통신보조기기용 어휘 예측 시스템의 구조

황인정, 김효진, 이은주, 민홍기
인천대학교 정보통신공학과

Design of word prediction system for Assistive Communication System

Ein Jeong Hwang, Hyo Jin Kim, Eun Joo Lee, Hong Ki Min
Information and Telecommunication engineering, University of Inchon
terry@hci.inchon.ac.kr

Abstract

본 연구에서는 청각장애인용 통신보조기기에 적용하기 위한 어휘예측 시스템의 기본구조를 제안한다. 통신보조기기의 어휘는 사용자의 환경을 고려한 어휘이므로, 어휘 예측 시스템도 사용자의 환경과 실생활에서 쉽게 이용할 수 있는 방향으로 고안되어야 한다. 따라서 어휘예측 시스템은 사용자의 환경을 정의하고, 중심어휘와 장소별 도메인에서의 어휘를 발췌한다. 발췌된 어휘는 말뭉치와 의미함축의 원리를 이용하여 분류한다. 분류된 어휘는 문법적 지식을 바탕으로 가상 네트워크를 구성한다. 가상네트워크에서의 어휘는 명사, 조사, 동사의 3부분으로 나눈 후 의미함축과 말뭉치로부터 파생된 어휘를 근접한 거리에 위치시킨다. 동일한 네트워크상에서 어휘의 위치는 문법적 연관성, 빈도수 등을 이용하여 정한다. 따라서 본 연구에서는 어휘예측은 명사, 조사, 동사에서 가장 근접한 어휘를 연결하여 간단한 문장을 작성할 수 있는 어휘 예측 시스템의 기본구조를 제안한다.

Ⅰ. 서 론

본 연구에서는 통신보조기기내에 사용할 수 있는 실용어휘를 이용하여 어휘예측의 기본구조를 제안한다. 통신보조기기란 일상생활에서 음성을 통한 의사소통이 불편한 사람들을 위한 것으로써 사용자의 의도에 맞는 문장을 구성하여 음성으로 들려주는 장치이다. 그러므로 그 구조는 휴대하기 편리하고, 일상생활에서 무리없이 사용할 수 있어야 하며, 실용어휘가 적절히 구성되어 있어야 한다. 구성된 어휘는 쉽고 빠르게 사용자의 의도에 맞는 문장을 만들어야 한다. 통신보조기기는 사용자의 환경에 따른 장치이므로, 가장 먼저 사용자의 연령과 환경에 따라 어휘를 발췌하여야 한다. 발췌된 어휘는 통신보조기기내에서 사용자의 의도에 맞는 문장을 쉽고 빠르게 만들기 위하여 어휘예측이 필요하다. 어휘예측은 한 두개의 어휘를 이용하여 완전한 문장을 만들기 위한 방안이다.

본 연구에서는 어휘를 발췌하고, 분류한다. 분류된 어휘를 이용하여 신뢰성 있는 어휘예측이 되도록 가상네트워크를 이용한 방법론을 제시한다.

가상네트워크를 이용한 방법론은 문법적 지식을 기반으로 하여 통계, 말뭉치, 의미함축등의 기법을 이용하여 구성한다.

Ⅱ. 어휘 발췌와 분류

어휘의 발췌 시에는 사용자의 환경을 고려해야 한다. 사용자의 환경으로는 연령, 어휘능력, 자주 이용하는 장소 등이 있다. 어휘의 발췌에 사용자가 자주 언급하는 어휘만을 사용한다면, 어휘능력의 향상을 이룰 수 없고, 사회적으로 널리 통용되는 어휘습득의 기회를 차단할 수 있으므로, 이러한 점을 고려하여 세심한 어휘발췌가 되어야 한다.

본 연구에서의 사용자는 어린이로 정의하였다. 그러므로 어휘발췌는 초등학교 교과서를 중심으로 발췌하였다. 어휘의 발췌는 중심어휘(core vocabulary) 와 장소별 도메인에 따른 어휘(fringe vocabulary)로 나누어 발췌한다. 중심어휘란 장소에 구애받지 않고, 널리 사용되는 어휘를 말한다. 표 1은 중심어휘의 예이다. 장소별 도메인에서의 어휘는 특정 장소에서 자주 언급되는 어휘를 말한다. 장소별 도메인은 학교, 가정(집), 식당 등으로 나눌 수 있다. 발췌된 어휘는 표 2에서와 같이 명사, 동사, 조사 등으로 나눈다.

