한국어 어휘의미망을 이용한 자동 수화 번역 시스템의 개발1)

김민호*, 최성기*, 권혁철*
*부산대학교 컴퓨터공학과
e-mail:karma@pusan.ac.kr

Development of Automatic Sign Language Translation System using Korean WordNet

Minho Kim*, Sung-Ki Choi*, Hyuk-Chul Kwon*
*Dept of Computer Science & Engineering, Pusan National University

요 약

한국어와 한국 수화 간 자동 번역을 위해서는 한국어-한국 수화 대역어 사전이 필요하지만, 현재 한국 수화 사전으로 가장 공신력 있는 한국 수화 사전은 등재 어휘 수가 약 12,000개에 불과하다. 이 때문에 한국어를 한국 수화로 자동 번역을 할 때 대치어가 없어 완벽하게 번역이 되지 않는다. 본 연구에서는 한국 수화 사전의 미등재어로 말미암은 번역률 저하를 최소화하고자 한국어 어휘의미망의 동의어와 상·하위어 정보를 이용한다. 또한, 자동 번역에서 빈번하게 발생하는 어의 중의성 문제도 한국어 어휘의미망의 정보를 이용하여 어의 중의성 해소 규칙을 일반화한다.

1. 서론

수화는 주로 손의 움직임과 표정 등과 같은 비수지 신호로 뜻을 전달하는 청각장애인에 의해 창조된 언어로서, 고유한 문법과 표현방법을 가진 하나의 독립된 언어이다[1, 2]. 청각장애인에게 한국어는 영어나 일본어와 같이 다른 문법 체계를 가진 언어이기 때문에 모든 청각장애인이 한국어로 된 문장을 읽거나 쓸 수 없다. 따라서 청각장애인과 비청각장애인이 서로 대화를 하거나 청각장애인에게 정보를 전달하려면 문자가 아닌 수화를 사용해야 한다.

청각장애인이 한국어를 배우거나 혹은 비청각장애인이 수화를 배우는 일은 많은 노력이 있어야 한다. 한국어와 한국 수화 간 자동 번역 시스템은 이러한 노력을 줄일 수 있으므로 많은 연구기관에서 자동 번역 시스템을 만들기 위한 연구를 진행하였다.

한국어와 한국 수화 간 자동 번역을 위해서는 한국어-한국 수화 대역어 사전이 필요하지만, 현재 한국 수화 사전으로 가장 공신력 있는 한국 수화 사전은 등재 어휘 수가 약 12,000개에 불과하다[3]. 이 때문에 한국어를 한국 수화로 자동 번역을 할 때 대치어가 없어 완벽하게 번역이 되지 않는다. 또한, 자동 번역에서 빈번하게 발생하는 어의 중의성 문제도 자동 번역의 번역률을 낮추는 가장 큰 이유 중 하나이다.

본 연구에서는 한국 수화 사전의 미등재어와 어의 중의성 문제로 말미암은 번역률 저하를 최소화하고자 한국어 어휘의미망의 동의어와 상·하위어 정보(이하 관계어 정보)를 이용한다. 본 논문의 2장에서는 국내·외 관련 연구와 본 연구에서 활용하는 한국어 어휘의미망을 설명한다. 3장에서는 본 연구에서 제안하는 한국어-한국 수화 자동 번역 시스템의 전체적인 구조와 한국어 어휘의미망의 활용 방안에 대해 설명하고 4장에서 한국어 어휘의미망의 활요 효과를 실험 결과를 통해 분석한다. 마지막 5장에서는 결론과 앞으로의 연구에 관하여 제시한다.

2. 관련 연구

2.1 국내·외 수화 자동 번역 연구

수화에 관한 연구는 수화가 청각 장애인의 언어라는 특수성과 낮은 경제성 때문에 활발히 진행되지 못하고, 산발적으로 이루어졌다. 국내에서는 1990년대 말부터 2000년대 초 사이에 몇몇 기관에서 관련 연구가 진행되었다. 그러나 한국어 언어분석기술의 부재로 국문법 체계의 문장을 수화 문법 체계로 변환하지 않고, 단순히 단어별로 대치하여 수화를 생성하여 수화가 자연스럽지 못하였으며 이마저도 한국 수화 사전의 어휘 수 부족으로 완벽한 수화 문장의 생성이 불가능하였다.

