요 약

한국어는 다른 자연어와 마찬가지로 많은 중의성을 가지고 있다. 한국어를 분석하는데 있어서 중의성의 해소는 매우 중요한 문제이며, 이 문제를 해결하지 않고 실용적인 한국어 분석 시스템을 개발한다는 것은 불가능하다. 한국어가 가지는 중의성 중, 구조적 중의성을 해소하기 위하여 많은 연구들이 있어왔다. 이들을 크게 분류하면 규칙을 이용한 접근 방법과 확률을 이용한 통계적 접근 방법이 있다. 확률을 이용한 방법에는 확률 문법을 이용한 방법과 확률 어휘정보를 이용한 방법이 있다.

일반적으로 확률 어휘정보는 개념 형태로 표현된다. 개념 수준의 확률 어휘정보는 자동으로 획득하는 것이 힘들어 구축에 많은 수작업을 요구하기 때문에, 기존의 많은 연구들은 코퍼스로부터 어휘 공기 정보를 추출하여 구조적 중의성 해소에 이용하여 왔다. 어휘 공기 정보는 개념이 아닌 어휘 자체로 표현되어 있어, 실제로 코퍼스에서 발생하지 않은 어휘에 대해서는 정보를 제공하지 못하여 자료 부족 문제가 발생할 수 있다. 이에 반해 개념으로 표현된 어휘정보는 코퍼스에서 실제로 공기하지 않은 어휘에 대해서도 적절한 정보를 제공하는 장점이 있다.

본 논문에서는 원시 코퍼스와 시소러스를 이용하여 자동으로 개념 수준의 어휘정보를 구축하여 구조적 중의성을 해소하는 방법을 제안한다. 원시 코퍼스를 자동 태거로 품사 태깅하여 단문으로 분리, 정확한 용언-명사-격조사 패턴을 추출하고, 자동으로 명사를 일반화시켜 용언-개념-격조사로 이루어진 어휘정보를 구축한다. 그리고 이 정보를 구조적 중의성 해소 과정에 적용한다.

약 800만 어절의 원시 코퍼스로부터 624,200개 가량의 공기된 용언-명사-격조사 어휘를 추출하였고, 이 어휘 공기 정보 중 명사를 시소러스를 이용하여 일반화 시켜 약 5,000개 정도의 용언-개념-격조사 쌍을 자동으로 구축하였다. 자동 구축된 어

휘정보를 명사구의 지배소 선택 실험에 적용한 결과, 실험 코퍼스에서 약 91.5%의 명사구 지배소 선택 정확도를 얻을 수 있었다.

목 차

그림 목차

표 목차

제 1 장

서 론

자연어 처리에서 구문 분석이란 주어진 문장의 구조를 주어진 구문 규칙에 따라 분석하는 작업을 말한다. 구문 분석 과정에서는 보통 하나 이상의 구문 구조가 구해지며, 이들 중 올바른 구문 구조를 선택하는 작업을 구조적 중의성의 해소 작업이라고 한다.

구조적 중의성을 해소하는 방법에는 규칙을 이용한 접근 방법과 통계를 이용한 접근 방법이 있는데, 최근에는 통계적 접근 방법이 널리 사용되고 있다. 통계적 접근 방식은 대량의 코퍼스로부터 구조적 중의성 해소에 필요한 확률 정보를 추출하기 때문에, 실제 사람들이 문장을 사용하는 경향을 쉽게 반영할 수 있으며, 지식 획득이 용이하다는 장점을 갖는다. 흔히 사용되는 방법에는 확률 문맥 자유 문법 (probabilistic context-free grammar)이나 확률 의존 문법 (probabilistic dependency grammar)과 같은 확률 문법을 이용하는 방법이 있다. 확률 문법을 사용한 구조적 중의성 해결은 매우 간단하다. 구문 트리의 확률값은 그 구문 구조에서 사용되어진 확률 규칙의 확률값의 곱이며, 이와 같이 얻어진 각 구문 트리의 확률값 중 가장 높은 값을 지닌 구문 구조가 입력 문장에 대해 가장 적절한 결과 트리로써 선택되어진다[이공주97].

구조적 중의성의 해소를 위하여 사용되는 통계적 접근 방식 중, 또 다른 방법

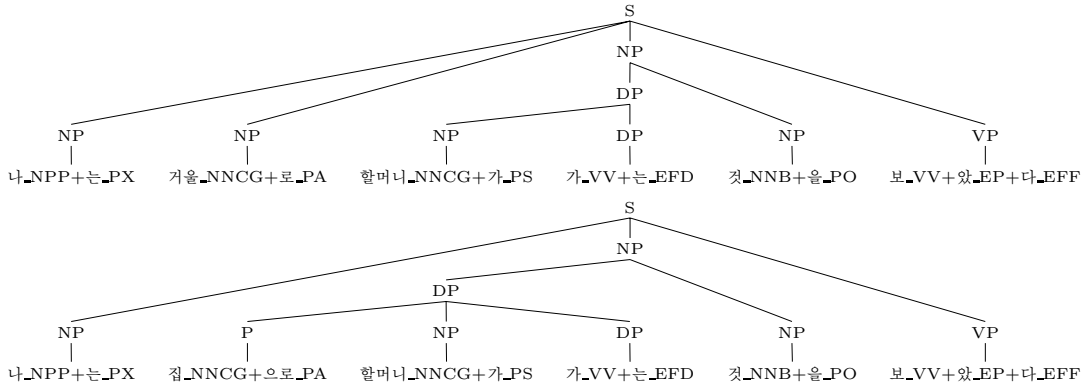


그림 1.1: 동일한 품사열을 가지나 구문 구조가 틀린 문장

에는 술어 하위 범주 정보와 격틀 정보, 어휘 공기 정보와 같은 어휘정보 (lexical information)에 확률을 부여하여 사용하는 방법이 있다. 어휘 정보란 어휘 자체가 가지는 특징들을 기술해 놓은 정보를 말한다.

두 방법 중 확률 문법을 이용한 방법은, 일반적으로 선호되는 구문 구조에 높은 확률값을 부여하기 때문에, 정확한 구문 분석을 하는데 부족함이 있다. 다음의 예문을 살펴 보자.

- 나는 거울로 할머니가 가는 것을 보았다.
- 나는 집으로 할머니가 가는 것을 보았다.

두 예문은 동일한 품사열로 구성되어 있으나, 구문 구조는 서로 다르다[그림 1.1] 하지만 확률 문법을 사용하여 위의 두 문장을 분석한다면 두 문장을 동일한 구조로 분석할 것이다. 일반적으로 확률 문법은 어휘는 고려하지 않고, 품사 태그만으로 규칙들을 표현하기 때문에 두 문장의 차이점을 구별해 내지 못하기 때문이다. 따라서 확률 문법을 보완할 추가의 정보가 필요한데, 확률 어휘정보는 이런 경우 올바른 구문 분석을 하는데 도움이 된다. “가다”라는 동사는 “거울로”와는 어울리지 않는다는 정보가 있으면 첫번째 예문을 두번째 구문 구조로 결정하지는 않을 것이다.

확률 어휘정보는 구축하기가 힘들다는 문제점이 있다. 확률은 코퍼스로부터 자동 학습이 가능하지만, 어휘정보 자체의 구축에는 많은 양의 수작업과 시간이 요구되기 때문이다. 영어권에서는 이미 수작업에 의해 FrameNet, COMLEX syntactic dictionary, LDOCE(Longman Dictionary of Contemporary English), Alvey NL Tools (ANLT) Dictionary 등의 기계 가독형 어휘정보가 구축되어 있으나, 국어에 대해서는 이러한 정보가 아직 구축되어 있지 않기 때문에, 국내에서는 구조적 중의성을 해결하기 위해서 단순히 어휘 공기 정보 확률을 코퍼스로부터 추출하여 사용하는 수준에 머물고 있다 [이공주97, 윤준태97]. 이런 어휘 공기 정보 확률은 어휘 자체의 공기 확률만을 고려하므로 실제로 중의성 해소에 적용시 자료 부족 문제 (data sparseness problem)를 일으킬 가능성이 매우 크다는 문제점을 가지고 있다. 따라서 어휘정보는 구조적 중의성을 해소하기 위해 필수적이며, 이러한 이유로 국내에서는 어휘 정보를 좀 더 쉽게 반자동으로 구축하는 방법에 대한 연구가 최근 활발히 진행되고 있고, 최근 어휘 공기 정보를 개념 패턴이라 불리는 어휘정보 형태의 정보로 자동으로 일반화하여 격 중의성을 해소하는 연구가 발표되었다[이휘붕98].

한편, 어휘정보가 구축되어 있지 않은 국내의 경우 뿐만 아니라, 이미 어휘정보가 상당량 구축되어 있는 영어권의 경우에도 어휘정보의 자동 구축 관련 연구는 최근 들어 활발히 진행되고 있다. 그 이유는 자동 구축된 어휘정보가 수동으로 작성된 어휘정보에 비해 다음과 같은 장점을 가지고 있기 때문이다[Manning93].

- 전문적인 분야에서 사용되는 특별한 용도를 위한 동사에 대한 정보는 수동으로 작성된 어휘정보 사전에서 구할 수가 없다.
- 수동으로 사전을 작성하는 것은 상당한 수작업량을 필요로 하기 때문에 많은 동사에 대하여 어휘정보를 작성하기가 힘들다.
- 코퍼스로부터 자동으로 어휘정보를 획득하면 정보의 갱신이 매우 쉬워지며, 전문 분야용 어휘정보 사전을 작성하는 것도 쉬워진다.

