의미 정보를 이용한 이단계 단문 분할 알고리즘

박현재*, 이수선*, 우요섭**
인천대학교 정보통신공학과
g9921091@lion.inchon.ac.kr

Two-Level Clausal Segmentation Algorithm using Sense Information

Hyun-Jae Park, Su-Seon Lee, Yo-Seop Woo
Dept. of Information and Telecommunication Engineering, University of Inchon

요 약 단문 분할은 한 문장에 중심어인 용언이 복수개인 경우 용언을 중심으로 문장을 나누는 방법이다. 기존의 방법은 정형화된 문장의 경우 효율적인 결과를 얻을 수 있으나 구문적으로 복잡한 문장인 경우는 한계를 보였다. 본 논문에서는 이러한 한계를 극복하기 위해 구문 정보의 단문 분할이 아닌 의미 정보를 활용하여 복잡한 문장을 효율적으로 단문으로 분할하는 방법을 제안한다. 정형화된 문장의 경우와 달리 일상적인 문장은 문장의 구조적 애매성이나 조사의 생략 등이 빈번하므로 의미 수준에서의 단문 분할이 필요하다. 본 논문에서는 의미 영역에서 단문 분할의 할 경우 기존의 방법들의 애매성을 해소할 수 있다는 점을 보인다. 이를 위해, 먼저 하위범주화 사전과 시소러스의 의미 정보를 이용하여 용언과 보어 성분간의 의존구조를 1 차적으로 작성하고 이후 구문적인 정보와 기타 문법적인 지식을 사용하여 기타 성분을 의존구조에 점진적으로 포함시켜가는 이단계 단문 분할 알고리즘을 제안한다.

제안된 이단계 단문 분할 방법의 유용성을 보이기 위해 ETRI-KONAN 의 말뭉치 중 20,000 문장을 반 자동적인 방법으로 술어와 보어 성분간의 의존구조를 태깅한 후 본 논문에서 제안한 방법과 비교하는 실험을 수행한다.

1. 서 론

한국어 문장은 용언을 중심으로하며 나머지 어휘 성분은 부분적으로 자유로운 어순적 특성을 가지고 있다.[3][8] 한국어의 단문 분할은 한 문장에 중심어인 용언을 기준으로 복잡한 문장을 여러 개의 단문으로 나누는 방법이다.

자연언어 처리에서 단문 분할의 중요성은 기존의 많은 연구들에서 제시되었다. 기존의 단문 분할 방법은 많은 부분을 간략화된 구문 분석이나 조사등 구문 정보에 의존하는 방법을 사용하였다. 이러한 방법은 문장이 복잡해질 때 구조적, 의미적 애매성으로 인해 적절한 결과를 얻지 못하는 경우가 많다.

이러한 애매성을 해소하기 위해 공기정보[4,5],패턴정보[8]나 부가적인 한국어의 구문적 특성 등을 이용한 방법[3,4,5,6,8]들이 기존에 연구되어 왔다. 그러나, 한국

어 문장의 경우 구조의 애매성이나 조사의 생략 등이 빈번하므로 이러한 방법으로는 정확한 단문 분할 결과를 얻는데 어려움이 있다.

본 논문에서는 구조적 애매성 해소를 격조사등에 의존하기보다는 대규모의 하위범주 패턴이나 개념 정보 등을 활용하여 단문을 추출하는 것이 유용하다고 판단하여 하위범주화 사전과 계층적 개념 시소러스를 적용하여 문장이 핵심적인 보어-술어 의존 관계를 파악하고 이를 기반으로 하여 첨어 등 기타 성분을 점진적으로 처리하여 의존 관계를 확장시켜 나가는 이단계 단문 분할 알고리즘을 제안하고자 한다.

2 장에서는 본 논문에서 사용하는 한국어 용언의 하위범주화 사전과 명사의 시소러스의 의미 정보에 대해 기술하고, 3 장에서는 이단계 단문 분할 알고리즘을 제안하며, 4 장에서 실험을 통해 알고리즘을 평가한다.

2. 의미 정보

본 연구에서는 단문 분할 시 발생하는 구조적 애매성을 해소하기 위해 조사의 격과 하위범주화 사전과 시소러스의 의미 정보를 이용한다.

논문에서 구축한 하위범주화 사전은 12,000 개의 술어에 대해 25,000 개의 패턴으로 정의되어 있고, 시소러스는 기존[11]의 12,000 개의 명사 계층 사전을 본 논문의 하위범주화 사전과 정합되도록 수정하여 사용하였다.

먼저 하위범주화 사전과 시소러스의 의미 정보를 살펴보고 의미 정보의 활용 방법에 대해 설명한다.

