신문 기사 요약문 생성을 위한* 구문 분석기 구현

정영규[0] 이현주 이상조
경 북 대 학 교 컴 퓨 터 공 학 과
{reraj, hjlee, sjlee}@sejong.knu.ac.kr

Implementation of a parser for news summarization

Young-Giu Jung[0] Hyun-Ju Lee Sang-Jo Lee
Dept. of Computer Engineering, KyungPook National University

요 약

본 논문은 문서요약 시스템의 일부로써 신문기사의 문장을 효율적으로 구문 분석할 수 있는 구문 분석기를 구현한다. 요약의 대상인 신문 기사의 문장은 보조동사, 화용조사, 인용동사 등 많은 동사들을 가지며, 이와 같은 동사들은 구문분석을 할 때 많은 문제점을 발생시킨다. 본 논문은 이러한 동사들을 단위화하고, 여기서 발생하는 주어 생략과 모호성 문제를 해결하는 방법을 제시한다. 그리고 단위화의 결과로 나온 의미적 중심용언을 이용하여 문장의 필수 성분을 추출한다.

1. 서론

일반적으로 구문 분석기는 문장을 구성하는 각 성분의 문법적 관계를 밝히는 것이다. 그러나 이것은 많은 모호성의 발생과 처리 속도의 저하 때문에 구문 분석기 구현에 많은 어려움이 있다. 따라서 실제적인 응용 시스템에서 사용되는 구문 분석기는 시스템에 적절한 형태의 결과 출력과 성능향상을 위해 구현되는 것이 바람직하다.

본 논문은 문서요약 시스템의 일부로써 신문기사의 문장을 효율적으로 분석할 수 있는 구문 분석기를 구현한다. 신문 기사는 일반적으로 사용되어지는 문장과는 많은 차이점이 있다. 그 중 가장 두드러진 것은 문장 안에서 용언들이 겹쳐서 나타나는 것이다. 이러한 용언 중 반 정도가 문장의 의미와 관련이 없으므로 이것을 배제하면 구문 분석 시 발생하는 모호성을 감소시키고, 사전 검색의 횟수를 줄여 시스템의 성능을 향상시킬 수 있다. 그리고 이러한 단위화에서 발생하는 주어 생략과 모호성 문제에 대한 해결방법을 제시한다. 마지막으로 단위화 처리의 결과로 찾은 의미적 중심용언을 이용하여 문장의 필수 성분을 추출한다.

2. 관련연구

초기에 의존 문법을 이용한 구문 분석기의 연구는 문장 내에 발생하는 중의성의 종류를 나열하고 이를 처리하는데 초점을 맞추었다[1]. 이러한 구문 분석기의 처리 결과는 문장 내에 있는 어형들 사이의 모든 관계를 밝히고 여기에서 발생하는 모호성이 있을 때 각각의 결과를 모두 내어 준다. 이 같은 시스템의 경우 많은 처리 시간과 결과의 오류 가능성이 존재함으로 실제적인 요약 시스템의 입력으로 쓰기에는 많은 문제점들이 있다.

오늘날 구문 분석기들에서는 처리의 단순화와 성능의 향상을 위한 연구가 진행되고 있다. 이것들 중 하나로 최장 묶음을 이용한 구문 분석 방법이 있다.[2] 여기에서는 사전을 이용하여 문장 내 어절들을 단위화하여 구문 분석의 성능향상과 모호성을 제거하고 있다. 이와 비슷한 연구가 단위(Chunk)분석이다. 문장을 적절한 크기의 명사구와 동사구로 단위화함으로써 구문 분석의 과정에 발생하는 모호성을 줄이고 시스템의 성능을 향상시킨다.[3]

본 논문은 신문 기사에 나타나는 동사의 겹침 현상을 이용하여 용언을 단위화하고 이것을 이용하여 신문기사의 구문 분석 시 발생할 모호성을 최소화한다. 그리고 의미적 중심용언을 이용하여 문장의 필수 성분을 추출한다.

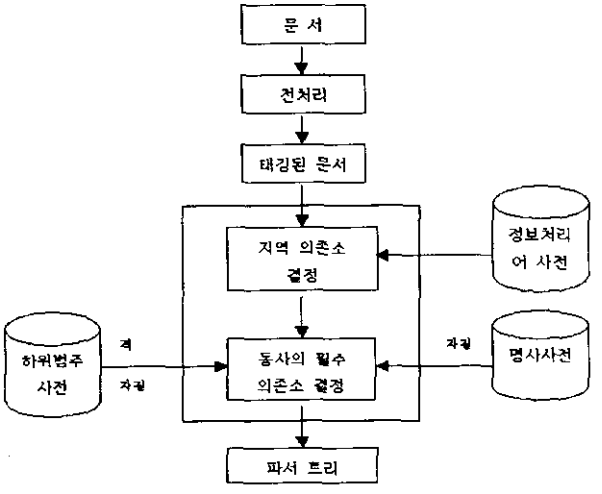
* 본 연구는 정보통신연구진흥원의 대학기초연구지원사업 과제 " Web 상에서 적확한 검색을 위한 문서의 대표 개념어 생성 및 요약 시스템" 의 일부로 수행되었음.