표 1. 중심어휘의 예

도와주십시오. 감사합니다. 죄송합니다. 저는 홍길동입니다.

표 2. 가정(집) 도메인에서의 어휘발췌의 예

명사	조사	동사
학교,시장,학원, 친구집,놀이터	에	다녀왔습니다. 다녀오겠습니다. 다녀올까요?
문,창문	을	열겠습니다. 열어주세요. 열려있습니다.열었습니다.
방,집,나의방,거실, 욕실,부엌	이	춥습니다. 덥습니다.
안녕히 주무세요. 일찍 깨워주십시오. 엄마, 아빠, 형, 누나		

말뭉치는 분류된 어휘들로 구성된다. corpus라 불리우는 말뭉치는 언어 현실을 총체적으로 보여줄 수 있는 언어 자료의 집합을 말한다. 그러므로 언어 자료들이 조직적인 정보의 형태로 모여 있는 것은 모두 말뭉치라고 할 수 있다. 좁은 의미로 본다면 컴퓨터처리가 가능한 컴퓨터가 읽을 수 있는 형태로 저장된 일정 규모 이상의 언어 자료를 가리키게 된다.

어휘예측에서 필요한 말뭉치는 현실생활에 존재하는 어휘와 그 어휘의 의미를 확장하여 나타내는 어휘사전으로 만들 수 있다. 기존의 인덱스 위주의 사전과는 다른 형태로 묶여있는 것을 말한다. 그림 1은 어휘예측을 위하여 계층적으로 분류한 말뭉치의 예이다.

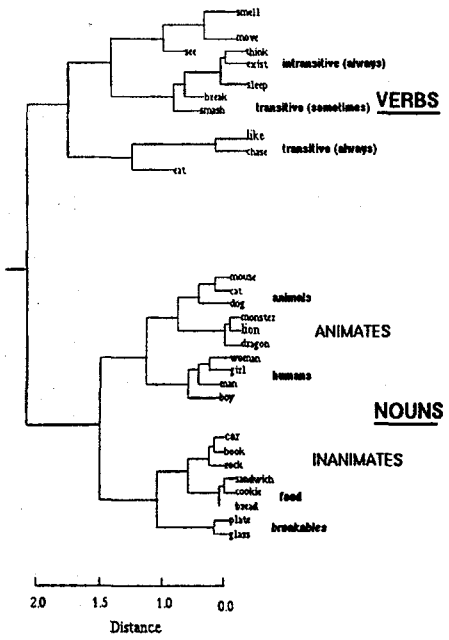

그림 1. 말뭉치(corpus)의 예

의미함축(semantic compantion)은 비슷한 의미의 어휘들을 확장하여 나타내는 개념이다. 본 연구에서는 말뭉치와 의미함축의 원리를 이용하여 발췌된 어휘를 분류하고 확장하였다. 명사는 단수와 복수, 비슷한 개념의 단어, 그리고 분류된 명사와 연결되어 자주 인용되는 어휘도 한 부류로 묶는다. 동사는 의문문, 부정, 청유, 현재, 과거, 미래등으로 묶는다. 표 3에서는 말뭉치와 의미함축을 이용한 어휘분류를 예를 보여주었다.

표 3. 말뭉치와 의미함축을 이용한 어휘분류의 예

명사	옷	옷옷	바지	입습니다.	
	신발	운동화	구두	신습니다.	
	학교	학원		공부합니다	
동사	있다.	있습니다.	있습니까?	있지않습니다.	있을까요?
	쉽다	쉽습니다.	쉽습니까?	쉽지않습니다.	쉬웠습니다.
	켜다	켜주세요	켰습니까?	켜지말아주세요.	켰습니다.

III. 어휘예측의 방법

어휘예측은 어휘발췌와 말뭉치, 의미함축등으로 확장된 어휘를 이용한다. 통신보조기기내에 사용될 수 있는 어휘예측시스템 개발의 전체적인 진행방향은 그림 2와 같다. 사용자의 범위를 확정하고, 그 사용자에 맞는 어휘를 발췌하고 분류하여 데이터베이스를 만든다. 분류된 어휘를 이용하여 가상네트워크를 구성한다. 문법적 지식과 말뭉치, 의미함축등의 이론을 이용하여 가상네트워크상에서의 어휘의 위치를 정한다. 지속적인 학습으로 수정과 보안을 거친 후 의미심볼과 연결하여 어휘예측 시스템을 완성한다.

