Maximum Entropy 모델을 이용한 나열 및 병렬형 인식

임수종[1],이창기[2], 허정[3], 장명길[4]
한국전자통신연구원 음성/언어정보연구센터[1234]
{isj[1],leeck[2],jeonghur[3], mgjang[4]}@etri.re.kr

Syntax Analysis of Enumeration type and Parallel Type Using Maximum Entropy Model

Soojong Lim[1], Changki Lee[2], Jeong Hur[3], Myoung-Gil Jang[4]
Speech/Language Information Research Center ETRI[1234]

요 약

한국어 문장을 구조 분석할 때에 모호성을 발생시키는 유형 중의 하나가 나열 및 병렬형이다. 문장 구조 복잡도를 증가시키는 나열 및 병렬형을 구조 분석 전에 미리 하나의 단위로 묶어서 처리하는 것이 문장 구조 분석의 정확도를 높이는데 중요하다.

본 연구에서는 형태소 태그를 이용한 기본 규칙으로 문장을 청크 단위로 분할하고 분할된 청크 중에서 나열형을 인식하여 해당되는 청크들을 하나의 나열 청크로 통합하여 청크의 개수를 줄인다. 병렬형에 대해서는 반복되는 병렬 청크의 범위와 생략된 용언을 복원한다.

이러한 인식은 첫 단계로 기호(symbol)를 중심으로 구축된 간단한 규칙으로 인식을 하고 이러한 규칙에 해당되지 않는 형태의 나열 및 병렬형은 Maximum Entropy 모델을 이용하여 적용한다.

ME 모델은 어휘자질, 형태소 품사 자질, 거리 자질, 의미자질, 구 단위 태그 자질(NP:명사구, VP:동사구, AP:형용사구), BIO 태그(Begin, Inside, Outside) 자질에 대한 ME(Maximum Entropy) 모델을 이용하여 구축되었다.

Keyword : 나열, 병렬, 문장구조 분석, 규칙 , Maximum Entropy 모델

1. 서 론

한국어 문장을 분석하기 위한 과정 중에서 문장의 구조와 문법적인 기능을 분석하는 완전한 구문분석(full parsing)은 성능 문제로 인해 실제 시스템에 널리 적용되지 못 하고 있다. 그러나 문장구조 정보는 한국어 문서를 다루는 번역, 정보추출, 검색 등의 응용 프로그램에서 중요한 정보이기 때문에 구묶음(chunk) 개념을 전처리 단계로 도입하여 완전 구문 분석이 가지는 제약을 넘어서려는 연구가 진행되었다[5, 6].

구묶음의 개념은 [11]에서 소개되었는데, 단어들의 의미있는 연속체로서 하나의 내용어와 몇 개의 인접한 기능어들로 이루어진 것으로 정의된다.

부분 어순 자유 언어인 한국어 문장을 구문분석할 경우 주로 의존문법을 많이 이용하는데 이 경우 너무 많은 의존 관계가 발생하는 것을 방지하기 위해서 앞에서 소개한 구묶음의 개념을 도입하여 성능향상을 보였다[1, 5]. 한국어에서는 조사와 어미가 명사구와 동사구 묶음에서 각각 중요한 역할을 하기 때문에 이런 정보를 사용하는 몇 개의 간단한 규칙만으로도 한국어 구묶음의 성능은 기계학습 기법이나 통계기반 방법등과 같은 추론 모델과 거의 비슷하게 될 수 있다[1, 4]. 그러나 이러한 접근 방법의 경우 구묶음이 잘못 묶일 경우에는 의존 문법을 적용하더라도 오류를 복구할 수 있는 방법이 없기 때문에 정확한 분석이 필요하다.