3. 구문 분석기의 구현

문장은 필수 성분과 부속 성분으로 나누어 진다.[4] 본 논문에서는 주어, 목적어, 필수 부사어, 서술어로 한정하고 그 외의 성분은 부속 성분으로 정의한다. 다음은 전체 시스템의 구조이다.

3.1 구문 분석기의 구조

구문 분석기는 전처리 단계와 구문 분석단계로 구성된다. 전처리 단계에서는 형태소 분석기와 태거를 사용하고, 이 단계의 처리결과로 <그림1>에서 보여지는 태깅된 문서를 제공한다. 구문 분석단계는 지역 의존소 결정단계와 필수 의존소 결정단계로 구성되며, 각 단계에서 사용하는 사전은 정보처리어 사전과 명사 사전, 하위 범주사전이다. 이러한 사전들은 다음 절에서 설명할 것 이다.



<그림 1> 구문 분석기의 구조

3.2 지역 의존소 결정 및 필수 성분 추정

본 논문에서 제안한 지역 의존소 결정은 동사구 단위화 단계와 명사구 단위화 단계로 구성된다. 첫 단계인 동사구 단위화 방법은 신문 기사상에서 빈번히 나타나는 용언의 겹침 정보를 이용한다. 이러한 용언들 중 화용조사, 보조용언, 간접인용, 직접인용을 ' 정보처리어' 라고 정의하고, 이것을 사전으로 구성한다. 정보처리어 사전의 사용은 문장 내에서 처리 할 용언의 수를 줄임으로써 사전 검색의 수를 현저히 감소시킨다. 다음은 정보처리어의 예로써 총 400여개로 이루어져있다.

● ~에 대해	화용조사
● ~ㄹ 것	보조용언
● 고 밝히	간접인용
● 고 예측하	간접인용
● " 라고 말하	직접인용
⋮	

두 번째 단계인 명사구 단위화는 품사 정보를 이용하여 격조사 및 연결어미, 관형형 전성어미, 보조사, 부사를 중심으로 단위화한다.

아래의 예는 지역 의존소 결정 알고리즘 수행 후 결과이다.

입력: 한국투자신탁은 3일 향후 장세는 수급악화와 대외 경제여건 악화등으로 대형주의 주가는 당분간 조정기간을 거칠 것으로 예상된다고 지적했다.

출력: 한국투자신탁은 3일 향후 장세는 수급악화와 대외 경제여건 악화등으로 대형주의 주가는 당분간 조정기간을 거치+(ㄹ 것으로 예상된다고 지적했다.)

* 중심용언은 " 거치다" 이다.

정보처리어 사전에 의한 단위화 처리는 한 가지 문제점을 가진다. 이러한 문제점은 배제된 정보처리어의 필수 의존소 결정 문제이다. 본 논문에서는 이와 같은 정보처리어 중 인용문장에서의 인용 주체의 생략과 모호성 문제에 대한 해결 방법을 제시한다. 인용문에서의 인용 주체 결정 문제는 인용 주체의 의미적 자질 정보를 이용한다. 인용문을 분석하여보면 인용 주체의 의미자질이 ' 단체' 나 ' 사람' 으로 한정됨을 알 수 있다. 이러한 특징을 이용하여 명사 사전 내의 의미 자질 분류와 하위범주 사전에서 각 격에 올 수 있는 의미 자질 정보, 그리고 필수 성분 결정 챠트 정보를 이용하여 한 개의 기사 내에서 인용문장의 주체로 올 수 있는 후보를 선정하고 이러한 후보들 중 인용문 내의 주어를 결정한다.

동사 단위화 처리의 결과로 문장 내 의미적 중심용언을 찾을 수 있다. 따라서 구문 분석을 할 때 의미적 중심용언의 필수 성분에 대한 문법적 관계만을 고려함으로써 구문 분석 시 발생하는 사전 검색 수의 감소로 인한 시스템의 성능 향상을 이룰 수 있다. 다음 절은 의미적 중심용언의 필수 의존소 결정 알고리즘에 대한 설명이다.

3.3 동사의 필수 의존소 결정

본 논문에서는 정보처리어의 단위화로 얻은 의미적 중심용언과 하위범주 사전을 이용하여 문장 내 동사의 필수 의존소들을 결정하는 알고리즘을 제안한다. 본 알고리즘에서 이용하는 하위범주 사전은 동사의 필수 격 정보와 각각의 격 내에 올 수 있는 명사 의미 자질로 구성되고, 하위범주 사전내의 동사가 여러 개의 격을 가질 때에 최장 성분 일치 격구조 원칙을

따른다. 다음은 알고리즘의 수행결과로써 제공되는 두 개의 챠트 구조에 대한 설명이다. 첫 번째 챠트는 6개의 속성으로 구성되어 있으며, 여기에 각 용언의 필수 의존소가 될 후보 의존소 정보를 기록한다. 두 번째 챠트는 첫 번째 챠트에서 기록된 각각의 정보에 대한 문장 성분 정보를 기록한다. 이렇게 구성된 두개의 챠트 정보는 문장 내 각 동사의 필수 의존소들을 결정하는 정보로 사용된다. 다음은 본 논문에서 제시한 여섯 개의 속성이다.