기존 한국어-한국 수화 자동 번역 연구에서 기술적 성과가 크지 못했던 이유는 한국어 통사 분석의 어려움과 한국어 어휘 의미 중의성 때문이다. 한국어는 영어와 달리 어순이 비교적 자유롭고, 문법적 관계에 있는 구성요소들

1) 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2012R1A2A2A06046730).

이 불연속적으로 나타날 수 있어 구문 분석의 난도가 높다. 또한, 하나의 단어가 문장에서의 쓰임에 따라 각각 다른 의미로 사용되는 어의 중의성 문제는 수화 번역을 더욱 어렵게 만든다.

국외는 국내와 비교하여 비교적 활발히 연구가 진행됨. 수화를 자연어로 자동으로 변환하는 연구와 자연어를 수화로 변환하는 연구 모두 진행되고 있다. 자연어 간 기계번역 기술을 토대로 자연어를 수화로 변환하는 연구가 조금 더 활발히 진행되고 있으며, 특히 통계에 기반을 둔 방법을 중심으로 연구가 진행되고 있다. 그러나 대규모 수화 말뭉치의 구축이 어려워 100~500개의 수화 단어를 중심으로 자연어-수화 번역이 이루어지고 있다.

본 연구에서는 한국 수화 사전의 미등재어로 말미암은 번역률을 최소화하고자 한국어 어휘의미망의 관계어 정보를 이용하여 한국어-한국 수화 정보를 확장한다. 또한, 한국어 어휘의미망의 정보에 기반을 둔 어의 중의성 해소 규칙을 구축하여 자동 번역에서 일어나는 어의 중의성 문제를 해결한다.

2.2. 한국어 어휘의미망

한국어 어휘의미망(Korean Lexico-semantic Network; 이하 KorLex)는 영어 워드넷(이하 PWN)을 참조모델로 하여 확장 개발된 대규모 언어자료이다. PWN은 어휘의미(word sense)를 기준으로 (그림 1)]의 {실과1, 과일1, 과실2}처럼 동의어를 묶은 신셋(synonym set; 동의어 집합)을 기본 단위로 삼아, 내용어(명사, 동사, 형용사, 부사)를 대상으로 신셋 간 계층구조(hierarchical structure)를 설정하였다.

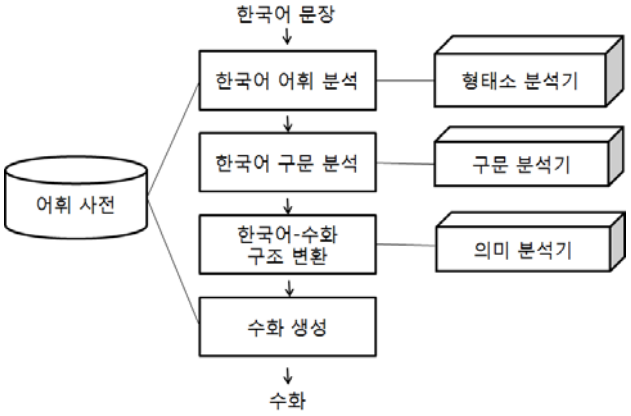
(그림 1) 어휘의미망의 계층 구조

예를 들어, 명사는 25개의 분야(음식, 동물, 식물, 인지, 상태, 행위, 사물, 인간, 기상 현상, 속성, 물질 등)의 약 8만 개의 신셋이 (그림 1)과 같은 계층구조에서 동의어, 상위어, 하위어, 형제어 관계를 형성하고 있다. KorLex는 PWN의 의미관계를 기본 골격으로 삼되, 영어 어휘의미에 경도된 일부 의미관계를 한국어 어휘의미에 맞게 수정 보완하고 확장한 대규모 어휘의미망이며, 그 크기는 <표 1>과 같이 중형 사전에 해당한다. 이는 PWN과 유사한 크기이며, 한국어에 발달한 "명, 분, 마리, 권" 등과 같은 분류사를 추가로 구축하고 명사와의 공기관계(cooccurrence) 등을 설정하였다.