이런 장점때문에 영어권에서는 90년대부터 코퍼스로부터 어휘정보를 자동 획

득하는 방법이 연구되었고[Brent91], 이에 확률을 부여하여 구조적 중의성 해소에 이용하고 있다. 국어의 경우에는 어휘정보를 자동으로 획득하여 구조적 중의성을 해결하는 연구가 아직 이루어지지 않았다.

외국의 연구와는 달리 국내에서 자동으로 어휘정보를 추출하는 연구가 늦어진 것은 적절한 시소러스(thesaurus)가 존재하지 않았기 때문이다. 시소러스는 어휘를 개념으로 확장하는데 필수적인 지식원이라고 할 수 있다.

또 다른 이유는 구문 태깅된 코퍼스가 부족하다는 것이다. 영어의 경우, The Penn TreeBank 프로젝트[Marcus93]에서 구축한 구문 태깅된 코퍼스(92년 11월 기준, 약 3,000,000 단어) 가 이미 널리 사용되고 있으나, 국내의 경우 97년에서야 30,000문장(796,449형태소)의 구문 태깅된 코퍼스가 발표되었다. 구문 태깅된 코퍼스가 없으면 올바른 의존 관계를 갖는 어휘를 추출하기가 힘들어지므로, 어휘 정보를 구축하는 것이 어려워 진다.

본 논문에서는 원시코퍼스와 시소러스를 이용하여 자동으로 확률 어휘정보를 구축하고, 구축된 어휘 정보를 구조적 중의성을 해소하는데 적용하는 방법을 제안하고자 한다. 여기서 말하는 어휘정보란 개개의 술어가 가지는 인자에 대한 정보를 말한다. 즉, 어떤 동사나 형용사가 가질 수 있는 격조사 및 그 격조사와 함께 나타나는 개념들에 대한 정보이다. 한국어에서는 술어의 역할이 매우 중요하기 때문에 술어-인자에 대한 어휘정보는 자연어 처리의 여러 분야에서 매우 유용하게 것이라고 기대된다.

제 2 장

관련 연구

본 장에서는 어휘정보의 구축에 관한 연구와, 이를 이용하여 구조적 중의성을 해결한 국내외의 관련 연구들을 살펴 본다.

2.1 영어 관련 연구

영어의 경우, 90년대 초부터 자동으로 동사 하위 범주 정보나 격틀 정보 형태의 어휘 정보를 자동으로 추출하는 연구가 진행되어 왔다. 구문 분석을 위해 일반적으로 많이 사용되던 확률 문법은 어휘에 대한 정보를 이용하지 못한다는 문제점이 있었고, 따라서 구문 분석기에서 사용할 수 있는 어휘에 관한 정보를 구축하는 작업이 필요했다[Ushioda93].

이런 이유로 [Brent91] 과 [Ushioda93]에서는 코퍼스를 이용하여 몇몇 동사에 대한 구문 구조 패턴을 자동으로 구축하는 시스템을 제안했다. 여기서 말하는 구문 구조 패턴이란 어떤 동사가 어떤 형태로 사용되느냐에 대한 정보이다. [Ushioda93]은 구문 구조 패턴에 통계 정보를 포함시켰는데, 품사 태깅된 코퍼스와 finite-state NP 파서를 이용하였다. [Manning93] 역시 품사 태깅된 코퍼스와 finite-state NP 파서를 이용하여 좀 더 많은 양의 구문 구조 패턴 정보를 추출했다. [Briscoe97] 에서는

자동 파서를 이용하여 동사의 문형 패턴을 추출한 후에, 이를 미리 정의한 구문 구조 패턴에 매핑시켜 동사 하위 범주 정보를 자동으로 구축하였다.

[Hindle93]에서는 구문적 중의성을 유발하는 전치사구 부착 (prepositional phrase attachments) 문제를 해결하기 위하여 구문 태깅된 코퍼스로부터 (명사, 동사, 전치사) 리스트를 추출, 확률을 부여하여 이용하였다. [Resnik93]은 [Hindle93]의 시도를 확장하여 데이터 요구도를 줄이기 위해서 WordNet이라는 온라인 시소러스를 사용해 단어 단위가 아닌 클래스(class) 단위로의 접근을 시도하였다. [Li95]에서는 패턴 매칭 방법을 이용하여 태깅된 코퍼스로부터 격틀 패턴을 추출한 후, MDL(Minimum Description Length) 원리와 WordNet을 이용하여 격틀 일반화를 시도하였고, 이 격틀을 전치사구 부착 문제에 적용하였다.

[Collins96]에서는 트리태깅된 코퍼스로부터 어휘 사이의 연관도를 계산한 후, 구문 분석 과정에 적용하였다. 어휘 일반화과정을 거치지 않음으로 발생하는 자료 부족 문제를 back-off 를 이용하여 완화시켰다.

2.2 한국어 관련 연구

[양재형94]에서는 한국어 분석기를 이용하여 코퍼스로부터 명사와 동사, 그리고 격관계(주격, 목적격) 패턴과 빈도를 추출하여 통계적으로 미지격 문제를 해결하였다. 그리고 이러한 패턴이 구조적 중의성의 해소에도 적용될 수 있다고 언급하였다. [엄미현96b]에서는 구조적 중의성을 해결하기 위하여 격 정보 및 어휘 사이의 상호정보를 이용하였다. 어휘 사이의 상호정보는 코퍼스로부터 뽑은 어휘공기정보를 이용하여 구하였다. 실제로 구조적 중의성 해소에 어휘 상호 정보를 적용하여, 상호 정보의 차가 클 때 분석 정확도가 높음을 실험으로 보여주었다.

[김선호96]에서는 원시 코퍼스 및 태깅된 코퍼스를 이용하여 반자동으로 (동사, 격조사, 명사) 형태의 패턴을 추출하였다. 그리고 시소러스를 이용하여 수동으로 어휘의 개념으로의 일반화를 수행한 후, 이 정보를 미지격 결정에 적용하였다. 또한 이 정보를 구조적 중의성에도 적용 가능하다고 언급하였다. [이공주97]에서는 확

률 구구조 문법으로 해결하지 못한 구조적 중의성 문제를 해결하기 위하여 확률 문법을 학습한 구문 태그 부착 코퍼스로부터 추출한 어휘 공기 정보를 이용하였다. [장명길97]에서는 [류법모97]에서 반자동으로 작성한 술어 하위 범주 정보를 활용하여 구문 분석을 수행하여 정확도를 높였다. [윤준태97]에서는 원시 코퍼스로부터 자동 추출한 어휘 공기 정보로 계산되는 어휘 연관도를 이용한 한국어 구문 분석 모델을 제안하였다. [윤준태98] 에서는 [윤준태97]과 같은 방법으로 추출한 어휘 공기 정보를 이용하여 연어 리스트를 생성한 후, 구문 분석 전처리 단계에서 적용하여 구문 분석 단계에서의 계산량을 감소시켰다.

이런 연구들의 공통점은 구조적 중의성을 해소하기 위하여 사용한 개념 수준(concept level)의 어휘정보의 경우, 모두 수작업 혹은 반자동으로 작성했다는 것이다. 그리고 개념 수준이 아닌 어휘 수준의 정보, 이를테면 어휘 공기 정보,들은 자동으로 코퍼스로부터 추출이 가능하다는 장점이 있으나, 실제로 구문 분석에 적용시 자료 부족 문제를 유발한다는 단점이 있다. 최근 발표된 [이휘붕98]에서는 코퍼스와 시소러스를 이용하여 자동으로 개념 수준의 어휘 정보를 추출하여 미지격 해소를 시도했는데, 이를 제외하면 어휘 수준의 공기 정보만을 자동으로 추출하여 이용하는 연구가 계속되어 왔다.

본 논문에서는 대량의 수작업이 요구되었던 개념 수준의 어휘정보 구축의 단점과 자료 부족 문제를 유발할 가능성이 있던 어휘 수준의 공기 정보의 단점을 보완하여 자동으로 개념 수준의 어휘정보를 구축하여 구조적 중의성을 해결하는 방법을 제안한다.

제 3 장

코퍼스와 시소러스를 이용한 확률 어휘정보의 자동 구축

본 장에서는 자동으로 확률 어휘정보를 구축하는 방법을 제안한다. 여기서 말하는 어휘정보란 용언과 그 용언이 요구하는 격조사 및 그 격조사와 함께 쓰이는 개념들에 대한 정보를 말한다. 예를 들어 "먹다"라는 동사의 목적어로는 "음식"이 온다는 정보를 말한다. 확률 어휘정보를 자동으로 구축하는 방법은 크게 두 단계로 나뉘어지는데, 첫 단계에서는 원시코퍼스로부터 (용언, 격조사, 명사) 형태의 어휘 공기정보 및 빈도를 추출하고, 두 번째 단계에서는 앞 단계에서 구한 공기 정보 중 명사 부분을 시소러스의 개념으로 일반화 하여 (용언, 격조사, 개념) 정보를 구축한다.

3.1 어휘 공기 정보의 추출

어휘 정보는 어휘 공기 정보를 기반으로 구축된다. 어휘 공기 정보란 코퍼스에서 술어-인자 관계를 가지며 함께 등장하는 (용언, 격조사, 명사)를 말한다. 예를 들어

밥을 먹었다.

라는 문장이 코퍼스에 있을 때, 이 문장에서 추출되는 공기 정보는 (먹다, 을, 밥)이 된다. 어휘 공기 정보를 추출하는데 고려해야할 사항들은 다음과 같다.

