한국어 동사의 격틀 정보를 이용한 구문분석 후처리기

전은희[0] 이성욱 서정연
서 강 대 학 교 컴 퓨 터 학 과
jeh92cool@hotmail.com, gospelo@nlprep.sogang.ac.kr, seojy@ccs.sogang.ac.kr

Post-processor of Parsing Results Using Case Frames

EunHee Jeon[0] Songwook Lee Jeonyun Seo
Dept. of Computer Science, Sogang University

요 약

언어를 컴퓨터로 처리하기 위한 방법으로 격문법(Case Grammar)을 사용하는 것이 있다. 격문법은 동사에 대한 격틀(Case Frame)을 기술함으로써 그 동사와 의미적으로 관계를 가지는 명사들에 대해 표현하는 것이다. 따라서 이러한 격 문법을 사용하기 위해서는 동사에 대한 격틀을 기술하는 것이 필수 과제이다. 본 연구에서는 동사에 대한 격틀을 기술하기 위해서 말뭉치에서 직접 사용된 명사-조사 쌍과 동사를 추출하여 이들의 격관계를 결정하고 이 자료들을 모두 동사의 격틀 정보로 사용하였다. 이렇게 구축된 격틀 자료를 구문분석의 후처리 단계에 적용하여 구문분석 결과 잘못된 명사-조사 쌍 의존관계를 수정하였다.

1. 소 개

격문법(Case Grammar)은 언어의 의미에 초점을 맞추어 언어의 문법성을 규명하고, 분석의 결과로 언어의 의미를 도출해 내는 문법이다. 이 방법은 동사구와 명사구 사이의 의미적 관계를 기술해 주기 때문에 비교적 형식이 자유로운 한국어를 처리하는데 유용한 방법이다. 격문법에서 문장의 주요 의미는 동사가 가지고 있고 그 외 다른 성분들은 동사에 대한 격으로 표현이 된다. 따라서 격문법을 사용하기 위해서는 동사별로 어떤 격을 갖게 되는지를 설명하는 격틀이 미리 정의되어 있어야 한다. 따라서 격틀을 이용하기 위해서는 한국어의 동사에 대한 격틀을 구축하는 것이 우선 과제이다.

따라서 격틀을 구축 방법에 관한 많은 연구들이 진행되어 왔다. 과기원 전문용어언어공학연구센터에서 구축하는 과기원 격틀[8]은 수동으로 직접 구축하는 격틀로 그 내용으로는 동사의 의미, 행위주체, 조사와 그에 따를 수 있는 개념, 하위 개념으로 실제 예를 기록하였으며, 격틀 자동구축과 격틀 평가방법에 관한 연구[7]에서는 기계 가독형 사전(Machine Readable Dictionary)과 말뭉치를 이용하여 자동으로 격틀을 구성하였다.

본 연구에서는 한국어의 동사에 대한 격틀 데이터베이스를 반자동으로 구축하고 이를 한국어 구문분석 후처리기에 응용함으로써 구문분석의 성능을 향상시켰다. 격틀 구축을 할 때 사용한 격 분류체계는 선행 연구[1]에서 정의한 30개의 격 분류체계를 사용하였다.

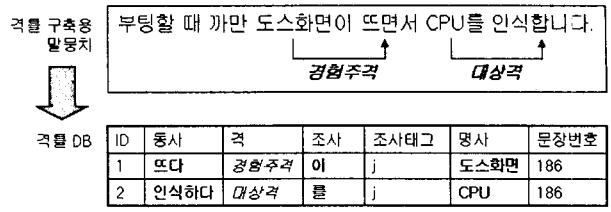
격틀 데이터베이스를 구축하는 방법은 사람이 직접 문장을 보고 격을 결정하는 방법을 사용하였으며, 이를 위해서 한국어 동사에 대한 격틀을 쉽게 구축하기 위한 도구를 제작하였다. 그 결과로써 나오는 격틀 정보와 시소러스를 이용하여 구문분석 후처리에 응용함으로써 구문분석의 성능 향상을 가져옴과 동시에 동사에 의존하는 명사에 대한 격을 결정할 수 있다.