본 논문에서 사용하는 하위범주화 사전과 시소러스 구조를 살펴보면 그림 1, 그림 2 와 같다. 하위범주화 사전은 문형 slot 와 의미 slot 를 가지고 있으며, 시소러스는 명사들간의 상하위의 계층적 구조를 가지고 있다.

논문에서 의미 정보를 활용 하는 방법은 하위범주화 사전의 의미 마커와 문장에 출현하는 명사의 시소러스 개념 정보와 비교하는 것으로 문장 성분의 의미 정보를 찾아낸다[1,2]. 정합은 시소러스 계층에서 상하위 정합을 기본으로 하지만, 정합이 되지않는 경우에는 개념 코드를 이용하여 거리를 계산하여 의미 정보를 찾아 낸다.

이러한 선택제약(Selectional Restriction) 방법을 본 논문에서 제안하는 이단계 단문 분할에 사용한다.

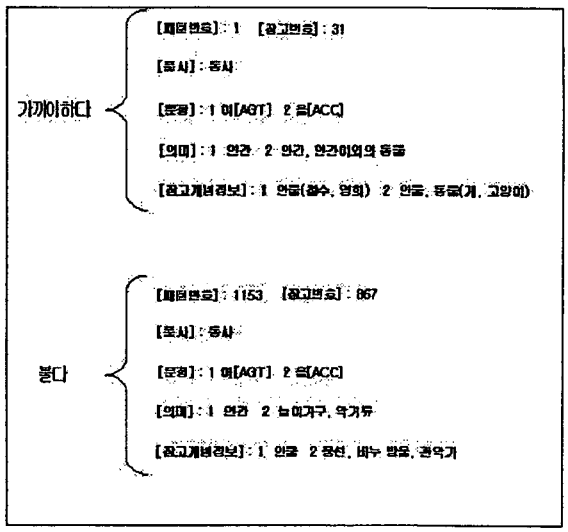

그림 1. 하위범주화 사전 구조

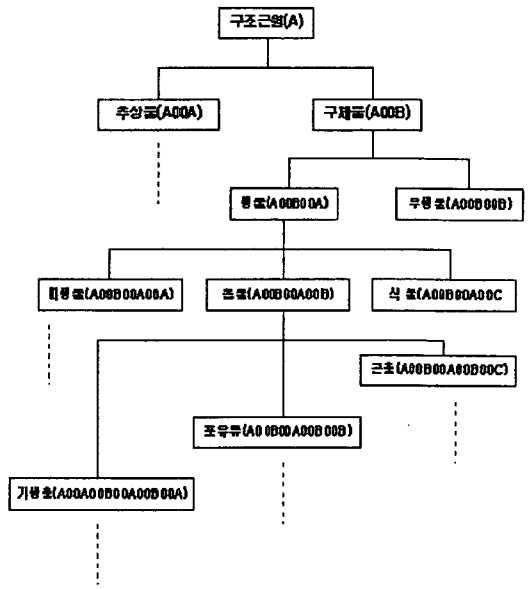

그림 2. 시소러스 구조

3. 단문 분할 과정

본 논문에서는 한 문장에서 각각의 용언들에 대한 필수 성분을 의미 정보을 이용하여 찾아내어 단문 분할 작업을 수행한다.

의미 성분을 이용하여 단문 분할을 하기 위하여 먼저 형태소 태깅을 하고, 형태소 태깅된 결과로부터 각각의 용언들에 대한 필수 성분을 찾아낸 후 그 외의 성분들을 가장 연관성이 높은 필수 성분에 연결시켜 단문 분할을 하는 이단계 단문 분할 알고리즘을 제안한다.

이단계 단문 분할 과정은 첫번째 단계로 용언들이 문장에서 이끄는 보어 성분을 찾아내는 핵심 구조 선택 과정과 두 번째 단계로 핵심 성분의 의존 구조에 첨어나 기타 성분을 점진적으로 포함시키는 의존 구조 확장 과정으로 설계하였다.

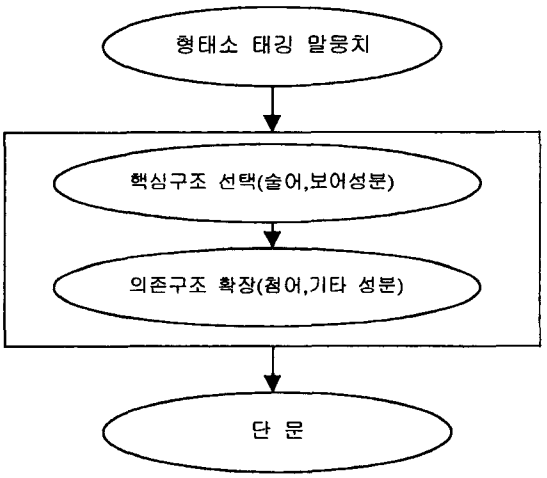

그림 3. 이단계 단문 분할 과정

3.1 핵심 구조 선택

이 과정에서는 한국어에서 가장 중요한 성분인 용언을 중심으로 문장에서의 용언이 이끄는 보어 성분을 찾아내고, 하위범주화 사전과 시소러스의 의미 정보를 사용하여 찾아낸 여러 가지 후보들 중에서 최적의 후보를 추출 한다.