3.1.1 의존 관계를 갖는 어휘 추출을 위한 고려 사항

어휘 공기 정보로 추출되는 술어, 격조사, 명사는 의존 관계를 가지고 있어야 한다. 코퍼스에서 술어-격조사-명사 사이의 의존 관계를 정확하게 알아내기 위해서는 구문 태깅된 코퍼스가 필요하다. 그러나 국내에 존재하는 구문 태깅된 코퍼스의 양이 적기 때문에, 원하는 어휘 공기 정보를 추출하기에는 부족하다.

유사한 방법으로 대량의 단문이 있다면 구문 분석 되어 있지 않아도, 정확한 의존 관계를 찾아 낼 수 있기 때문에, 공기 정보를 추출하는데 유용할 것이다. 하지만, 실제로 코퍼스에 등장하는 문장들은 단문의 형태로 존재하기 보다는 거의 복문의 형태를 취하고 있다. 그러므로 코퍼스로부터 단문 형태로 이미 존재하는 것만을 추출하는 경우 현재 구축되어 있지 않은 너무 광대한 코퍼스를 필요로 한다 [양단희98].

이에 본 논문에서는 원시 코퍼스의 문장들을 자동 품사 태거로 태깅한 후, 다음과 같은 방법으로 정확한 의존 관계를 갖는 술어, 격조사, 명사를 추출한다.

- 용언의 연결어미를 기준으로 복문을 단문으로 분리한다.
- 각 단문의 마지막 용언과 그 앞 용언 사이에 있는 인자들만을 어휘 공기 정보 추출 대상으로 삼는다.

이런 방법을 사용하면 잘못된 용언-격조사-명사 공기 정보를 추출할 가능성이 배제된다. 이 단계에서 오류를 포함할 경우 다음 단계인 명사 일반화 단계에서 오류를 파생시킬 수 있기 때문에 추출되는 정보의 양이 적더라도 정확히 의존 관계를 갖는 어휘들만을 추출하는 것이 바람직하다.

원시코퍼스 : 학교에서 열린 학회에 참석하고, 우리는 밥을 먹었다.

⇓

자동 태깅 : 학교_NNCG+에서_PA 열리_VV+ㄴ_EFD 학회_NNCG+에_PA
참석_NNCV+하_XSVV+고_EFC+,_SS, 우리_NPP+는_PX
밥_NNCG+을_PO 먹_VV+었_EP+다_EFF+._SS.

⇓

단문 분할 : 학교_NNCG+에서_PA 열리_VV+ㄴ_EFD 학회_NNCG+에_PA
참석_NNCV+하_XSVV+고_EFC

+

우리_NPP+는_PX 밥_NNCG+을_PO 먹_VV+었_EP+다_EFF+._SS.

⇓

어휘공기정보 : (참석하, 학회, 에) (먹, 밥, 을)

그림 3.1: 공기 정보 추출의 예

[그림 3.1]는 원시 코퍼스의 한 문장으로부터 용언-격조사-명사 정보를 추출하는 예이다. 예문의 "학교에서"라는 명사+격조사 어절은 지배 술어가 "열린"인지 "참석하지"인지 알 수가 없으므로 (열리, 학교, 에서)라는 어휘 쌍은 오류의 가능성이 있다고 여겨져 추출되지 않는다. 또 "우리는"이라는 어절은 술어 "먹었다"와 의존 관계를 갖는 것이 분명하나, 격조사가 아닌 보조사가 사용되어 격관계를 알 수 없으므로 (먹, 우리, 는) 어휘 쌍 또한 추출되지 않는다.

3.1.2 술어에 대한 고려 사항

술어 중 용언(동사와 형용사)만을 정보 추출 대상으로 삼는다. 지정사라고 불리는 "명사+서술형 전성 어미" 형태의 술어는 어휘 공기 정보 추출 대상으로 삼지 않는다. 다음은 공기 정보 추출시에 고려한 사항이다.

- 동사나 형용사로 태깅된 술어는 술어의 어간을 추출한다.
  예 : 아름답다(아름답_VJ+다_EFF) ⇒ 아름답
- 동작성 명사나 상태성 명사가 사용된 술어는 명사+하(용언화 접미사)의 형태로 추출한다.
  예 : 공부하다(공부_NNCV+하_XSVV+다_EFF) ⇒ 공부하
- "용언 + 연결어미 + 보조용언"으로 이루어진 복합 술어의 경우 복합 술어 전체를 추출한다.
  예 : 변해왔다(변하_VV+어_EFC+오_VX+었_EP+다_EFF) ⇒ 변해오

3.1.3 격조사에 대한 고려 사항

격조사란 체언과 다른 말과의 관계를 나타내는 조사를 말한다[조규빈95]. 격조사는 크게 주격 조사, 목적격 조사, 보격 조사, 관형격 조사, 부사격 조사, 호격 조사와 같은 일반 격조사와 서술격 조사로 나뉘는데, 본 논문에서 격조사를 추출하는 이유는 용언이 필요로 하는 격조사를 알아내기 위한 것이므로, 용언과 관계를 갖지 않는 관

역할	대표 조사	해당 조사
주격	가	가, 이, 께서, ...
목적격	를	를, 을
방향/처소격	에	에, 에게, 한테, 에는, 에도, ...
도구/시발격	로	로, 으로, 으로는, ...
처소격	서	서, 에서, 에서는, ...

표 3.1: 격조사의 문법기능에 따른 분류

형격 조사, 호격 조사, 서술격 조사는 추출 대상에서 제외한다. 또 현재 자동 품사 태거가 보격 조사와 주격 조사를 제대로 구별해내지 못하기때문에 보격 조사 역시 정보 추출 대상에서 제외하고 주격 조사, 목적격 조사, 부사격 조사 만을 고려한다.

본 논문에서는 명사가 용언에 대해 가지는 문법 기능을 중심으로 주격 조사, 목적격 조사, 부사격 조사를 분류하여 [그림 3.1]과 같이 5가지로 분류하였다. 이 5가지 분류의 격조사는 명사와 용언과의 문법 기능 관계 중 가장 많이 나타나는 것이다. 분류에 포함되지 않은 부사격 조사들은 추출 대상에서 제외한다.

3.1.4 명사에 대한 고려 사항

여기서 "명사"라 함은 술어에 대한 인자 역할을 하는 어절에서 격조사의 앞에오는 어휘를 말한다. 즉, 일반적인 명사 뿐만 아니라, 고유 명사, 대명사, 수사까지 포함하는 의미이다. 명사에 대한 공기 정보 추출시 다음과 같은 사항을 고려한다.

- 복합 명사의 경우, 머리어 역할을 하는 가장 마지막 명사만을 추출한다.
  예 : 학술대회(학술_NNCG+대회_NNCG) ⇒ 대회
- "들" 이나 "끼리" 같은 접미사가 명사에 붙은 경우, 접미사는 제외하고 명사만 기록한다.
  예 : 사람들(사람_NNCG+들_XSNN) ⇒ 사람

- 한 술어에 두 개 이상의 동일한 격조사가 의존 관계를 갖는 것이 확실한 경우, 앞 쪽에 위치하는 격조사 및, 명사는 정보 추출 대상에서 제외하였다.
  예 : 그 사람이 손이 잘렸다. (사람_NNCG+이_PS, 손_NNCG+이_PS) ⇒ 손-이-잘리 ("사람-이-잘리"는 추출되지 않음)

- 품사 태거가 "명사추정(NN?)"이라고 분석한 것은 추출 대상에서 제외한다.

위와 같이 고려하여 원시 코퍼스의 모든 문장으로부터 어휘 공기 정보를 추출하면 [윤준태97] 에서 사용한 어휘 공기 정보와 유사한 형태의 정보가 추출된다[표 3.2]. 이 자체만으로도 구문 분석에서 유용하게 사용될 수 있지만 어휘 수준 (lexical-level)의 정보만을 포함하고 있으므로 자료 부족 문제를 유발할 가능성이 있다.

3.2 명사의 일반화

명사의 일반화란 앞 단계에서 추출한 (용언, 명사, 격조사) 어휘 공기 정보에서 명사부분을 시소러스의 개념 항목으로 대체하는 것이다. 시소러스란 명사들을 개념별로 분류하여 계층화 시켜놓은 사전을 말한다. 그런데 명사를 시소러스의 개념항목으로 대체하기 위해서는 명사가 사용된 의미(word sense)를 알아야 한다. 예를 들어 [조평옥97]에서 생성한 시소러스에서 명사 '방문'은 '문'이라는 개념의 하위 항목인 의미와, '글'이라는 개념의 하위 항목인 의미, 두 가지[1]를 갖는데, 어떤 의미로 사용되었는지를 모른다면 적절한 개념으로 일반화해 줄 수가 없기 때문이다. 명사가 사용된 의미를 결정했다면 같은 (용언, 격조사) 쌍과 함께 사용된 명사들을 하나로 묶어 (clustering), 개념으로 바꿔주는 작업을 수행해야 한다.

3.2.1 명사의 의미 결정

명사의 의미를 알아내기 위해 본 논문에서는 다음과 같은 가정을 사용하였다.