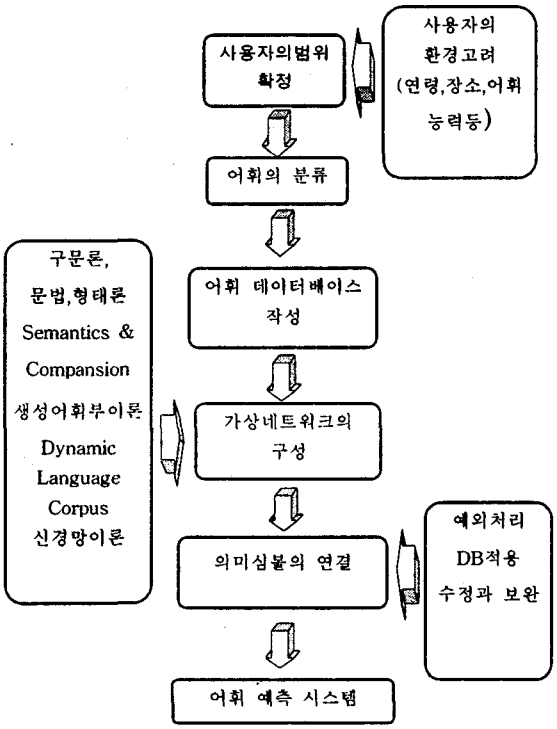

그림 2. 어휘예측시스템의 진행방향

IV. 가상 네트워크의 구성

어휘예측의 방법은 문법적 지식을 기반으로 통계, 말뭉치, 의미함축등의 원리를 이용하여 가상네트워크 구성을 제안한다. 가상 네트워크는 명사, 조사, 동사부분으로 나누어 한 문장을 완성할 때 가장 근접한 어휘로 연결되도록 한다. 사용자가 원하지 않는 어휘가 예측되었을때는 지속적인 학습에 따라 보안되며, 근접한 위치에 따라 순위를 정하여 사용자에게 보여줄 수 있을 것이다. 그림 3은 제안한 가상 네트워크의 전체적인 보기이다.

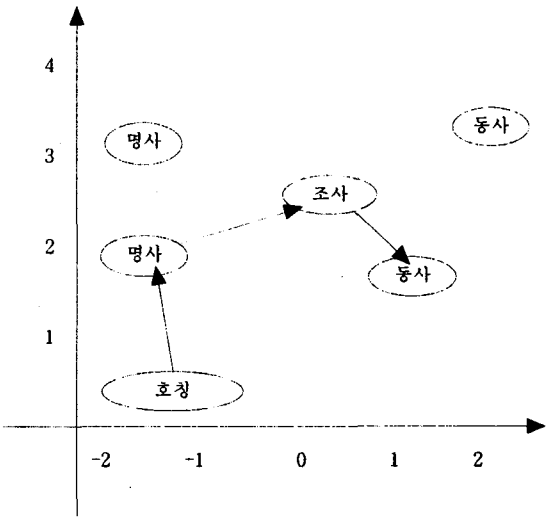

그림 3. 제안한 가상네트워크

그림 3의 제안된 가상네트워크에서 가로축과 세로축은 연관도를 나타낸다. 숫자가 적을수록 연관도가 높은 것이다. 그림 4는 어휘를 이용하여 제안한 가상 네트워크를 구성해 보았다.