2. 격틀 구축기

본 논문에서 사용한 격틀 자료를 구축하기 위해서 구문분석된 말뭉치를 사용하여 격틀을 구축하였다. 격틀 구축 도구를 사용하여 격틀을 구축하게 되는데, 격틀 구축 도구는 사용자가 쉽고 편리하게 격틀을 구축할 수 있도록 도와주는 도구이다. 격틀 구축은 문장 단위로 이루어지며, 격틀 구축 도구가 각 문장에서 출현한 동사와 그 동사에 의존하는 명사-조사 쌍을 추출하여 가능한 후보 격 리스트를 보여주면 사용자가 가장 적합한 격을 결정하는 과정이 반복된다.

그러나 격틀을 구축하기 전에 선행되어야 할 과제가 있는데, 바로 구문분석의 결과가 정확한 것인지 검사하는 작업이 필요하다. 격틀 구축시 출현한 동사들에 대해서 사용자에게 후보로 제시되는 명사들은 의존 관계를 이용하여 추출되기 때문에 의존관계가 올바로 결정 되어야만 올바른 명사들을 제시할 수 있고, 제대로 된 격틀을 구축할 수 있기 때문이다. 이렇게 격틀을 구축하기 전에 어절간의 의존관계를 검사하고 잘못된 의존관계를 수정할 수 있는 도구가 격틀 구축 도구에 포함된다. 따라서 사용자가 격틀을 구축하기 전에 반드시 의존관계 검사 도구를 사용하여 의존관계를 검사 하여야 하며, 격틀 구축은 의존관계가 끝난 문장에 대해서 수행된다.

[그림 1]은 이러한 방식으로 격틀을 구축하는 예를 보여준다.



ID	동사	격	조사	조사태그	명사	문장번호
1	뜨다	경험주격	이	j	도스화면	186
2	인식하다	대상격	를	j	CPU	186

[그림 1] 격틀 구축의 예

3. 구문분석 후처리기

구문분석 후처리 과정에서는 격틀 데이터를 사용하여 구문분석의 결과를 향상시키고 동시에 명사와 동사와의 격 관계를 결정하는 작업을 수행한다.

구문분석 후처리는 문장 단위로 수행되는데, 구문 분석 후처리는 문장 단위로 수행된다. 문장을 각 어절 단위로 읽어 들이면서 어절에 명사, 조사 쌍이 발견되면 이를 기억해 둔다. 그리고 어절에 동사가 발견 되면 기억해둔 명사, 조사 쌍들과 동사와의 격 확률을 계산한다. 격 확률을 계산할 때에는 동사에 대해 구축된 모든 격에 대해서 이 명사, 조사 쌍이 그 격으로 결정될 확률을 계산한다.

만약, 명사, 조사, 동사 쌍이 모두 일치하는 것이 격 DB에 있을 경우 격 확률을 1로 설정하고 그렇지 않을 경우 시소러스를 참고하여 유사도를 구하고 이를 격 확률로 사용한다. 그리고 격을 결정하고, 의존 구조를 수정할 때에는 격 확률과 함께 그 격에 대해서 현재 대상이 되는 명사, 조사의 빈도수, 조사의 빈도수 등을 참고로 한다. 어떤 명사, 조사 쌍의 격 확률 계산 결과 문장 내의 두가지 이상의 동사에 대해서 같은 격 확률은 갖게 되면, 명사, 조사의 빈도수가 큰 격틀의 동사로 결정한다. 그리고 빈도수마저 동일하다면, 구문분석 결과로 결정된 의존관계에 따라서 격을 결정하고 의존관계 수정은 하지 않는다.

만약, 해당 명사, 조사 쌍을 동사에 대한 격틀 DB에서 검색할 수 없을 경우에는 동사의 격틀 DB 각 엔트

리에 출현한 명사와 대상 명사와의 유사도를 구해서 이를 격 확률로 이용한다.

계산된 격 확률을 이용하여 선행하는 동사에 이미 의존관계와 격이 결정된 명사, 조사 쌍이라 할지라도 현재 서술어에 의존할 확률이 더 높으면 현재 동사에 대해서 의존관계와 격을 결정하게 된다.

4. 실 험

컴퓨터 관련 FAQ에서 문장을 수집하여 이를 격틀 구축 및 구문분석 후처리를 위한 실험 데이터로 사용하였다. 격틀을 구축할 때 사용된 말뭉치와 구문분석 후처리의 실험 데이터로 사용된 말뭉치는 <표 1>과 같다.