[1]방문 : 1. 방으로 드나드는 문 2. 널리 알리기 위하여 길거리 등에 써붙이는 글

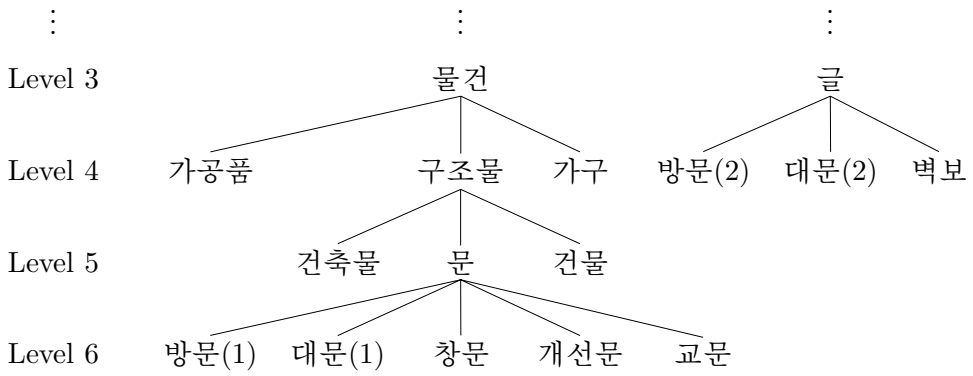


그림 3.2: 시소러스의 예

- 가정 : 같은 격조사 및, 같은 술어와 함께 사용되는 명사들은 유사한 의미 자질을 갖는다.

두 단어가 유사한 의미 자질을 갖는다는 것은 시소러스에서 두 단어 사이에 MSCA[2]가 존재하며, 그 단어들 사이의 거리가 가깝다는 것을 의미한다. 예를 들어, 용언-격조사 쌍인 (열리다, 가)와 함께 사용된 명사들 "교문"과 "대문"은 시소러스에서 두 단어 사이에 MSCA가 존재하고, 두 단어 사이의 거리가 가깝다는 것이다.

그런데 시소러스에서 두 명사의 MSCA 깊이가 깊을 수록 두 명사의 의미가 더욱 유사하다는 것을 고려해야 하므로, 시소러스에서 두 명사 n과 m 사이의 상대적 거리 $d(n,m)$ 를 다음과 같이 정의한다.

$$d(n,m) = \frac{dist\_in\_thesaurus(n,m)}{depth(MSCA(n,m))} \tag{3.1}$$

위 식에서 $dist\_in\_thesaurus(n,m)$ 은 시소러스에서 두 개념 사이의 실제 거리를 의미하고, $depth(w)$는 시소러스에서의 주어진 개념 w의 깊이를 의미한다. [그림 3.2]에서 "대문"과 "개선문"사이의 실제 거리는 2(대문-문-개선문)이고, MSCA

[2]MSCA(Most Specific Common Abstraction) : 시소러스에서 두 단어가 가지는 공통된 개념 중 가장 구체적인 개념

인 “문”의 깊이는 5이므로, 두 단어 사이의 상대적 거리 d(대문,창문)는 $\frac{2}{5} = 0.4$가 된다.

문장에서 사용된 명사 n의 의미를 결정할 때는 다음과 같은 식을 이용하여 두 명사 사이의 상대적 거리 d가 최소가 되는 명사의 의미 i를 선택한다.

$$sense(n) = \text{argmin}_i d(n_i, m_{j,k}) \qquad (m_j \in M) \tag{3.2}$$

식 (3.2)에서 M은 n과 같은 용언-격조사와 사용되는 명사들의 집합이다. n_i란 명사 n의 i번째 의미에 해당하는 시소러스에서의 개념을 말하고, $m_{j,k}$란 m_j가 가지는 k 번째 개념을 말한다. 가정에 의해서 위 수식의 결과인 i가 명사 n의 의미가 된다. 이를 이용하면 앞 단계에서 추출한 용언-명사-격조사 리스트를 가지고 리스트 내 명사들의 의미 결정을 할 수 있게 된다.

명사의 의미를 결정하는 예를 살펴보자. 용언, 격조사 쌍인 (열리다, 가)와 함께 등장하는 명사에는 “방문”, “창문”, “문”, “회의” 등이 있다[표 3.2]. 이 중에서 두가지 의미를 가지고 있는 명사 “방문”의 의미를 식 (3.2)를 이용하여 결정해보겠다.

n : 방문 $\qquad (0 \le i \le 1)$

M : {문, 창문, 대문, 회의, ...}

- j = 1 : m_j가 “문”이 된다. 문은 하나의 의미만을 가지고 있기 때문에 k 값을 고려할 필요가 없다.
  d(방문(1),문) $= \frac{1}{5} = 0.2$
  d(방문(2),문) 는 두 개념 사이에 MSCA가 존재하지 않으므로 구할 수가 없다.

- j = 2 : m_j가 “창문”이 된다. 창문도 하나의 의미만을 가지고 있기 때문에 k 값을 고려할 필요가 없다.
  d(방문(1),창문) $= \frac{2}{5} = 0.4$
  d(방문(2),창문) 는 MSCA가 존재하지 않으므로 구할 수 없다.

- j = 3 : m_j가 “대문”이 된다. 대문은 2개의 의미를 가지고 있다.

d(방문(1),대문(1)) $= \frac{2}{5} = 0.4$
d(방문(1),대문(2))는 MSCA가 존재하지 않으므로 구할 수가 없다.
d(방문(2),대문(1))는 MSCA가 존재하지 않으므로 구할 수가 없다.
d(방문(2),대문(2)) $= \frac{2}{2} = 1$

- j = 4 : m_j가 "회의"가 된다. 회의는 하나의 의미 만을 가지고 있다. d(방문(1),회의) 는 MSCA가 존재하지 않으므로 구할 수 없다. d(방문(2),회의) 는 MSCA가 존재하지 않으므로 구할 수 없다.

- j=5 . . . : 생략

식 (3.2)의 결과가 최소인 경우는 j가 1인 경우이며(d=0.2), 따라서 명사 "방문"은 "문"의 하위 개념인 첫번째 의미로 사용되었다고 결정된다. 이와 같은 방법으로 모든 명사에 대하여 의미 결정을 해줄 수 있다.

3.2.2 명사의 군집화

명사의 의미를 결정했다면, 비슷한 개념을 가지는 명사들을 묶어 일반화하는 작업이 필요한데, 이는 시소러스에서 찾은 명사들의 MSCA를 이용한다. 동일한 술어-격조사를 가지는 모든 명사들을 쌍으로 묶어 클러스터로 만든 후, 두 클러스터의 MSCA 사이에 조상-자손 관계가 존재하는 클러스터들을 하나의 클러스터로 묶으면 된다. 앞 서 살펴본 명사들 "문", "방문", "대문", "창문"은 "문"이라는 개념으로 군집화 된다.

모든 동사와 격조사 및 격조사와 함께 추출된 명사에 대해서 이 작업을 수행하면 일반화된 어휘정보를 얻을 수 있다. [표 3.3]은 위와 같은 과정을 거쳐 얻은 동사 "열리"의 일반화된 어휘정보이다.[3]

[3]각 어휘 및 격조사가 갖는 확률은 표시되어 있지 않다.

술어	격조사	명사	빈도	술어	격조사	명사	빈도
열리	주격	방문	26	열리	부사격(에)	손	2
열리	주격	문	112	열리	부사격(에)	뜻밖	2
열리	주격	창문	8	열리	부사격(에)	위	2
열리	주격	바위	1	열리	부사격(에)	때문	6
열리	주격	세계	4	열리	부사격(에)	전	3
열리	주격	회의	40	열리	부사격(에)	끝	2
열리	주격	대문	9	열리	부사격(에)	나무	4
열리	주격	성문	1	열리	부사격(로)	세계	3
열리	주격	교문	1	열리	부사격(로)	위	4
열리	주격	대회	51	열리	부사격(로)	적	7
열리	주격	뒷문	1	열리	부사격(로)	주최	17
열리	주격	윗미닫이	1	열리	부사격(로)	구호	1
열리	주격	들창	2	열리	부사격(에서)	누상	1
열리	주격	모임	4	열리	부사격(에서)	댁	1
열리	주격	잔치	5	열리	부사격(에서)	집	5
열리	주격	말문	7	열리	부사격(에서)	취리히	1

표 3.2: 동사 “열리”에 관한 어휘 공기 정보 중 일부

열리다_VV	
격조사	클래스(의미코드) : 어휘들
가	문(0) : 문, 대문, 창문, 방문
	모임(0) : 청회, 공청회, 연주회, 토론회, 전시회, 음악회, 대회, 회의
	기관(0) : 국회, 위원회
	열매(0) : 과일, 열매
로	앞(1) : 처음, 앞
서	건물(0) : 집, 전당, 회관
	지역(0) : 각지, 지역

표 3.3: 동사 “열리”에 대한 어휘정보

제 4 장

구조적 중의성의 해소

구조적 중의성이란 주어진 문장이 여러 구조로 분석될 수 있는 성질을 의미한다. 본 장에서는 본 논문에서 해결하려는 구조적 중의성의 형태와, 이를 해결하기 위하여 구문 분석 단계에서 어휘정보를 사용하는 방법을 설명한다.

4.1 구조적 중의성

한국어는 다른 자연어와 마찬가지로 많은 중의성을 가지고 있다. 한국어를 분석하는데 있어서 중의성의 해소는 매우 중요한 문제이며, 이 문제를 해결하지 않고 실용적인 한국어 분석 시스템을 개발한다는 것은 불가능하다 [엄미현96a].

본논문에서는 의존 문법을 이용한 한국어의 구문 분석 과정에서 발생하는 구조적 중의성 중, 명사구의 지배소 선택 중의성을 해결하고자 한다. 이는 어휘정보를 사용하지 않고서는 해결하기 어려운 문제이다.