한국어 백과사전 질의응답 시스템의 일부로 구묶음을 이용한 의존 파서[7, 12]를 개발하면서 생기는 오류는 구묶음의 경우 대부분 긴 문장이 차지하고 있으며 이러한 긴 문장의 대부분은 백과사전 문서의 특성상 나열형과 병렬형으로 구성이 된다. 종래에는 나열형과 병렬형에 대한 고려 없이 긴 문장을 단문 단위로 나눠서 분석을 하는데 이러한 방법은 단지 여러 개의 절로 구성된 복문에서는 효과를 발휘하지만 나열형이나 병렬형으로 구성된 긴 문장에서는 단문 단위로 분할이 되지 않기 때문에 모호성을 감소시킬 수 없다.

본 논문에서는 그동안 알려진 규칙에 기반한 한국어 구묶음 기법을 이용하여 구묶음을 수행하고 나열형과 병렬형을 정확히 묶어 주기 위해서 간단한 규칙과 최대 엔트로피 모델을 적용하고자 한다.

2. 관 련 연 구

한국어 구묶음에 대한 연구는 규칙을 사용하는 방법과 기계학습을 사용하는 방법 두 가지로 분류할 수 있다.

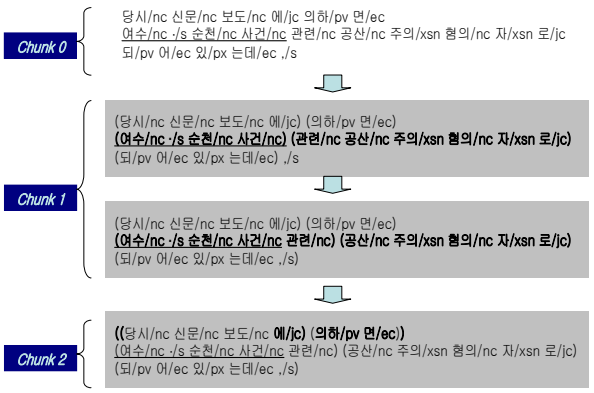
규칙을 사용하는 방법으로는 정규식으로 명사구를 묶고 장벽 알고리즘을 사용하여 동사구를 묶는 접근 방법[1, 4], 구문패턴 사전과 간단한 규칙을 사용한 방법[8], 세 단계로 구묶음을 수행하는 방법[10]이 있다. [10]은 먼저 형태소 특성을 이용하여 기본적인 구를 묶은 후, 문맥의존문법을 이용하여 명사구를 인식하고 마지막 단계로 말뭉치에서 추출한 연어 패턴을 사용하여 동사구를 인식하였다. 이런 방법들은 주로 명사구와 동사구만을 인식하기 위해 적용되었고 규칙만 사용하였기 때문에 모든 언어 현상을 모두 표현하기가 힘든 단점이 있다.

기계 학습을 사용하는 방법으로는 결정트리 학습과 최대 엔트로피 모델을 결합하는 방법[2]이 있고 규칙과 통계를 결합하여 각각의 단점을 보완하고자 하는 접근 방법[3]이 있다.

3. 기 본 구 묶 음

구묶음 인식은 개체명 인식 결과와 규칙 기반 접근 방법으로 문장을 구묶음 단위로 분리하는 방식으로 진행된다. 구묶음 인식 과정은 모두 3 단계로 구분할 수 있다. 개체명 인식 결과를 Chunk0 이라 하고, 형식형태소 단위별 구묶음 분리인 BaseNP, BaseVP, BaseADP, BaseIP 를 찾는 과정과 이들의 예외사항을 처리한 후의 구묶음 인식 결과를 Chunk1 이라 한다. 기존 구묶음 인식 연구는 대부분 Chunk1 을 대상으로 한다. 이후 추출된 구묶음을 규칙기반 접근방법을 취하여 속격 NP 와 어미 확장 사전을 통한 VP 확장인식 후의 결과를 Chunk2 결과라 하였다.

구묶음 인식 예제는 <그림 1>과 같다



[그림 1] 구묶음 인식 예제

기본 구묶음 인식 방법은 어절 단위 및 형식형태소를 기준으로 하여 진행한다.우선, 문장을 구성하는 청크들의 내부 형태소 I 와 외부 형태소 O 를 구분하는 작업을 진행한다.