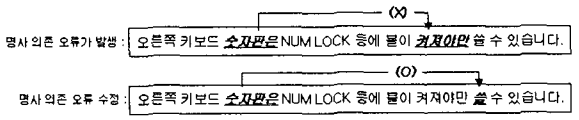
<표 1> 실험에 사용된 말뭉치

격틀 구축 말뭉치		구문 분석 후처리 실험 말뭉치		합 계	
어절수	문장수	어절수	문장수	어절수	문장수
6397	613	1490	128	7887	741

구문분석 후처리 실험용 말뭉치는 변환규칙 학습기를 이용한 한국어 의존 구조 분석기[4]를 사용하여 구문분석을 수행하고 이 결과를 받아서 구문분석 후처리를 수행하였다.

격틀 구축 결과로 총 237개의 서술어에 대한 격틀 데이터를 구축하였고 격틀 데이터베이스는 총 1272개의 항목을 갖는다. 실험 말뭉치 128문장, 1490어절에 대해서 구문분석을 수행한 결과 62개의 어절에서 명사-조사 쌍 의존관계 오류가 발생하였다.

다음 <그림 2>는 명사 의존 오류가 발생한 경우 그 오류를 수정한 예를 보여준다.



<그림 2> 명사-조사 쌍 의존 오류 수정의 예

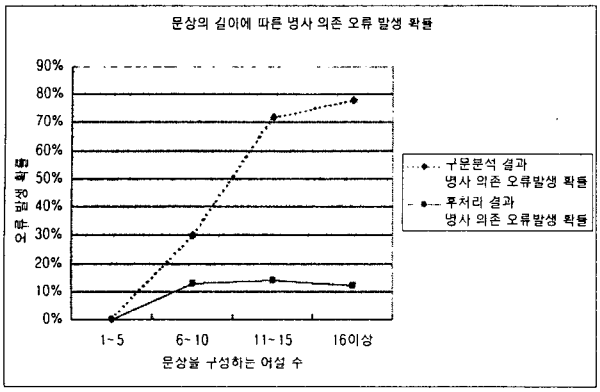
다음 <표 2>는 명사-조사 쌍 의존 오류가 발생한 어절과 이 오류를 수정하기 위해서 수행한 후처리 결과이다.

<표 2> 구문분석 후처리 실험 결과

명사 의존관계 오류 어절의 수	62 어절
명사 의존관계 오류 수정 어절의 수	51 어절
올바르게 수정한 어절의 수	46 어절
정확률	**90%**
재현율	**74%**

구문분석 후처리 결과, 명사-조사 쌍 의존관계 수정의 정확률은 90%, 재현율 74%의 성능을 보였다. 결과를 분석한 결과 재현율이 낮은 이유는 격틀 자료의 부족과, 시소러스의 부족이 주요 원인이었다. 즉, 구축된 격틀 정보가 없는 동사인 경우 격 확률을 계산할 수 없었고 시소러스에 등록되지 않은 명사의 경우 유사한 의미를 가지는 명사를 포함한 격틀 정보가 있다고 하더라도 두 명사간에 유사도를 계산할 수 없었다. 따라서 격틀 자료를 확장하고 시소러스도 확장 시켜서 실험을 수행한다면 좀 더 나은 결과를 얻을 수 있을 것이다. 명사-조사 쌍 의존 관계를 수정한 어절 중에서 올바르게 수정하지 못한 경우는 격틀 자료와 시소러스의 부족과 더불어 또 한가지 문제 때문에 발생하였다. 즉 한 문장에 있는 여러 동사들이 비슷한 격틀 정보를 가지는 경우에 격 확률만으로 의존관계를 수정할 경우 오류가 발생하였다.

구문분석의 결과를 분석해 보면, 문장이 길어질 수록 문장의 복잡도가 증가하여 구문분석의 정확률이 떨어지는 경향이 있다. 다음 <그림 3>은 문장의 길이에 따른 명사-조사 쌍 의존관계 오류가 발생할 확률을 구문분석 후처리 전, 후로 나누어 비교한 결과이다.



<그림 3> 문장의 길이에 따른 명사-조사 쌍 의존관계 오류 발생 확률

구문분석을 수행한 결과 문장의 길이가 길어질 수록 명사의 의존관계 오류가 발생할 확률이 높아진다. 실험 말뭉치를 구문분석한 결과를 분석해 보면, 1~5어절로 구성된 문장의 경우 명사의 의존관계 오류가 없었으나, 6~10어절로 구성된 문장의 경우 문장별 오류 발생 확률이 30%, 11~15어절로 구성된 문장은 72%, 16어절 이상으로 구성된 문장은 78%로 증가하였다. 즉, 구문분석 후처리를 수행하기 전에는 문장이 길어질 수록 명사-조사 쌍 의존관계 오류가 발생할 확률이 선형적으로 증가하였으나, 후처리를 수행한 결과 이 오류 발생 확률은 문장의 길이에 관계없이 12~13%로 안정화 되었다. 따라서 본 시스템은 문장의 길이에 상관 없이 명사 의존관계 오류 발생 확률을 일정한 수준으로 낮출 수 있기 때문에 분석해야 하는 문장이 길고 복잡한 경우 이 시스템의 유용성은 더욱 커질 것이다.