의존 문법에서 의존 관계는 두 언어요소 사이에 존재하는데, 이 중 한 언어요소는 지배소(governor)가 되며 다른 한 언어요소는 의존소(dependent)가 된다. 의존 관계에 있는 두 언어요소중 지배소는 의미의 중심이 되는 요소를 말하며, 의존소는 지배소가 갖는 의미를 보완해 주는 요소를 말한다[나동렬94]. 의존 문법은 의존 규

	동사	형용사	일반명사	고유명사
일반명사	○	○	○	×
주격조사	○	○	×	×
보격조사	○	○	×	×
목적격조사	○	○	×	×
보격조사	○	○	×	×
관형격조사	×	×	○	○

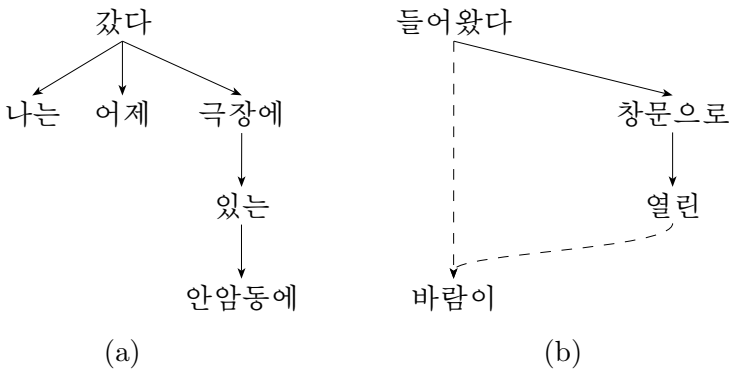
○, ×는 의존관계의 유무를 표시

표 4.1: 의존 규칙의 일부

칙들에 의해서 정의되며, 의존 규칙은 의존 관계를 맺을 수 있는 언어 요소들을 정의한다[표 4.1]. 일반적으로 한국어에서는 술어가 인자들을, 피수식어가 수식어를 지배한다고 정의된다. 또한 한국어의 특성상, 지배소가 의존소보다 뒤에 온다고 가정한다.

의존문법을 이용하여 문장 "나는 어제 안암동에 있는 극장에 갔다"를 분석하면 [그림 4.1]의 (a)와 같다. "갔다"는 전체 문장의 지배소가 된다. "나는", "어제", "극장에"는 "갔다"라는 지배소의 의존소들이다. 또, 어절 "있는"은 "극장에"의 의존소이며, "안암동에"는 "있는"의 의존소가 된다.

명사구의 지배소 선택 문제란 의존 문법을 사용하는 구문분석기에서 지배 술어가 확실치 않은 명사구의 올바른 지배소를 선택해 주는 문제를 말한다. 즉, [그림 4.1](b)에서 명사구 "바람이"의 지배소가 "열린"이 되는지, "들어왔다"가 되는지를 결정해 주는 문제이다. 이 문제는 의존 문법에서 뿐만 아니라 구구조 문법에서도 그대로 발생한다. [그림 1.1]에서 보듯이, 동일한 품사열을 가지고 있더라도, 어떠한 어휘가 사용되었냐에 따라서 구문 구조가 틀려지므로 어휘정보를 사용하지 않고는 해결하기가 불가능하다.



(a) 의존 문법으로 구문 분석한 트리

(b) 구조적 중의성을 유발하는 어절 “바람이”

그림 4.1: 의존 트리

4.2 어휘 연관도

본 논문에서는 연관도 함수 $Assoc()$를 통하여 자동 구축한 확률어휘정보를 구문분석기에 적용하여 명사구의 지배소 선택 문제를 해결한다. $Assoc()$는 다음과 같이 계산한다[윤준태97] [김선호96].

$$Assoc(v,n,j) = \alpha \times \overline{Assoc}(v,n,j) + (1-\alpha) \times \overline{Assoc}(v,j) \qquad (0.5 \leq \alpha \leq 1) \quad (4.1)$$

연관도 함수 $Assoc(v,n,j)$는 용언, 명사, 격조사가 통계적으로 어느 정도 관련이 있는지를 표현하는 함수이다. 그런데, 연관도 값을 구하려는 용언, 격조사, 명사가 코퍼스에서 함께 나타나지 않는 자료 부족 문제가 발생할 수 있다. 격조사의 수는 일정하고, 용언의 수도 명사의 수보다는 월등히 적기 때문에, 이런 경우 보통 문제가 되는 것은 명사이다. 따라서 명사를 제외한 용언, 격조사 공기 확률 값을 함께 고려해줌으로써, 자료 부족 문제를 완화시킬 수 있다 (smoothing). 만약 연관도를 구하려는 명사가 함께 주어진 용언-격조사와 한번도 함께 발생하지 않았고, 명사가

속하는 어떠한 개념에서도 어휘정보를 찾을 수 없다면, 용언과 격조사 연관도만을 가지고 연관도 값을 구한다.

연관도 값은 특정한 "용언, 명사, 격조사"가 연관된 정도를 나타내는 $\overline{Assoc}(v,n,j)$ — 술어, 명사, 격조사 연관도 — 와 용언과 특정한 격조사가 연관된 정도를 나타내는 확률인 $\overline{Assoc}(v,j)$ — 술어, 격조사 연관도 — 의 합으로 이루어진다. 변수 α 는 "술어, 명사, 격조사 연관도"와, "술어, 격조사 연관도" 를 더하는 비율을 결정하는데, 0.5보다 커야 한다. $\overline{Assoc}(v,n,j)$와 $\overline{Assoc}(v,j)$는 각각 다음과 같이 계산한다.

$$\begin{aligned}
\overline{Assoc}(v,n,j) &= \max(P(n,j|v), \frac{P(class(n),j|v)}{N}) \\
&\approx \frac{\max(f(v,n,j), \frac{P(v,class(n),j)}{N})}{f(v)} \qquad (when f(v) \neq 0) \quad (4.2) \\
\overline{Assoc}(v,j) &= P(j|v) \\
&\approx \frac{f(j,v)}{f(v)} \qquad (when f(v) \neq 0) \quad (4.3)
\end{aligned}$$

$\overline{Assoc}(v,n,j)$는 특정한 용언, 명사, 격조사가 관련이 있을 확률을 나타내는데, 개념으로 일반화된 어휘정보와 일반화되어 있지 않은 어휘 수준의 공기 정보 중, 빈도가 높은 정보를 이용하여 구한다. $class(n)$이란 명사 n이 포함된 개념을 의미하고, N은 개념 $class(n)$에 속한 명사들의 갯수를 말한다. 예를 들어, n이 "대문"이라면, $class(n)$은 시소러스에서 "대문"의 상위 개념인 "문", "구조물", "물건" 혹은 "글" 등이 될 수 있다. [그림 3.2] 그런데, 실제로 $Assoc()$값을 구할 때, 명사 n이 가지는 의미적 중의성때문에 $class(n)$을 결정하기가 힘든 경우가 생긴다. 이럴 때는 $P(class(n),j|v)$의 값을 최대로 해주는 $class(n)$을 올바른 n의 개념으로 선택해 준다.

$\overline{Assoc}(v,j)$는 주어진 용언과 격조사가 연관될 확률을 나타내는데, 이는 용언이 주어졌을 때, 격조사가 발생할 확률로 계산한다. 동사 "들어오"와 명사 "바람", 주격조사 "이" 사이의 어휘 연관도는 다음과 같이 계산한다.

$Assoc$(들어오, 바람, 이) $= \alpha \times \overline{Assoc}$(들어오, 바람, 이) $+ (\alpha - 1) \times \overline{Assoc}$(들어오, 이)

4.3 어휘 연관도의 적용

구조적 중의성의 해소를 위해서는 우선 가능한 모든 구문 분석 결과를 구한 후, 각각의 구문 분석 결과에 확률을 부여한다. 그리고 확률값이 최대가 되는 분석을 올바른 구문 구조라고 결정한다. 의존 관계 $relation_i$의 확률을 $P(relation_i)$라고 할 때, 파스 트리 T의 확률 $P(T)$는 식 (4.4)과 같다.

$$P(T) = \prod_i P(relation_i) \tag{4.4}$$

즉, 파스 트리의 확률은 파스 트리를 구성하는 모든 의존 관계의 확률의 곱이다. 그런데, 본논문에서 해결하려는 구조적 중의성 문제가 명사구의 지배소를 선택하는 문제이기 때문에, 명사구와 지배소 후보 사이의 의존 관계에만 확률을 부여해주고 다른 의존 관계에는 동일하게 1이라는 확률을 부여해도 명사구의 지배소를 선택할 수가 있다.

명사구와 지배소 후보 사이의 의존 관계에 확률을 부여할 때에는 연관도 함수값을 사용한다. 명사구를 구성하는 중심 명사 및 격조사와, 지배소의 중심 술어 사이의 연관도 값을 구하면 된다. 예를 들어 “창문으로 들어왔다.”라는 문장에서 어절 “창문으로”와 “들어왔다”사이의 의존관계가 갖는 확률 값은 $Assoc$(들어오, 창문, 으로)가 된다. 다음 예문을 살펴보자.

바람이 열린 창문으로 들어왔다.

위 예문에서 가능한 구문 분석 결과는 [그림 4.2]와 같고, 각각의 파스트리가 갖는 확률은 다음과 같이 구할 수 있다.