또한, 핵심성분을 찾을 때에 용언의 선방향뿐만 아니라, 후방향의 일정 범위 성분들까지 비교함으로써 관형절의 경우도 단문분할이 가능하게 하였다.

핵심구조 선택 알고리즘

1) 문장내에서 용언을 추출한 후 하위범주화 사전을 검색하여 문장내의 필수격 조사와 대응되는 명사 추출
2) 용언을 중심으로 선방향과 용언의 후방향 문장성분 중 가장 가까운 2개의 필수격 명사에 대해 조사 비교
3) 선방향과 후방향의 조사가 동시에 일치 시 선방향의 조사를 우선
4) (2),(3) 과정의 반복으로 가능한 모든 후보를 추출
5) 추출된 후보들 중 최적 후보 결정 : 추출된 후보들의 조사가 이끄는 성분의 의미를 시소러스에서 찾아서 하위범주화의 의미 정보와 일치 여부를 판단한 후 가장 일치 정도가 높은 후보 선택
   (1) 조사를 가지고 있지 않는 성분의 시소러스 의미가 하위범주화 사전의 의미 마커와 정합되면 일치 성분으로 추가
   (2) 기호, 수식, 외래어의 경우에는 뒤의 조사에 따라 격을 결정
   (3) 조사가 일치했으나 의미가 틀린 경우, 하위범주화 사전의 의미역 정보가 인명/지명만의 의미를 갖는 경우에는 고유 명사로 추정
   (4) 조사가 일치된 후보가 없을 때는 하위범주화 사전의 의미 영역을 비교한 후 최적 후보 선택
   (5) 의미 영역의 비교시 시소러스의 상위 개념의 일치가 없을 때에는 시소러스 개념 유사도가 높은 후보를 선택
   (6) 문장이 내포문의 형태를 포함하는 경우는 내포문의 형태의 후보를 최적 후보로 선택 (내포문 형태 : 용언+[~라고,~고,~는 것을 등]+주격성분)

(7) 문장이 간단한 관형절의 형태를 포함하는 경우는 관형절의 형태의 후보를 최적 후보로 선택 (관형절 형태 : 용언+ㄴ어미+주격 성분, [하는, 등]+주격 성분 등)

3.1.2 의존 구조 확장 단계

본 논문에서는 핵심 구조를 선택하는 과정에서 최적 후보를 선택한 후 남은 첨어나 기타 성분들을 비핵심 성분으로 정의한다.

핵심 성분의 의존 구조에 비핵심 성분을 점진적으로 포함시키는 단계를 거쳐 단문을 추출하는 과정의 알고리즘은 다음과 같다.

의존 구조 알고리즘

(1) 비핵심 성분 중 보어 성분과 결합하여 시소러스나 복합명사 사전과 비교 후 사전에 등록되어 있는 복합명사의 경우에는 비핵심 성분을 보어 성분에 복합어로 포함

(2) 비핵심 성분 중 보어 성분과 대등적인 관계에 있는 경우에는 보어 성분에 비핵심 성분을 포함 ('와/거ㅣ', ',', '그리고'와 같이 연결된 명사구의 경우)

(3) 보조용언은 왼쪽 최인접 용언에 포함

(4) 관형어의 경우에는 왼쪽으로 가장 가까운 보어 성분에 이존

(5) 부사, 접속부사와 그 외의 성분들은 핵심성분이 아니므로 오른쪽으로 가장 가까운 보어 성분에 의존 (접속부사의 경우 맨 처음 나오면 가장 마지막 용언에 의존)

이상의 이단계 알고리즘을 밑의 예를 가지고 각 단계의 결과를 살펴보면 그림 4 와 같다.

