성 출력 부분은 본 연구에서 제외하였다. 그림 4는 저장된 목록 중 명사를 바꾸어 출력하기 위한 그림이며, 그림 5는 명사를 바꾸어 바뀐 문장이 출력된 그림이다.



(그림 4.) 저장목록 사용의 예 (명사바꿈)



(그림 5) 저장목록사용의 예 (새로운 문장출력)

Ⅳ. 결 론

어휘정보를 이용한 문장구성은 통신보조기기에 사용하기 위해 사용자 인터페이스를 고려하여 구현되어야 한다. 문장구성을 위한 방법으로는 도메인에 따른 어휘발췌, 문장패턴 분류, 시소러스와 말뭉치, 하위시소러스 사전의 구축이었다. 기본 자료를 이용하여 문장구성과 재사용, 에러처리를 하였다.

향후과제로는 제한된 어휘의 한계를 벗어나기 위한 방안으로 웹을 통한 어휘의 다운로드 환경과 문장구성에서 발생할 수 있는 문맥상 어색한 문장을 올바른 문장으로 재구성할 수 있는 에러처리용 데이터베이스의 구축이 필요하다.

* 본 연구는 인천대학교 멀티미디어 연구센터의 일부 지원에 의하여 수행되었음.

참고문헌

[1] S. L. Glennen and Decoste, The Handbook of Augmentative and Alternative Communication, Singular Publishing Group, Chapter 3, 1996

[2]수화사랑"http://myhome.naver.com/minirose/"

[3] 조성문, "한국어 초급 교재의 기초 어휘 선정에 관하여",한양어문, Vol.15, No.0, pp.317-348, 한국언어문화학회,1997

[4] 추교남, 개념 기반 정보 검색을 위한 한국어 어휘의 의미분석, 인천대학교 석사학위논문, 1998.12,

[5] 김영택외 10인, '자연 언어 처리', 교학사, 제7장, 1994