5. 결론 및 향후 과제

본 연구에서는 한국어 동사에 대한 격틀을 구축하고 이를 구문분석 후처리기에 응용하는 시스템을 구현하였다. 격틀은 237개의 동사에 대해서 구축되었으며, 이 정보를 명사-조사 쌍과 동사와의 의존관계 오류가 발생한 문장에 적용하여 의존관계 수정을 위한 구문분석 후처리를 수행하였다. 의존관계 수정 결과 정확률 90%, 재현율 74%로 재현율이 다소 낮으나, 이는 격틀 자료와 시소러스 자료의 부족으로 인한 결과이므로, 격틀 자료와 시소러스를 확장시킨다면, 재현율을 충분히 향상될 수 있다. 의존관계를 결정할 때 격 확률만을 가지고 결정할 경우, 비슷한 격틀 정보를 가지는 동사들의 경우에는 분별력이 떨어지게 된다. 따라서 이러한 경우는 격 확률 뿐만 아니라 어절간의 거리, 문맥정보 등을 고려하여 의존 관계를 결정하는 것이 필요하다. 본 시스템은 구문분석으로 해결하기 힘든 명사와 동사와의 의존 관계를 결정할 수 있는 방법을 제시하였으며, 이를 통해 문장의 길이에 상관 없이 안정적으로 명사의 의존관계 오류를 감소시킬 수 있었다.

참고문헌

[1] 최영림, 신경망을 이용한 한국어 격 분석기의 구현, 한국과학기술원, 전산학과 석사학위논문, 1995

[2] 김창현, 한국어 구문 분석을 위한 오른쪽 우선 챠트 파서, 한국과학기술원, 전산학과, 석사학위논문, 1993

[3] 박경진, 통계적 결정 트리를 이용한 한국어 구문 분석, 서강대학교, 전자계산학과, 석사학위논문, 1997

[4] 이성욱, 변환 규칙 학습기를 이용한 한국어 의존 구조 분석기, 서강대학교, 전자계산학과, 석사학위논문, 1997

[5] 윤준태, 공기 관계 기반 어휘 연관도를 이용한 한국어 구문분석, 연세대학교, 컴퓨터과학과, 박사학위논문, 1997

[6] 김형근, 확률적 의존문법과 한국어 구문분석, 한국과학기술원, 전산학과, 석사학위논문, 1995

[7] 최용석, 격틀 자동구축과 격틀 평가 방법에 관한 연구, 한글 및 한국어 정보처리 학술발표논문집, pp.272-279, 1999

[8] 송영빈, 동사의 애매성 해소를 위한 구문 의미사전의 구축, 한글 및 한국어 정보처리 학술발표논문집, pp.280-287, 1999

[9] 이창기, WordNet을 이용한 한국어 시소러스 자동구축, 한글 및 한국어 정보처리 학술발표논문집,

pp.156-163, 1999
[10] 이문열, 격틀 기반의 패턴매칭을 이용한 구어체 문장분석, 서강대학교, 전자계산학과, 석사학위논문, 1997
[11] Haim Gaifman, Dependency Systems and Phrase Structure Systems, Information and Control, Vol.8, pp.304-337, 1965
[12] Ellen Riloff, An Empirical Approach to Conceptual Case Frame Acquisition, WVLC-98
[13] Kemal oflazer, A Constraint-based Case-frame Lexicon, COLING-96
[14] Farreres, X., Using WordNet for Building WordNets, In proceedings of COLING-ACL Workshop, 1998
[15] Atserias, J., Combining Multiple Methods for the Automatic Construction of Multilingual WordNets, In *Proceeding of the Conference n Recent Advances on* NLP, 1997
[16] Magerman, D., Statistical Decision-Tree Models *for Parsing, Proceedings of the* 34th *Annual Meeting* of the Association for Computational Linguistics, 1995
[17] *Collins, M. J., A New Statistical Parser Based on* Bigram Lexical Dependencies, Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, 1996