3. 한국어-한국 수화 자동 번역 시스템

3.1 한국어-한국 수화 자동 번역 시스템의 구조

전통적인 기계 번역 시스템의 일반적인 구조에 따라 본 연구에서 개발한 '한국어-한국 수화 자동 번역 시스템'의 구조는 (그림 2)와 같다.



(그림 2) 한국어-수화 자동 번역 시스템의 구조

자동 번역을 위해서는 다양한 유형과 단계의 사전 정보, 규칙 정보 등이 필요하다. 이를 이용하여 한국어 문장을 분석하고, 수화를 생성하기 위한 부분들로 구성된다. (1) 한국어 어휘 분석은 입력된 문자열을 분석하여 형태소라는 자연언어 분석을 위한 기본단위로 분류하는 것이다. 이를 위해 형태소 분석기는 어휘 사전을 바탕으로 입력 문자에 형태소 결합 규칙을 역으로 적용하여 형태소를 분석하고, 각 형태소가 가진 품사 정보를 어휘 사전으로부터 추출하여 함께 출력한다. (2) 한국어 구문 분석은 통사 규칙에 따라 문장 내에서 각 형태소가 가지는 역할, 혹은 상호 관계를 분석하는 것이다. 여기서는 어휘 분석의 결과를 입력으로 받아서 입력 문장에 대응하는 파스 트리(parsed tree)를 출력한다. (3) 변환 단계는 구조적 모호성과 의미 모호성을 해결한다. 즉, 한국어 문장과 수화 사이에 존재하는 구조적인 차이를 해소하고 분석된 단위에서 사용된 어휘의 적절한 의미를 선택한다. 마지막으로 (4) 수화 생성 과정은 변환 단계를 거친 후의 한국어 내부 표현으로부터 수화를 생성해 내는 과정이다.

3.2 한국어-한국 수화 대역어 정보 확장

(그림 2)에서 수화 생성을 위해서는 한국어-한국 수화

대역어 정보가 필요하다. 기존 연구에서는 대역어 정보로서 한국 수화 사전의 정보를 이용하지만, 한국 수화 사전에 등재된 어휘 수가 부족하여 미등재어는 지화[2]로 표시하였다. 본 연구에서는 KorLex의 관계어 정보를 이용하여 미등재어를 처리한다.

(a) *눈이 흩날리다.*

문장 (a)를 수화로 번역하려면 한국어 '*눈*'과 '*흩날리다*'에 대응하는 한국 수화가 있어야 한다. 그러나, 한국 수화 사전에 '*흩날리다*'라는 단어가 없으므로 기존의 연구에서는 '*흩날리다*'를 지화 '*ㅎ,ㅡ,ㅌ,ㄴ,ㅏ,ㄹ,ㄹ,ㅣ,ㄷ,ㅏ*'로 번역한다. 이는 수화문의 길이를 지나치게 길게 만들어 의사 전달을 어렵게 만든다. 본 연구에서는 이를 해결하고자 미등재어의 동의어, 하위어, 상위어 순으로 검색하여 해당하는 단어가 존재하면 그 단어로 대치한다. KorLex에서 '*흩날리다*'는 '*날리다*'의 하위어이므로 문장 (a)는 문장 (b)로 변환된다.

(b) 눈이 날리다.

'*흩날리다*'가 '*날리다*' 보다 더 세분화된 의미이긴 하지만 '*흩날리다*'를 지화로 표현하기보다는 비슷한 의미인 '*날리다*'의 수화로 표현하는 것이 의미 전달에서 더 효과적이다.