- 단어 일치
- 단어의 속성 및 격 정보 일치
- 격 정보 일치
- 전체 문장 내 단어 일치
- 전체 문장 내 단어의 속성 및 격 정보 일치
- 전체 문장 내 격 정보 일치

다음은 의미 자질 정보를 이용한 필수 성분 추출 알고리즘이다.

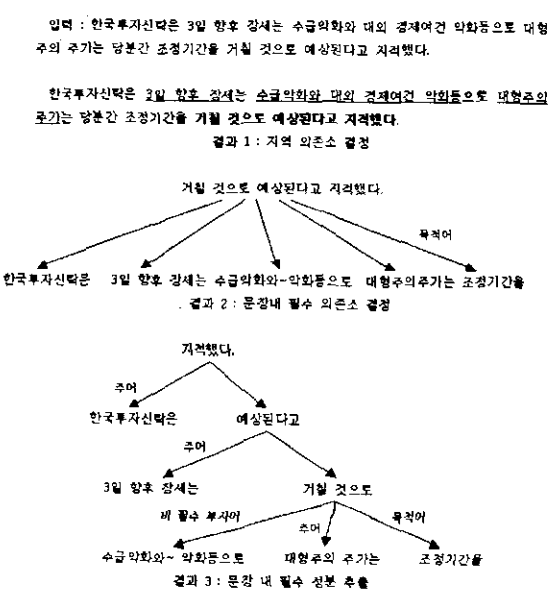
알고리즘 : 필수 성분 추출 알고리즘

단계 1. 정보처리어 배제;
단계 2. 연결어미의 위치를 찾음;
단계 3. for(연결 어미의 수)
단계 4. 현재 연결의미의 위치에서 앞쪽에 위치하는 동사의 후보 의존소를 찾음;
단계 5. 후보 의존소의 명사자질과 하위범주 사전의 자질과 비교;
단계 6. 후보 의존소의 6개 속성에 값을 저장;
단계 7. 두개의 챠트를 만듬;
단계 8. 두 챠트의 등록된 정보를 이용하여 동사의 필수 의존소 결정;

4.실험 및 결과

본 논문에 구문 분석의 대상은 신문 기사 100편을 대상으로 한다. 신문 기사에서 추출한 정보처리어의 수는 400개로써 코퍼스의 양이 늘어나더라도 정보처리어의 수는 일정 양으로 수렴한다. 따라서 구문 분석 시 정보처리어의 처리는 시스템의 성능을 향상 시킬 수 있다.

한 문장에서 보여지는 평균 용언의 수가 약 3.57개이고 포함된 정보처리의 수가 약 1.5개정도 되므로 이러한 용언을 미리 처리함으로써 구문 분석 과정에서 발생하는 사전의 검색 횟수를 반 정도 감소시킨다. 그리고 의미적 필수 용언만을 처리함으로써 시스템 전반적인 모호성을 감소 시킬 수 있다. <그림2>는 입력 문장에 대한 명사구 단위화와 동사구 단위화를 처리하고, 정보처리어의 결과로 나온 의미적 동사로부터 필수 성분을 추출한다. 그리고 보조사의 특성과 정보처리어의 주체가 갖는 명사 자질 정보를 이용한 필수 성분 생략 및 모호성 처리가 끝난 결과이다.



<그림 2> 시스템 실행 결과

5. 결론

요약의 대상인 신문 기사의 문장은 보조동사, 화용조사, 인용동사 등 많은 동사들을 가지며, 이와 같은 동사들은 구문분석을 할 때 많은 문제점을 발생시킨다. 본 논문은 이들을 정보처리어라 정의하고, 이를 구문 분석의 대상에서 제외함으로 구문 분석시 발생하는 모호성을 감소시키고 시스템 성능을 향상 시켰다. 그리고 정보처리어의 단위화에 따른 주어 생략 및 모호성 문제를 제기하고 이를 해결하는 방법을 제시한다. 마지막으로 정보처리어의 단위화 결과로 찾은 의미적 중심 용언을 가지고 문장의 필수 성분을 추출하는 구문 분석기를 구현하였다.

6. 참고 문헌

[1]. 홍영국, " 의존 문법에 기반을 둔 한국어 구문 분석기의 설계 및 구현" MS. Thesis, 1993.
[2]. 박상규, 정창민, 조준모, 이상조 "최장 묶음을 이용한 효과적인 한국어 구문 분석기" 제 22회 한국정보과학회 봄 학술 발표논문집, pages 961-964, 1995
[3]. 김미영, 강신재, 이종혁 " 단위(Chunks)분석과 의존문법에 기반한 한국어 구문 분석" 제27회 한국정보과학회 학술 발표논문집, pages 327-329, 2000.
[4]. Michael A. Covington " A Dependency Parser for Variable-word-Order Language" Research Report AI-1990-01, Artificial Intelligence Pragrames, Univ.of.Georgia, 1990.
[5]. I.A.Mel' cuk," Dependency Syntax: Theory and Practice" State Univ.of New York Press, 1988