- 내부 형태소 목록: nc, np, mm, nn, pv, pa, mag, maj, ii, xsn, xp
- 외부 형태소 목록: jm, s, nb, jc, jx, jj, etm, px, co, ef, ep, xsv, xsm, ec, etn

이후, 외부로부터 로딩된 기본 청크 예외 규칙을 이용하여 청크의 I, O 태그를 재조정하고 B 태그를 할당한다. 여기서 사용되는 단서어휘 목록은 아래와 같다.

[표 1] 단서어휘 목록

단서어휘	설명
‘ ’ ' 《 》 · () 〈 〉 “ ”	심볼은 CHK_O 옆 단서 어휘는 CHK_I 로 할당
후 뒤 때 다음 이후 직후 가운데 전 이상	한 어절의 마지막 형태소일 경우 CHK_O 할당
소속	현재 형태소가 prev 를 AT 로 갖고 word 어휘와 일치할때 CHK_O
년 월 일 대 것 해 번째	CHK_I 할당
언제 매년 스스로 각각 이로써	현재 형태소가 mag 나 maj 이고 아래 목록에 속하면 CHK_I
어떻 같	prev 가 null 이 아니고 prev 가 m_etm_prev_list 에 존재할 때 CHK_I
제일	prev 가 mag 현재 어휘는 mag 가 아닐 때 현재 어휘가 m_mag_modify_list 에 있을 경우, CHK_B
처음 결과 이듬해 역시 오랫동안	prev 가 null 이 아니고 prev 와 현재 어휘가 같은 word 가 아닐 때 CHK_B 할당
차	CHK_O 할당
이후 후 이상 밑 때	연속해서 포함 어휘들은 청크에 포함

기본 청크 확장은 어미 공기사전과 조사공기사전을 이용하여 기본 청크로 결정된 청크들을 확장하는 역할을 한다. 즉, 어미와 조사는 일부 공기하는 어휘를 포함하여 하나의 어미와 조사로 간주하며, 이들 목록을 미리 준비하여 청킹에 활용한다. 이들 목록 역시 외부 규칙으로 독립하여 청킹 블록이 로딩할 수 있게 처리한다.

4. 나 열 및 병 렬 구 묶 음

기본 청크 확장된 결과에 대해서 나열 및 병렬형을 처리하고자 한다.

4-1 나열 및 병렬 정의

한국어 문장에서 나열형이란 한 단위의 체언류가 공백, 컴마(,) 가운데점(·)과 같은 기호로 구분되어 동격으로 죽 벌여놓은 부분을 포함하는 것으로 다음과 같은 문장에서 밑줄 친 부분을 말한다.

산관(算官) · 율관(律官) · 화원(畵員) 등은 하급기술관이었고, 악생(樂生) · 악공(樂工) · 상도(尙道) · 화사(畵史) · 공제(工製) · 재부(宰夫) · 선부(膳夫) 등은 잡직기술관이었다.

기계의 출력, 회전수, 단위중량당 출력, 신뢰도 등 개개 지표마다 비교하는 것은 어렵지 않다.

본 연구에서 목표로 하는 나열형 패턴은 다음의 표 2 와 같다.

표 [2] 나열형 패턴

패턴	예문
~와 ~	(1)야구와 테니스 (2)영국과 프랑스
~와 ~ 그리고 ~	(1)후손 둘리, 집주인 고길동, 외조카 희동이, 악동들인 도우너와 또치 그리고 이웃집 가수 지망생 마이콜
~·~	(1)리틀엔젤스회관·리틀엔젤스예술단 (2)청룡(靑龍:東) ·주작(朱雀:南) ·백호(白虎:西) ·현무(玄武:北)의 4 가지
~ 및 ~	(1)향찰(鄕札) 및 이두(吏讀)
혼합형 (~와/과 및, 도트의 혼합)	(1)태평양 ·대서양 및 인도양 배우 ·무대 ·관객, 그리고 희곡의 4 가지를 든다.
~,~	(1)한국은 폴란드(38 위), 미국(13 위), 포르투갈(5 위)과 함께 D 조에 (2)보라돌이, 뚜비, 나나, 뽀(실제이름은 팅키윙키, 딥시, 라라, 포) 등 4 명

한국어 문장에서 병렬형이란 두 단위 이상의 체언류가 공백, 컴마(,) 가운데점(·)과 같은 기호로 구분되어 동격으로 죽 벌여놓은 부분을 포함하는 것으로 아래 문장에서 밑줄과 괄호로 병렬형을 표시하였다.