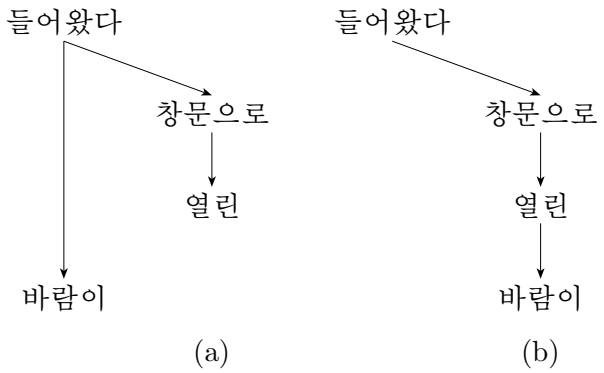


그림 4.2: 동일 문장에 대한 서로 다른 파스 트리

$P(T_a) = Assoc(\text{들어오, 창문, 으로}) \times 1 \times Assoc(\text{들어오, 바람, 이})$
$P(T_b) = Assoc(\text{들어오, 창문, 으로}) \times 1 \times Assoc(\text{열리, 바람, 이})$

예문의 "열린"과 "창문으로"사이의 의존 관계는 명사구와 지배 술어 사이의 관계가 아닌 수식 관계이므로, 확률값 1을 부여한다. 위 수식에서 $Assoc$(들어오, 창문, 으로)는 공통된 항목이므로, 결국 올바른 구문 트리를 결정하는 것은 "바람이"와 지배 술어 후보간의 연관도 값이 된다. α 값을 0.999로 주었을 때, 이 값들은 다음과 같이 구해진다.

$Assoc(\text{열리, 바람, 이}) = 0.999 \times 0 + 0.0001 \times 0.560 = 0.0005$
$Assoc(\text{들어오,바람, 이}) = 0.999 \times 0.03 + 0.0001 \times 0.389 = 0.0037$

계산된 결과 값은 "바람이"라는 어절이 동사 "열리" 보다 동사 "들어오"와 더 잘 어울린다는 것을 표현해주고 있다. 즉, 이는 어절 "바람이"의 지배술어를 "들어왔다"로 결정([그림 4.2]의 (a))하는 것이 더 바람직함을 나타낸다.

어휘 연관도 함수는 명사구의 지배소 결정 문제뿐만 아니라 미지격 결정등의 구조적 중의성 해소 작업에도 유용하게 사용될 수 있다.

제 5 장

실험 및 평가

이 장에서는 어휘정보가 구문 분석에서 발생하는 중의성 해소에 어느 정도 도움을 줄 수 있는지를 평가하고 검증해본다.

어휘정보를 추출한 코퍼스는 신문 기사, 잡지 기사, 소설, 설명문 등으로 구성된 약 800만 어절의 원시 코퍼스이며, 이 원시 코퍼스를 자동 품사 태거[김진동97]로 품사를 붙여준 후, 앞에서 설명한 방법으로 용언-명사-격조사 리스트를 생성하였다. 이렇게 얻은 용언-명사-격조사 쌍이 약 624,200 개의 분량이다.

일반화에 사용한 시소러스는 [조평옥97]에서 bottom-up 방식으로 생성한 시소러스로서, 12,833 개의 명사 항목으로 구성되어 있으며 최대 깊이는 17이다. 추출된 어휘정보의 명사를 일반화시킬 때는 빈도가 3 이상인 명사를 대상으로 했으며, 두 명사 사이의 상대적 거리가 0.85 이하일 때만 두 명사가 유사한 의미를 가진다고 고려하여 일반화를 시켰다. 결과 5000개 정도의 용언-개념-격조사 쌍이 추출되었다.

추출된 어휘정보는 의존 파서를 통하여 구문 분석 과정에 적용되었다. 이 의존 파서는 일반적으로 사용되는 지배소 후위 제약, 투영의 제약, 지배소 유일 제약, 일문 일표층격 제약 등이 가해졌으며, 연관도 함수 $Assoc()$을 이용하여 최적의 분석 결과 하나 만을 출력한다. α 값으로는 0.999를 사용하였다.

5.1 어휘정보의 평가

어휘정보를 구조적 중의성 해소에 사용하기 전에, 본 논문에서 제안한 어휘정보 구축 방법이 어느 정도 타당성이 있는지를 실험해 본다.

반자동으로 어휘정보를 구축한 [김선호96]에서는 미지격의 격결정 실험을 통하여 반자동으로 작성한 어휘정보의 유효성을 검증하였다. 미지격이란 술어와 명사구, 두 언어 요소 사이에 숨겨진 격관계를 말하는데, 다음과 같은 경우에 발생하게 된다.

- 격조사가 생략된 경우
  예 : 문 열었다. (명사 "문"과 동사 "열" 사이에 목적격 관계가 숨겨져 있다.)
- 보조사가 사용된 경우
  예 : 문은 열렸니? (명사 "문"과 동사 "열리" 사이에 주격 관계가 숨겨져 있다.)
- 술어가 관형형으로 사용되어 명사를 수식하는 경우
  예 : 어제 열린 학회 (명사 "학회"와 동사 "열리" 사이에 주격 관계가 숨겨져 있다.)

이런 미지격을 알아내는 작업은 어휘 정보가 없이는 해결이 어려운 분야이다. 본 논문에서는 미지격 결정 실험을 통하여 수작업으로 작성한 어휘정보와 본 논문에서 제안한 방법으로 작성한 어휘정보의 비교를 수행하고 제안한 방법의 타당성을 알아 본다.

미지격 결정을 위해서는 앞장에서 설명한 $Assoc()$ 함수를 사용한다. 미지격 관계인 명사와 술어가 모든 격조사와 어울릴 확률을 구하여 가장 적절한 격조사를 선택해주면 된다.

바람이 열린 창문으로 들어왔다.

에서는 관형형으로 사용된 술어 “열린”과 피수식어 “창문” 사이에 미지격 관계가 존재하는데, 이는 다음과 같이 미지격을 알아낼 수 있다.

$Assoc$(열리, 창문, 이) $= 0.999 \times 0.025 + 0.0001 \times 0.560 = 0.0261$
$Assoc$(열리, 창문, 을) $= 0.999 \times 0 + 0.0001 \times 0.0165 = 0.00001$
$Assoc$(열리, 창문, 에) $= 0.999 \times 0 + 0.0001 \times 0.0884 = 0.00008$
$Assoc$(열리, 창문, 로) $= 0.999 \times 0 + 0.0001 \times 0.101 = 0.0001$
$Assoc$(열리, 창문, 서) $= 0.999 \times 0 + 0.0001 \times 0.232 = 0.0002$

고려하는 모든 격조사에 대해서 $Assoc$(열리,창문,격조사) 값을 구한 후, 최대 어휘 연관도 값을 가지는 격조사를 선택하면 “열린”과 “창으로”의 미지격 관계가 주격으로 결정됨을 알 수 있다. 같은 방법으로 격조사가 생략된 경우나, 보조사가 사용된 경우에도 미지격을 결정해 줄 수 있다.

미지격 결정 실험을 위해서 [김선호96]의 실험에서 사용된 동사를 포함한 문장 50개씩을 학습 및 실험 코퍼스에서 골랐다. 총 500개의 문장에서 나온 미지격을 포함한 의존 관계는 534개였다. [표 5.1]는 실험 결과이다.

표에서 보듯이 동사에 따라서 [김선호96]의 실험 결과와 어느 정도의 차이가 있지만, 전체 평균 정확도로 봤을 때, 본 논문의 방법으로 자동 생성한 어휘정보를 사용한 경우나, 사람이 개입하여 작성한 [김선호96]의 어휘정보나, 미지격의 결정을 하는 데 큰 차이가 없음을 알 수 있다. 실험 대상 문장이 [김선호96]의 실험에서 사용한 문장과 동일하지 않기때문에 단순 정확도 비교만을 할 수는 없지만, 어휘정보를 작성하는 데 들어가는 엄청난 수작업의 양을 고려할 때, 본 실험은 본 논문에서 제안하는 자동화된 방법의 유효성을 보여준다.

5.2 명사구의 지배소 선택 실험

구조적 중의성 해소에서 술어-인자 어휘정보가 사용될 수 있는 부분은 명사구의 지

동사	[김선호96]의 정확도	본 논문의 정확도
내리	81.81%	83.01%
만들	85.71%	81.48%
먹	85.14%	90.38%
받	86.43%	85.96%
보내	90.90%	86.79%
쓰	91.37%	89.65%
앉	88.24%	78.84%
열/열리	77.27%	90.00%
짓	77.78%	96.22%
타	88.89%	79.24%
평균	86.26%	86.16%

표 5.1: 미지격 결정 실험결과

	평균 중의성 정도	어휘공기정보만 사용한 경우의 정확도	어휘공기정보와 개념정보를 함께 사용한 경우의 정확도
학습코퍼스	2.52개	90.1%	90.1%
실험코퍼스	2.49개	85.9%	91.5%

표 5.2: 명사구의 지배소 선택 실험 결과

배소를 선택하는 부분이다. 이 문제를 해결하기 위해 관련 연구에서는 코퍼스로부터 어휘 공기 정보를 추출하여 사용하였다[이공주97]. 본 실험에서는 개념으로 일반화된 어휘정보를 이용하여 명사구의 지배소를 선택할 때의 정확도와 어휘 공기 정보만을 이용하여 명사구의 지배소를 선택할 때의 정확도를 비교하여 본 논문에서 제안한 방법이 우수함을 보인다.