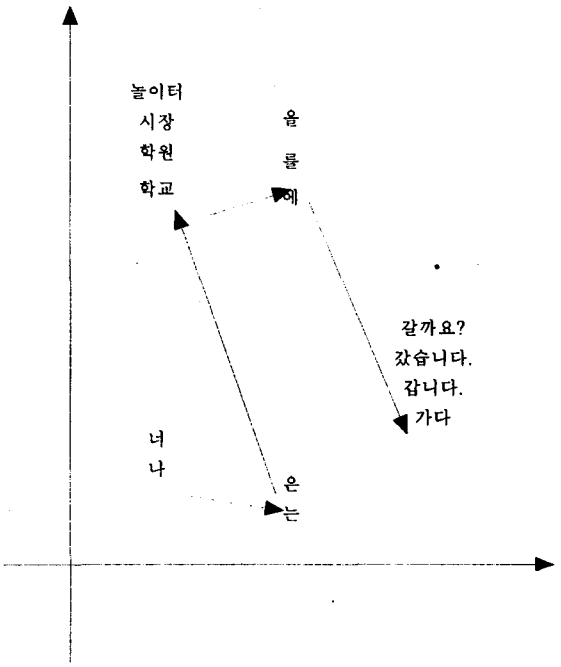

그림 4. 제안한 어휘를 이용한 가상네트워크

실용어휘에서 간단한 문장은 명사+조사+동사로 이루어져 있다. 장소 도메인에서의 어휘는 제한적이고, 자주 반복되어 나타나며, 일정한 패턴을 가진다. 가상 네트워크안에서의 어휘들은 문장패턴에 따라 자주 반복되어 나타나는 것을 연관성이 큰 것이라고 가정한다. 그러면 패턴에 따라 문장을 이루는 어휘들은 근접한 위치에 나타난다. 이러한 연관성에 착안하여 반복 수행한다면, 한 개의 어휘선택만으로도 연관도에 따라 완전한 문장을 제시해 줄 수 있을 것이다. 가상 네트워크에서 어휘의 위치는 문법적인 지식과 빈도수와 패턴등의 요소를 이용하여 정할 수 있을것이다. 문법지식은 어느 정도의 패턴을 만들 수 있으며, 의미함축과 빈도수에 따라 연관성의 위치를 정할 수 있을 것이다.

V. 결 론

본 연구는 통신보조기기내에 사용할 수 있는 어휘예측 시스템을 구성하기 위한 방법론을 제시하였다. 어휘를 장소별로 발췌하고, 의미함축과 말뭉치의 원리를 이용하여 분류한다. 분류된 어휘는 문법적 지식과 통계등을 이용하여 가상 네트워크를 구성하며, 가상네트워크 상에서 근접한 위치에 있는 어휘를 연결하여 간단한 문장을 만들 수 있었다. 어휘를 쉽게 인식하고, 효과적으로 표현하기 위해서는 어휘의 의미를 적절히 표현할 수 있는 의미심볼로 연결하여 사용하여야 한다.

어휘예측 시스템은 한 두 가지의 방법으로 완성될 수 있는 것이 아니므로, 많은 시행착오와 지속적인 학습, 예외처리 데이터베이스등이 계속 보완되어야 할 것이다. 이러한 점이 보완된다면 좀 더 복잡한 문장을 예측할 수 있을것이라 생각된다.

본 연구는 인천대학교 멀티미디어 연구센터의 일부지원에 의하여 수행되었음.

참고문헌

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

[2] 홍재성 외, 현대 한국어 동사구문 사전, 두산동아, 1997

[3] R. T. Corss and L. S. Valot " Using core Vacabulary in Activity-Based Learning", Proceedings of the 17th Annual Southeast Augmentative Communication Conference, Birmingham : SEAC Publications, pp.25-30, 1996

[4] 이정민,"한국어 술어 중심의 의미구조", 한국인지과학회 춘계학술발표논문집, pp.32-40, 1997

[5] K. F. McCoy, "Simple NLP Techiques for Expanding Telegraphic Sentences", In Proceedings of Natural Language Processing for Communication Aids, an ACL/EACL '97 Workshop, Madrid, Spain, July 12, 1997

[6] K. F. McCoy and P. Demasco. "Some Applications of Natural Language Processing to the Field of Augmentative and Alternative Communication", In Proceedings of the IJCAI-95 Workshop on Developing AI Applications for People with Disabilities, Montreal, Canada, August, 1995.

[7] W. M. Zickus, K. F. McCoy, P. W. Demasco, and C. A. Pennington, "A Lexical Database for Intelligent AAC Systems", In Proceedings of RESNA '95 18th Annual Conference, Vancouver, B.C., June 1995.

[8] J. L. Elman, "Language as a dynamical system", Mind as Motion: Explorations in the Dynamics of Cognition. Cambridge, MA: MIT Press. pp. 195-223, 1995