3.3 어의 중의성 해소

자연언어처리에서 어의 중의성 해소(word sense disambiguation; WSD)란, 하나 이상의 의미가 있는 어휘(이하 중의성 어휘)가 문맥에서 어떤 의미로 사용되었는지를 판단하여 정확하게 구분하는 작업이다. 어의 중의성 해소는 형태소 분석과 통사 분석과 마찬가지로 자연언어처리에서는 반드시 필요한 작업으로 여러 응용분야에서 중요한 역할을 담당한다. 기계번역에서 어의 중의성 해소는 주어진 어휘의 올바른 대역어(translation)를 선택하는 데 있어서 매우 중요하다. 예를 들어, 문장 (b)의 한국어 '*눈*'은 '*eye*', '*snow*' 등으로 번역될 수 있는데 이들 중에서 문맥상 가장 올바른 의미를 선택하여야 한다.

기존 연구에서 어의 중의성 해소를 위해 가장 많이 사용되는 방법은 어의 중의성 해소 규칙을 이용하는 것이다. 규칙을 이용하는 방법은 통계를 이용하는 방법과 비교하여 재현율은 낮지만, 정확도가 매우 높고 쉽게 적용 가능하므로 상용화된 시스템에서 많이 이용한다. 어의 중의성 해소 규칙에서 가장 중요한 부분은 '문맥 정보'로서, 기존 연구에서는 이 '문맥 정보'를 단어로 표시한다.

<표 1>은 기존 연구에서 어의 중의성을 해소하기 위해 만든 규칙을 간략하게 표시한 것이다. 해당 규칙에서 보는 바와 같이 중의성 어휘 '*눈*'의 문맥에 '*내리다*', '*날리다*', '*흩날리다*'와 같은 단어(R1)가 나타나면 'snow'의 뜻으로 번역한다. 이러한 규칙은 문맥에 나타난 어휘 간의 정확한 일치를 전제로 하므로 규칙의 재현율(recall)이 낮다.

<표 1> 중의성 어휘 '*눈*'을 위한 규칙

눈 CW → snow	
(R1)	CW = [내리다 | 날리다 | 흩날리다]

본 연구에서는 문맥 정보를 확장하여 규칙의 재현율을 높이고자 문맥 정보로 단어 대신 KorLex의 신셋 정보를 활용한다. 즉, '내리다'라는 단어 대신 '내리다'에 해당하는 신셋 번호를 사용하고, 중의성 어휘의 주변에 나타난 어휘가 규칙의 문맥 정보에 잇는 신셋의 동의어, 하위어이면 해당 규칙이 적용되도록 한다.

<표 2> 중의성 어휘 '*눈*'을 위한 수정한 규칙

눈+P CW → snow	
(R1)	Context = B1
(R2)	CW = [02674938 | 02041026 | 02197925]
(R3)	P = [주격, 보격, 목적격, 보조사]
(R4)	Conjugation = 1001 + 2001

<표 2>는 본 연구에서 사용하는 중의성 해소 규칙의 간단한 예이다. 중의성 해소 규칙은 정규문법으로 표현되며, 사용될 때는 유한상태기계로 바뀌어 작동한다. 각 규칙이 작동하는 핵심 어휘(위의 예에서는 '눈')를 기준으로 해당 규칙을 찾게 되고, 문맥 정보를 통해 문맥에서 사용된 의미(위의 예에서는 'snow')로 교정한다. <표 2>의 교정규칙은 문맥 정보(CW)와 조사 제약(P)을 기술한다. 각 세부 규칙은 크게 ① 핵심어를 기준으로 해당 문맥의 위치(R1), ② 신셋 번호로 표시된 문맥 정보(R2), ③ 조사 제약 정보(R3), ④ 문맥의 활용형 정보(R4)에 관한 제약 등을 포함한다. 예를 들어, "*눈이 나풀대다*"라는 문장이 있을 때 중심어 "눈"의 문맥 첫 번째 요소(R1)에 "02041026"의 하위어인 '*나풀대다*'가 나타나고(R2), 중심어 "*눈*"은 주격조사와 결합(R3)하므로 해당 규칙이 동작한다.