실험 대상 100 문장은 어휘정보를 추출한 학습 코퍼스에서 뽑은 문장 50개, 학습 코퍼스 이외의 신문 기사, 초등학교 교과서, 소설 코퍼스에서 뽑은 문장 50개이다. 각 문장은 "명사+서술격조사" 형태의 지정사를 술어로 포함하고 있지 않다. 자동 추출한 어휘 정보에 지정사에 관한 정보는 포함되어 있지 않기때문에, 지정사가 포함된 문장의 경우 제대로 명사구의 지배소를 선택하는 것이 불가능하기 때문이다. 학습코퍼스의 평균 문장 길이는 8.54 어절이며, 최대 문장 길이는 17 어절이고, 최소 문장 길이는 4 어절 이다. 실험코퍼스의 평균 문장 길이는 8.52 어절이며, 최대 문장 길이는 16 어절이고, 최소 문장의 길이는 4 어절이다. "명사+격조사" 어절이 결합할 수 있는 지배소 후보가 단 하나인 경우에는 중의성이 존재하지 않으므로, 이 경우는 정확도 계산에서 제외했다. 이렇게 했을 때 학습 코퍼스와 실험 코퍼스에 각각 71개의 중의성을 갖는 "명사+격조사" 어절이 남았다. 의존 파서를 통해 구조적 중의성 해소 실험을 수행한 결과, 정확도는 [표 5.2]와 같았다. 평균 중의성 정도는 각 명사구가 가지는 평균 지배소 후보의 갯수를 말한다.

표에서 보듯이 자동 추출한 어휘정보를 학습 코퍼스에 적용한 경우, 개념으로 일반화된 어휘정보를 어휘 공기 정보와 함께 사용한 경우나 어휘 공기 정보를 그

대로 사용한 경우나 중의성 해소율에 차이가 없었으나, 실험 코퍼스에 본 논문에서 제안한 방법을 적용했을 때, 약 6%의 정확도가 향상되었다.

명사구의 지배소 선택을 제대로 하지 못한 경우들을 살펴보자.

- 음악을 가르치는 선생님으로서 뛰어난 업적을 남기셨습니다.[1] :

  $Assoc$(뛰어나,선생님,으로) $= 0.999 \times 0 + 0.001 \times 0.1142 = 0.00011$

  $Assoc$(남기,선생님,으로) $= 0.999 \times 0 + 0.001 \times 0.0936 = 0.00009$

  어절 "선생님으로서"는 "뛰어난"과 "남기셨습니다"의 두 지배소 후보를 가진다. 그런데 $\overline{Assoc}$(뛰어나, 선생, 으로)와 $\overline{Assoc}$(남기, 선생,으로)가 모두 0의 값을 가지므로, 술어와 격조사 사이의 연관도 값만을 가지고 지배소를 결정하게 된다. $\overline{Assoc}$(뛰어나, 으로) < $\overline{Assoc}$(남기, 으로)이므로, 명사구 "선생님으로서"의 지배소가 "남기셨습니다"로 잘못 결정된다. 두 술어가 비슷한 격조사를 요구해서 적은 확률의 차이로 지배 술어가 잘못 결정되는 예이다.

- 또한 작업 도중에 중요한 팁이나 사용법 등에 대한 정보가 제공되기도 한다.[2] :

  $Assoc$(중요하,도중,에) $= 0.999 \times 0 + 0.001 \times 0.1528 = 0.00015$

  $Assoc$(대하,도중,에) $= 0.999 \times 0.0009 + 0.001 \times 0.8585 = 0.00140$

  $Assoc$(제공되,도중,에) $= 0.999 \times 0 + 0.001 \times 0.1111 = 0.00011$

  $Assoc$(하,도중,에) $= 0.999 \times 0.0001 + 0.001 \times 0.0617 = 0.00020$

  어절 "도중에"는 "중요한", "대한", "제공되기도", "한다"를 지배소 후보로 갖

[1] 음악_NNCG+을_PO 가르치_VV+는_EFD 선생님_NNCG+으로서_PA 뛰어나_VV+ㄴ_EFD 업적_NNCG+을_PO 남기_VV+셨_EP+습니다_EFF+._SS.

[2] 또한_AA 작업_NNCV 도중_NNCG+에_PA 중요_NNCJ+하_XSVJ+ㄴ_EFD 팁_NNCG+이나_PN 사용법_NNCG 등_NNB+에_PA 대하_VV+ㄴ_EFD 정보_NNCG+가_PS 제공_NNCV+되_XSVV+기_EFN+도_PX 하_VV+ㄴ다_EFF+._SS.

는다. “도중에”와 어휘 연관도가 가장 높은 어절은 “대한”이지만, 어절 “등에”가 “대한”과 같는 연관도 값이 더 높기 때문에($Assoc$(대하, 등, 에) $= 0.0153 <$ $Assoc$(대하, 도중, 에) $= 0,001$) 중복격 금지 원칙에 의해 “도중에”의 지배소는 “한다”로 결정된다. 이 경우 “도중에”는 시간 부사절인데, 모든 동사와 형용사를 수식할 수 있기때문에 통계 정보만으로 시간 부사절의 지배소를 결정하는 것은 문제가 있다는 것을 보여주는 예이다.

- 대구의 한낮 기온이 올해 들어 가장 높은 35.3도를 기록하는 등 남부지방 역시 폭염에 시달렸다. [3] :

  $Assoc$(들,기온,이) $= 0.999 \times 0 + 0.001 \times 0.343 = 0.00034$

  $Assoc$(높,기온,이) $= 0.999 \times 0.005 + 0.001 \times 0.651 = 0.00613$

  $Assoc$(기록하,기온,이) $= 0.999 \times 0 + 0.001 \times 0.083 = 0.00008$

  $Assoc$(시달리,기온,이) $= 0.999 \times 0 + 0.001 \times 0.071 = 0.00007$

  파서는 어절 “기온이”의 지배소를 “기록하는” 대신에 “높은”으로 잘못 선택했다. 실제로 “기온이”가 “높다”라는 술어와 자주 사용되지만, “기온이 x도를 기록하다.”라는 문형도 존재한다. 어휘 정보를 추출할 때, 구문 분석된 코퍼스를 이용하지 않고, 원시 코퍼스만을 가지고 정확히 의존 관계를 갖는 어휘들만을 추출하기 때문에 이와 같은 형태의 문형 정보는 얻기 힘들다. 개개의 어휘간 확률만을 가지고는 해결하기 어려운 형태의 예이다.

위와 같은 경우들을 살펴 볼 때, 어휘정보가 구조적 중의성의 해소에 도움이 된다는 것은 분명하지만, 어휘정보만을 가지고는 해결하지 못하는 구조적 중의성 문제가 존재한다는 것도 알 수 있다. 이를 해결하기 위해서는 확률 어휘정보를 적용

[3] 대구_NNP+의_PD 한낮_NNCG 기온_NNCG+이_PS 올해_NNCG 들_VV+어_EFC 가장_AA 높_VJ+은_EFD 35_SCD+._SS.+3_SCD+도_NNBU 를_PO 기록_NNCV+하_XSVV+는_EFD 등_NNB 남부_NNCG+지방_NNCG 역시_AA 폭염_NNCG+에_PA 시달리_VV+었_EP+다_EFF+._SS.

할 때에 연관도 값의 차이가 일정 크기 이상인 경우에만 구조적 중의성 해소를 수행하고, 나머지 경우에는 다른 정보를 이용하여 구조적 중의성의 해소를 시도하는 것이 바람직하게 보인다.

또한 확률 어휘정보를 확률 문법과 함께 사용하는 것이 바람직하다고 생각된다. 확률 문법은 선호되는 구문 구조를 반영하기때문에 지배소 후보 술어들이 비슷한 개념의 인자인 명사구를 요구할 때도, 명사구가 올바른 지배소 술어를 선택하는데 도움이 될 것이다.

본 실험에서 사용한 시소러스의 문제점때문에 오류가 발생하기도 했다. 예를 들어 '사람'이라는 개념이 '물건'의 하위 개념으로 분류되어 있기때문에 일반화 과정에서 오류를 유발했다. 시소러스의 정련이 요구된다.

본 논문에서는 동사와 형용사만을 술어의 대상으로 하여 어휘정보를 작성하였으나, 실제 많은 문장들에서 명사+서술격조사 형태의 술어가 사용되고 있다. 이러한 문장들을 확률적으로 해결하기 위해서는 명사+서술격조사 형태의 술어에 대한 어휘정보의 작성이 요구된다.

제 6 장

결 론

대량의 코퍼스를 이용하여 구조적 중의성을 해소하는 연구들이 있어왔다. 그러나 일반화 된 어휘정보를 사용하지 못하고 어휘 공기 정보만을 사용하여 구조적 중의성 해소에 적용시켜 왔다.

본 논문에서는 이러한 문제점을 해결하기 위해서 원시 코퍼스와 시소러스를 이용하여 자동으로 어휘정보를 구축하여 구조적 중의성을 해소하는 방법을 소개하였다. 원시 코퍼스를 자동 태거로 품사 태깅하여 단문으로 분리, 정확한 용언-명사-격조사 패턴을 추출하고, 자동으로 명사의 의미를 결정하고 일반화시켜 용언-개념-격조사로 이루어진 어휘정보를 구축하는 방법을 제안하였고, 이를 구조적 중의성 해소에 적용하는 방법을 제안했다.

약 800만 어절 크기의 원시 코퍼스로부터 624,200개 가량의 공기된 용언-명사-격조사 어휘를 추출하였다. 이 어휘 공기 정보 중 명사를 12,833개의 명사 항목으로 구성된 시소러스를 이용하여 일반화 시켜 약 5,000개 정도의 용언-개념-격조사 쌍을 자동으로 구축하였다. 자동 구축된 어휘정보를 명사구의 지배소 선택 실험에 적용한 결과, 실험 코퍼스에서 약 91.5%의 명사구 지배소 선택 정확도를 얻을 수 있었다.