예) **농사에 관심을 가진 지식인들은 책을 많이 썼고, 이 책을 널리 알리려고 임금에게 바치기도 하였다.**

```
:: 형태소태깅

        농사에 농사/NN + 에/JO
        관심을 관심/NN + 을/
        가진 가지/VV + ㄴ/EM
        지식인들은 지식인/NN + 들/SF + 은/JO
        책을 책/NN + 을/JO
        많이 많이/AD
        썼고 쓰/VV + 었/EP + 고/EM

        . ./SY
        이 이/DT
        책을 책/NN + 을/JO
        널리 널리/AD
        알리려고 알리/VV + 려고/EM
        임금에게 임금/NN + 에게/JO
        바치기도 바치/VV + 기/EM + 도/JO
        하였다 하/VX + 었/EP + 다/EM
        . ./SY

:: 핵심 구조 선택 단계를 거친 결과

        지식인들은 관심을 가지다.
        지식인들은 책을 쓰다.
        지식인들은 책을 알리다.
        지식인들은 책을 임금에게 바치다

:: 의존 구조 확장 단계를 거친 결과

        지식인들은 농사에 관심을 가지다.
        지식인들은 책을 많이 쓰다.
        지식인들은 이 책을 널리 알리다.
        지식인들은 책을 임금에게 바치기로 하였다
```

그림 4. 이단계 단문 분할의 단계별 결과

4. 실험 및 결과

이단계 단문 분할 알고리즘의 실험에는 25,000 개의 용언을 가지고 있는 하위범주화 사전과 시소러스[11]를 이용하였다. 제안된 알고리즘의 평가를 위하여 형태소 분석된 말뭉치를 이단계의 단문 분할 알고리즘을 거쳐서 나온 단문과, ETRI-KONAN 의 말뭉치 중 20,000 문장을 반 자동작인 방법으로 술어와 보어 성분간의 의존구조를 태깅한 결과를 비교하는 실험을 수행하였다.

실험 대상 20,000 만 문장 중 용언의 수는 43,067 개이었으며, 본 실험의 이단계 단문 분할 알고리즘의 핵심 구조 선택 단계에서 41,090 개의 용언이 성공하여 95%

정도의 정합률을 보였다. 여기서 정합에 실패한 경우는 하위범주화 사전에 용언이 존재하지 않는 경우(1%이내)가 포함되며 하위범주화 사전의 의미와 시소러스 개념이 잘 정합되지 않거나 알고리즘이 미비하여 보완이 필요한 경우가 있었다.

다음 단계로 의존 구조 확장 단계에서는 핵심 구조 선택 단계에서 정합된 41,090 문장 중 37,820 개가 성공하여 92.2%의 정합률을 보였다.

5. 결 론

본 연구에서는 한국어 문장을 단문으로 분리하는 이단계 단문 분할 알고리즘을 제안하고 검증하였다. 제안된 알고리즘은 단문 분할 과정에서 구문 정보뿐만 아니라, 의미정보와 선택 제약을 이용함으로써 단문 분할 시 나타나는 애매성을 줄일 수 있고, 보다 정확한 단문 분할이 가능하다고 판단된다.

본 연구에서 사용하는 하위범주화 사전의 격과 의미 정보의 정확성은 단문 분할 과정의 정확도에 큰 영향을 주기 때문에, 추후 하위범주화 사전의 수정 및 용언 추가가 필요하며 단문 분할에 활용 가능한 구문 정보를 보다 다양하게 활용하는 방향의 연구를 수행할 예정이다.

이 논문은 정보통신부 대학 기초 연구 지원 사업의 연구비 지원에 의해 수행되었음.

참고문헌

[1] 추교남, "개념 기반 정보 검색을 위한 한국어 어휘의 의미 분석", 인천대학교 정보통신공학과 석사 학위 논문 1998

[2] 추교남, 우요섭, "어휘와 구문의 중의성 해소를 위한 한국어 하위범주화사전 구축", 정보처리학회 98 추계 학술발표논문집, 1998

[3] 김광진, 송영훈, 이정현 "한국어 내포문을 단문으로 분리하는 시스템 구현" 한글 및 국어정보처리 학술발표 논문집, 1994

[4] 이현아, 이종혁, 이근배, "구문분석과 공기정보를 이용한 개념 기반 명사구 색인 방법" 제 7 회 한글 및 한국어 정보처리 학술발표논문집,1995

[5] 이현아, 이종혁, 이근배 "단문 분할을 통한 명사구 색인방법", 정보과학회논문지(B) 제 24 권 제 3 호,1997

[6] 양단희, 송만석 "말뭉치로부터 격틀 구축에 필요한 학습 데이터 추출", 제 10 회 한글 및 한국어 정보처리 학술대회

[7] 김나리, "패턴 정보를 이용한 한국어 구문 분석", 서울대학교 컴퓨터공학과 박사학위논문, 1997

[8] 박성배, "문장분할을 이용한 한국어 분석", 서울대학교 컴퓨터공학과 석사학위논문, 1996

[9] 이호, 백대호, 임해창, "분류 정보를 이용한 단어 의미 중의성 해결", 정보과학회 논문지(B) 제 24 권 제 7 호, 1997

[10] 김영택, "자연 언어 처리" 교학사 1994

[11] 우요섭 "토큰 기반 한국어 분석기 개발 - 한국어 의미 분석 사전 및 하위범주화사전 구축", 한국전자통신연구원, 1997