나열형과 병렬형의 다른 점은 나열형은 기본적으로 체언류를 죽 늘어놓은 형태이지만 병렬형의 경우에는 격조사를 포함하고 있고 단순히 체언류를 늘어놓은 것이 아닌 아래의 예처럼 용언이 생략됐다는 점이다.

이러한 노력의 결실로 (제 16 회전에서는 《고담(古談)》 으로 최고상,)*을 수상하였고* (17

회전에서는 《하일(夏日)》로 조선총독상을) 수상하였다.

이러한 나열형과 병렬형은 신문기사를 중심으로 하여 많은 문서에서 사용되고 있지만 자연언어처리 응용 시스템에서 필수적으로 거쳐야 하는 구문 분석은 이러한 형태의 문장에서 많은 문제점을 안고 있다.

4-2 규칙에 의한 구 묶음

규칙은 아래와 같은 형태로 형태소 태그와 기호를 사용한다. 아래는 사용된 규칙의 한 예이다.

{n+sym}$^{+}${nc+j}
n = {자립명사(nc), 의존명사(nb), 대명사(np)}
j = {격조사, 보조사, 접속조사}
sym = { ‘,’ , ‘.’ , ‘ ‘}
‘+’ : 한번 이상 반복
‘+’ : 앞과 뒤가 같은 어절. 공백으로 분리되지 말아야 함.

위와 같은 규칙을 사용하여 대상으로 판단이 됐을 경우에는 인식용 통계 정보를 이용하여 나열형인지 병렬형인지 판단한다.

4-3 최대 엔트로피를 이용한 구 묶음

최대 엔트로피(Maximum Entropy, ME) 모델[9]은 주어진 제약 조건을 만족하는 여러 확률 분포 중에서 가장 균일한 분포 상태를 가지는 모델이다. 바꾸어 말하면, ME 모델은 주어진 제약 조건 하에서 최대 엔트로피를 가지는 확률 분포를 가지고 있다. 이를 수식으로 나타내면 다음과 같다.

$P = \{ \text{models consistent with constraints} \}$
$H(p) = \text{Entropy of } p,\ p \in P$
$P_{ME} = \text{argmax}_{p \in P} H(p)$

여기서 P_{ME} 가 최대 엔트로피 확률 분포를 가지는 모델이다.

ME 모델의 매개변수 추정에 사용되는 알고리즘에는 Generalized Iterative Scaling(GIS), Improved Iterative Scaling(IIS), 그리고 Limited Memory BFGS(L-BFGS) 등 잘 알려진 것이 몇가지 있다. 본 연구에서는 GIS 알고리즘을 사용하였다.

ME 모델의 가장 두드러진 특징은 모델의 특성을 완전히 드러내는 후보 자질들을 선택해 주기만 하면 되는데 다음에 자질에 대한 확률값으로 구성되어 있다.

어휘자질, 형태소 품사 자질, 거리 자질, 의미자질, 구 단위 태그 자질(NP:명사구, VP:동사구, AP:형용사구), BIO 태그(Begin, Inside, Outside) 자질을 사용하여 Maximum Entropy(ME) 모델을 이용하여 구축되었다. 원래 ME 모델은 이진 결정을 내리는 경우에 더 적합하지만 이진 결정과 함께 범위를 결정하기 위해서 BIO 태그를 자질로 포함하여 문제점을 보완하고자 하였다. 수작업으로 학습 데이터를 구축하여 나열형과 병렬형에 대한 자질에 대한 확률값을 미리 구해 놓고 문장에서 앞에서 언급한 자질을 통계 모델에 넣어서 나열이나 병렬 값을 0 과 1 사이의 확률값으로 결과를 얻을 수 있다. 그리고 각각의 구단위에 대해서 나열형의 시작과 끝을 알 수 있는 태그를 얻을 수 있다.