실험 과정을 통하여 확률 어휘정보를 이용한 구조적 중의성 해소의 문제점도 들

어났다. 명사구의 지배소 결정시, 비슷한 인자를 요구하는 술어들이 지배소 후보로 있는 경우 오류가 발생하기도 했다. 또한 자료 부족 문제를 해결하기 위하여 사용하는 술어-격조사 연관도가 오류를 유발하기도 했다. 이런 문제를 해결하기 위해서는 후보들 사이의 연관도 값의 차가 일정 크기 이상인 경우에만 구조적 중의성 해소를 수행하고, 나머지 경우에는 다른 정보를 이용하여 구조적 중의성의 해소를 시도하는 것이 바람직하게 보인다.

한국어에서 술어의 역할이 매우 중요하기 때문에 술어-인자에 대한 어휘정보는 구문 분석 과정에서 유용하게 사용될 수 있을 것이다. 또 본 논문에서 사용한 의존문법에서 뿐만 아니라, 구구조 문법, 어휘 기반 문법에까지 본 논문에서 제안한 방법이 응용될 수 있을 것이다.

향후 연구로는 본 논문에서 사용한 확률 어휘정보와 함께 다른 형태의 정보도 함께 사용하여 구문 분석 단계에서의 중의성 해소를 시도할 계획이다.

참고 문헌

[김선호96] 김선호, “통계 정보를 기반으로 한 어휘 관계 예측”, 연세대학교 대학원 컴퓨터학과, 석사 학위 논문, 1996.

[김진동97] 김진동, 임희석, 임해창, “Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델”, *한국정보과학회논문지*, Vol. 24, No. 12, pp. 15-2–1512, 1997.

[나동렬94] 나동렬, “한국어 파싱에 대한 고찰”, *한국정보과학회지*, Vol. 12, No. 8, pp. 33-46, 1994.

[류법모97] 류법모, 장명길, 박수준, 박재득, 박동인, “구문구조부착 말뭉치를 이용한 술어의 하위범주화 정보 구축”, *제 9회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 116–121, 1997.

[양단희98] 양단희, 송만석, “말뭉치로부터 격틀 구축에 필요한 학습 데이터 추출”, *제10회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 287-292, 1998.

[양재형94] 양재형, “통계 정보를 활용한 한국어 미지격 명사구의 문법기능 결정”, *한국정보과학회논문지*, Vol. 21, No. 5, 1994.

[엄미현96a] 엄미현, 신대규, 나동렬, “한국어의 구조적인 애매성”, *한국정보과학회 봄 학술발표 논문집*, Vol. 23, No. 1, pp. 911 – 914, 1996.

[엄미현96b] 엄미현, 신대규, 임병준, 나동렬, “다중파스 여과에 기반한 한국어의 구조적 중의성 해소”, *제 8회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 443 – 451, 1996.

[윤준태97] 윤준태, “공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석”, 연세대학교 대학원 컴퓨터학과, 박사 학위 논문, 1997.

[윤준태98] 윤준태, 최기선, 김선호, 송만석 “구문 분석에서의 어휘간 공기 정보의 활용”, *제10회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 276 – 280, 1998.

[이공주97] 이공주, 김재훈, 김길창, “중심어간의 공기 정보와 구문 규칙을 기반으로 한 확률적 한국어 구문 분석”, *제 9회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 332-338, 1997.

[이휘붕98] 이휘붕, 강인수, 이종혁, “개념패턴과 통계정보를 이용한 한국어 미지격의 구문관계 결정방법”, *제10회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 261 – 266, 1998.

[장명길97] 장명길, 류법모, 박재득, 박동인, 맹성현, “통계/의미 정보를 이용한 한국어 의존 파싱”, *제 9회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 313–319, 1997.

[조규빈95] 조규빈, “하이라이트 고교문법 자습서”, *지학사*, 1995.

[조평옥97] 조평옥, 옥철형, “한국어 명사 의미 계층 구조 구축”, *제 9회 한글 및 한국어 정보처리 학술발표 논문집* , pp. 129-135, 1997.

[Brent91] M. Brent, “Automatic acquisition of subcategorization frames from untagged text”, In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991.

[Briscoe97] T. Briscoe, J. Carroll, "Automatic extraction of subcategorization from corpora", In *Proceedings of the Conference on Applied Natural Language Processing*, 1997.

[Collins96] M. Collins, "A New Statistical Parser Based on Bigram Lexical Dependencies", In *Proceedings of the 34th Annual Meeting of the Asssciation for Computiotional Linguistics*, pp. 79-86, 1996.

[Hindle93] D. Hindle, M. Rooth, "Structural Ambiguity and Lexical Relations", *Computational Linguistics*, 1993.

[Li95] H. Li, N. Abe, "Generalizing case frames using a thesaurus and the MDL principle", In *Proceedings of Recent Advances in Natural Language Processing*, pp. 239–248, 1995.

[Manning93] C. Manning, "Automatic acquisition of a large subcategorization dictioanry from corpora", In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 235-242, 1993.

[Marcus93] M. P. Marcus, B. Santorini, M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*, Vol. 19, No. 2, 1993.

[Resnik93] P. Resnik, "Selection and Information: A Class-based Approach to Lexical Relationships", University of Pennsylvannia, PhD Thesis, 1993.

[Ushioda93] A. Ushioda, D. Evans, T. Gibson, A. Waibel, "The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora", In *Proceeding of ACL Workshop on the Acquisition of Lexical Knowledge from Text*, 1993.

부록 A

품사태그 리스트

- NNCG :체언(N):명사(N):보통(C):일반(G)
- NNCV :체언(N):명사(N):보통(C):동사(V)
- NNCJ :체언(N):명사(N):보통(C):형용사(J)
- NNB :체언(N):명사(N):의존(B)
- NNBU :체언(N):명사(N):의존(B):단위(U)
- NNP :체언(N):명사(N):고유(P)
- NPP :체언(N):대명사(P):인칭(P)
- NPI :체언(N):대명사(P):지시(I)
- NU :체언(N):수사(U)
- XSNN :접사(X):접미(S):체언(N):명사(N)
- XSNND :접사(X):접미(S):체언(N):명사(N):관형(D)-적
- XSNP :접사(X):접미(S):체언(N):대명사(P)

- XSNU :접사(X):접미(S):체언(N):수사(U)
- XSNPL :접사(X):접미(S):체언(N):복수(PL)-들
- XPNN :접사(X):접두(P):체언(N):명사(N)
- XPNU :접사(X):접두(P):체언(N):수사(U)
- PS :조사(P):주격(S)
- PC :조사(P):보격(C)
- PO :조사(P):목적격(O)
- PD :조사(P):관형격(D)
- PA :조사(P):부사격(A)
- PV :조사(P):호격(V)
- PN :조사(P):접속(N)
- PX :조사(P):보조(X)
- DA :관형사(D):성상(A)
- DI :관형사(D):지시(I)
- DU :관형사(D):수(U)
- XSD :접사(X):접미(S):관형사(D)-적
- AA :부사(A):성상(A)
- AP :부사(A):서술(P)
- AI :부사(A):지시(I)

- AC :부사(A):접속(C)
- AV :부사(A):동사(V)
- AJ :부사(A):형용사(J)
- XSA :접사(X):접미(S):부사(A)
- XSAH :접사(X):접미(S):부사(A):히(H)-히
- C :감탄사(C)
- I :서술격조사(I)
- VV :용언(V):동사(V)
- VVX :용언(V):동사(V):보조(X)
- VJ :용언(V):형용사(J)
- VJX :용언(V):형용사(J):보조(X)
- XSVV :접사(X):접미(S):용언화(V):동사(V)
- XSVJ :접사(X):접미(S):용언화(V):형동사(J)
- XSVJD :접사(X):접미(S):용언화(V):형동사(J):답(D)-답
- XSVJB :접사(X):접미(S):용언화(V):형동사(J):기타(B)-롭,스럽
- EFF :어미(E):어말(F):종결(F)
- EFC :어미(E):어말(F):연결(C)
- EFN :어미(E):어말(F):명사(N)
- EFD :어미(E):어말(F):관형(D)

- EFA :어미(E):어말(F):부사(A)
- EP :어미(E):선어말(P)
- NN? :체언(N):명사(N):추정(?)
- V? :용언(V):추정(?)
- SS. :기호(S):문장(S):온점(.)
- SS? :기호(S):문장(S):물음표(?)
- SS! :기호(S):문장(S):느낌표(!)
- SS, :기호(S):문장(S):반점(,)
- SS/ :기호(S):문장(S):빗금(/)
- SS: :기호(S):문장(S):쌍점(:)
- SS; :기호(S):문장(S):반쌍점(;)
- SS‘ :기호(S):문장(S):왼쪽따옴표(‘)
- SS’ :기호(S):문장(S):오른쪽따옴표(’)
- SS(:기호(S):문장(S):왼쪽괄호(()
- SS) :기호(S):문장(S):오른쪽괄호())
- SS- :기호(S):문장(S):줄표(-)
- SSA :기호(S):문장(S):줄임표(A)
- SSX :기호(S):문장(S):기타(X)
- SCF :기호(S):문자(C):외국(F)-ASCII

- SCH :기호(S):문자(C):한자(H)
- SCD :기호(S):문자(C):숫자(D)
- SPACE :공백(SPACE)