4. 실험 결과 및 분석

KorLex를 활용한 한국어-한국 수화 번역률과 어의 중의성 해소율을 측정하고자, 'KBS 1TV 뉴스 930 일기예보' 스크립트 1년 치 데이터(2012)를 대상으로 실험하였다. <표 3>은 KorLex 사용 전후의 번역률을 비교한 것이다.

<표 3> KorLex 사용 전후의 번역률 비교

말뭉치 구분	말뭉치 크기	번역률	
		KorLex 적용 전	KorLex 적용 후

2) 수화(手話)에서, 한글 자모음이나 알파벳, 숫자 하나하나를 손가락으로 표시하는 방법.

내부 말뭉치	82,303어절	95.87%	96.17%
외부 말뭉치	1,448어절	88.60%	94.68%

내부 말뭉치는 수화 번역 시스템 개발 당시 분석을 위해 사용한 문장이며, 외부 말뭉치는 시스템 개발 후 순순한 번역률 분석을 위해 추가로 수집한 문장이다. 어의 중의성 해소 역시 빈번하게 나타나는 중의성 어휘가 포함된 문장 1,000개에 대한 어의 중의성 해소 정확도가 75%에서 86%로 많이 증가하였다.

5. 결론 및 앞으로 연구

본 연구에서는 한국 수화 사전의 미등재어와 어의 중의성 문제로 말미암은 번역률 저하를 최소화하고자 한국어 어휘의미망의 동의어와 상·하위어 정보를 이용하였다. 미등재어가 나타났을 때 미등재어의 동의어, 하위어, 상위어 순으로 KorLex에서 검색하여 해당하는 단어가 존재하면 그 단어로 대치하였다. 또한, 어의 중의성 해소 규칙을 만들 때 문맥 정보를 단어가 아닌 신셋 정보를 사용하여 문맥 정보를 확장하였다.

아직 연구 초기이기 때문에 '일기예보'라는 특정 도메인으로 대상으로 실험을 진행하였으나, 앞으로 연구에서는 좀 더 일반적인 도메인에서 연구를 진행하여 한국어-한국수화 간 번역률을 높일 예정이다.

참고문헌

[1] 이준우, 수화통역입문, 인간과 복지, 2004.03.

[2] 이준우, 김연식, 송재순, 한기열, 홍유미, 한국수화 회화 첫걸음, 나남출판사, 2010.03.

[3] 한국 수화 사전, http://222.122.196.111/, 국립국어원

[4] 김대진, 김정배, 장원, 변증남, "TV 자막 신호를 이용한 한글 수화 발생 시스템의 개발," 전자공학회논문지-CI 39(5), pp.32-44, 2002.09.

[5] 정상윤, 장은영, 박종철, "수화 자동 생성을 위한 한국어 동음이의어 분석과 처리," 한국정보과학회 학술발표논문집 36(1C), pp.315-320, 2009.06.

[6] 김준엽, 정진우, 박종철, "한국어-수화 자동 변환을 위한 수화 문장의 시제 표현 분석", 한국정보과학회 학술발표논문집 39(2B), pp.121-123, 2012.11.

[7] Jan Bungeroth, Hermann Ney, "Statistical Sign Language Translation" Workshop proceedings of Representation and Processing of Sign Languages, pp.105-108, 2004.05.

[8] M. Huenerfauth, "A Multi-Path Architecture for Machine Translation of English Text into American Sign Language Animation" Proceedings of Student Workshop at Human Language Technolgies conference 2004.03.

[9] R. San-Segundo, R. Barra, R. Córdoba, L.F. D'Haro, F. Fernández, J. Ferreiros, J.M. Lucas, J. Macías-Guarasa, J.M. Montero, J.M. Pardo (2008) "Speech to sign language translation system for Spanish Speech Commun. Vol.50, pp.1009-1020, 2008.01.

[10] H.-Y. Su, Y.-H. Chiu, C.-H. Wu, C.-J. Cheng "Joint optimization of word alignment and epenthesis generation for Chinese to Taiwanese sign synthesis", IEEE Transactions of Pattern Analaysis, Vol.29, pp.28-39, 2007.01.