병렬형이라고 판단이 됐을 경우에는 병렬문장 복원을 하여야 한다. 구단위 분석 결과와 나열 및 병렬 인식에 의해 생성된 결과를 이용하여 반복되고 있는 병렬구 단위를 인식하고 생략되어진 용언을 복원한다. 먼저 범위 안에 있는 구 단위의 개수와 간단한 조사를 중심으로 한 형태소 패턴을 사용하여 반복되는 병렬구를 발견한다. 범위 내의 구 단위는 6 개 존재한다고 하면 가능한 반복 패턴은 2 개의 NP 가 반복하여 3 개의 반복 패턴이 나오는 경우와 3 개의 NP 가 반복하여 2 개의 반복 패턴이 나오는 경우가 가능하기 때문에 다음과 같이 비교를 해보면

(jc+jx):NP_B (jc):NP_I
(s):NP_I (jc+jx):NP_I
(jc):NP_I (jc):NP_O

(jc+jx):NP_B (jc):NP_I
(s):NP_I
(jc+jx):NP_I (jc):NP_I
(jc):NP_O

3 개의 NP 가 반복하여 2 개의 반복 패턴이 나오는 경우를 선택하게 되고 (s):NP_I 구 단위 뒤에 용언이 복원되어야 한다. 용언을 복원하기 위해서 먼저 복원대상 용언 후보를 선정한다.

선정하는 기준은 한국어의 경우에는 반복 패턴의 뒤쪽에 출현하는 용언만이 생략되는 특징이 있기 때문에 뒤쪽에 출현하는 용언만을 대상으로 한다. 예문에서는 하나의 용언 '수상하다' 밖에 없지만 만약 2 개 이상의 후보가 있을 경우에는 용언 후보를 선택할 수 있는 학습 데이터를 구축한 후에 앞에서 언급한 ME 모델을 사용하여 통계 모듈을 구축한 후 각각의 용언이 가지는 자질을 이용하여 확률 값을 얻은 후에 최종적으로 확률 값이 높은 용언을 복원 용언으로 선택하게 된다.

5. 실 험

5-1 실험 방법

실험을 위해서 백과사전에서 임의로 1324 문장을 선택하였다. 단, 병렬이나 나열형이 발생할만한 문장을 추출하기 위해서 20 어절 이상의 긴 문장만을 대상으로 하였다. 이 중에서 1024 문장을 최대 엔트로피 모델의 학습 데이터로 사용하였다. 학습 데이터는 형태소 분석기와 기본 구묶음을 이용하여 오류를 수정하지 않은 상태로 구축되었다. 학습 데이터 예는 다음과 같다.

0 verb_lex=배포하다 cla_rel=00
chk1234=00-1-1-1 lex1234=은
pos1234=28 sense1234=행동 chk13=-1-1
lex13= pos13= sense13= chk24=00-1
lex24=은 pos24=28 sense13=행동
chk4=0 lex4=은 pos4=28 sense4=행동

나머지 300 문장을 이용하여 평가를 하였고 평가는 형태소 분석, 기본 구묶음의 오류를 포함하여 수행하였다. 평가는 각 구묶음별로 구묶음이 나열형, 병렬형 혹은 둘다 아닌지 여부로 판단하였다. 300 문장에서 총 8,020 개의 구묶음이 발생하였고 이중 나열 및 병렬형에 속하지 않는 구묶음은 총 7,533 개로 93.93%이다.

5-2 실험 결과

실험의 베이스 라인을 모든 구묶음에 대해서 나열 및 병렬형에 속하지 않는다고 인식을 했다고 가정을 하면 93.93%의 정확률을 보이게 되는데 실제 실험 결과는 96.95%의 정확률로 베이스 라인보다 약 3.02%의 성능 향상을 보였다. 이러한 결과를 나열 및 병렬형 인식이라는 측면에서 평가하기 위해서 나열 및 병렬형인 구묶음 487 개에 대해서 평가를 하였을 경우에는 267 개의 나열 및 병렬형 구묶음으로 인식을 하고 이 중에 245 개에 대해 올바른 답을 제시함으로 해서 정확률 91.76, 재현률 50.31, F-Score 64.99 가 되었다. 그리고 인식된 병렬형에 대해서 용언을 복원하는 경우에는 병렬형을 인식하기만 하면 복원에서는 100% 성능을 보였다.

6. 결 론

본 논문에서 구묶음을 이용한 한국어 구문분석의 성능을 저하시키는 요인 중의 하나인 나열 및 병렬형 구묶음에 대해서 간단한 규칙 방법과 규칙 방법의 단점을 도와줄 수 있도록 통계 모델을 결합한 방법을 제시하였다. 제안하는 방법은 베이스라인보다 약 3.02 만큼 정확률이 향상되었다.

나열형 인식은 대상 자체가 명사구에 한정이 되어 있기 때문에 다음과 같은 용언을 포함하는 형태도 인식할 수 있도록 해야 한다.

염장법에는 (식품에 소금을 뿌리는 살염법(撒鹽法)과) (식품을 소금물에 담그는 염수법(鹽水法)이) 있는데

그리고 위와 같은 형태를 포함해서 복합적으로 조사 및 기호가 반복되는 나열형 인식도 필요하다.

참고문헌

[1] 김미영, 강신재, 이종혁, “규칙과 어휘정보를 이용한 한국어 문장의 구묶음(chunking)”, 제 12 회 한글 및 한국어 정보처리 학술대회, pp.103-109, 2000

[2] 박성배, 장병탁, “최대 엔트로피 모델을 이용한 텍스트 단위화 학습”, 제 13 회 한글 및 한국어 정보처리학술대회, pp. 130-137, 2001

[3] 박성배, 장병탁, “한국어 구 단위화를 위한 규칙기반 방법과 기억 기반 학습의 결합”, 정보과학회논문지:소프트웨어 및 응용 제 31 권 제 3 호, pp. 369-378, 2004

[4] 신효필, “최소자원 최대효과의 구문분석”, 제 11 회 한글 및 한국어정보 처리 학술대회, pp.242-247, 1999

[5] 박의규, 조민희, 김성원, 나동열, “구묶음과 구간분할을 이용한 의존 관계 추출 기법”, 제 16 회 한글 언어 인지 학술대회, pp.131-137, 2004

[6] 양재형, 심광섭, “중한번역에서 구묶음을 이용한 파싱 효율 개선”, 정보과학회논문지: 소프트웨어 및 응용 제 31 권 제 8 호, pp. 1083-1091, 2004

[7] 임수종, 정의석, 장명길, “백과사전 질의응답을 위한 격틀 기반 의존관계 분석”, 제 16 회 한글 언어 인지 학술대회, pp.167-172, 2004

[8] 임지희, 최호섭, 이정철, 옥철영, “자동 구축된 구문패턴사전과 규칙을 이용한 구묶음”, 제 16 회 한글 언어 인지 학술대회, pp.35-39, 2004

[9] Berger, A., Della Pietra, S. and Della Pietra, V., “A maximum entropy approach to natural language processing”, Computational Linguistics, 22(1):39-71, 1996.

[10] J.T.Yoon, K.S. Choi and M.S. Song, “Three types of chunking in Korean and dependency analysis based on lexical association,” In Proceedings of the 18th International Conference on Computer Processing Languages, pp. 59-65, 1999

[11] S. Abney, “Parsing by Chunks”, in Berwick, Abney, Tenny eds, Principle-Based Parsing, Kluwer Academic Publisher, pp. 257-278, 1991

[12] S.J. Lim, E.S. Jung and M.G. Jang, “Dependency Relation Analysis Using Caseframe for Encyclopedia Question-Answering Systems”, IECON, Korea, 2004