제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

구 분할을 이용한 명사구기반 색인의 성능향상

이충희, 김현진, 장명길
한국전자통신연구원(ETRI)
e-mail : forever@etri.re.kr

Improvement of phrase-based indexing performance using phrase segmentation

Chung-Hee Lee, Hyun-Jin Kim, Myung-Gil Jang
Electronics and Telecommunications Research Institute (ETRI)

요 약

정보검색의 정확률을 높이는 것이 최근 정보검색 연구의 추세이며, 정확률을 높일 수 있는 방법 중 하나로 명사구단위 색인이 있다. 명사구 색인을 하는 방법에는 구문분석기를 이용하는 방법과 패턴 규칙을 이용하는 방법으로 나눌 수 있다. 구문분석기를 이용하여 전체 문장을 분석한 후 명사구단위 색인을 할 경우, 범용적으로 이용할 수 있지만 속도와 정확도가 떨어진다는 문제점이 있으며 패턴 규칙을 이용하는 경우는 속도는 빠르지만 정확도 및 확장성에 문제를 가지고 있다. 이런 문제들을 해결하기 위해 본 논문에서는 문장으로부터 명사구를 분할한 후, 분할된 명사구를 완전 구문 분석하여 색인하는 방법을 제안한다. 명사구는 속격어구와 관형형 명사구를 대상으로 하였고, 구 분할은 속격조사와 관형형어미를 중심으로 주변 형태소와 품사를 고려하는 규칙을 만들어 실행하였다. 실험대상은 짧은 문장, 중간문장, 긴 문장을 각각 25 개를 선정하여 실험하였고, 구 분할을 이용할 경우 평균 재현율은 86%, 평균 정확률은 74% 정도의 성능을 보였다. 긴 문장의 경우, 구 분할을 이용하지 않는 경우에 비해서 정확도 및 속도에서 월등한 성능향상이 있었다.

1. 서론

인터넷의 정보량이 많아지면서 정보검색 정확도의 중요성이 커지고 있다. 정보검색 정확도를 높이기 위한 방법으로 구 단위 색인이나 문장 단위 색인을 이용하는 방안이 있고, 구 단위 색인을 위해서 구문분석기를 이용하거나 규칙이나 통계정보를 이용하는 별도의 구 묶음 모델을 이용한다.

구문분석기를 이용하는 방법은 처리할 문장이 길어질수록 구문분석기의 속도와 정확도가 많이 떨어지는 문제를 가지며 패턴 규칙을 이용하는 구 묵음 모델을 이용하는 방법은 속도는 빠르지만 규칙을 뽑는 데이터에 의존적이고 확장성에 문제를 가지며 색인을 위해서 구 묵음 내부를 다시 분석해야 하는 문제를 가진다.

이러한 문제들을 해결하기 위해서 본 논문에서는 구문분석을 하기 전에 구 분할을 수행하여 구문분석기의 속도와 정확도를 향상시키고자 한다. 즉, 구문분석기의 입력으로 전체문장을 주는 것이 아니라 처리하고자 하는 명사구만을 분리하여 입력으로 넣어주는 것이다.

구 분할은 명사구를 대상으로 하였고 명사구로는 속격어구와 관형형 어구만을 고려하였다. 구 분할은 규칙에 기반해서 실행되는데 9 가지 규칙을 일반 말뭉치에서 사용되는 문장들을 고려하여 설정하였다. 실제 상용시스템에서 사용될 것을 고려하여 규칙은 최대한 단순화 하였고 추출되는 명사구 또한 단순한 형태만을 추출한다. 복잡한 명사구 형태를 추출해야 할 경우도 있지만 그런 경우는 의외로 희박하므로 단순한 명사구만을 추출하여도 실험결과 91%이상의 명사구 추출 정확률을 보였다.

구문분석기의 알고리즘이 다양하고 각각의 성능도 차이가 있지만 입력되는 문장이 상당히 짧은 경우에

는 알고리즘에 상관없이 대부분 정확한 결과와 빠른 분석속도를 가지므로 특정 구문분석기를 가지고 있고 별도의 구 묶음 모델이 없을 경우, 본 논문에서 제안하는 규칙을 이용한 구 분할을 구문분석기의 전처리기로 사용한다면 구 기반 색인을 이용한 정보검색에서 상당한 성능향상을 얻을 수 있을 것이다.

본 논문의 구성은, 2 절에서는 구 묶음에 대한 기존 연구들을 알아보고, 3 절에서 본 논문에서 사용된 구 분할 알고리즘을 상세히 설명한다. 4 절에서는 실험에 사용된 전체 시스템을 설명하고, 구 분할을 이용한 색인 실험에 대하여 언급하고, 실험결과를 분석한다. 5 절에서 결론을 내리고 앞으로의 연구방향에 대해서 설명한다.

2. 관련 연구

구문분석기를 이용한 구 단위 색인의 경우, 구문분석기 자체의 성능에 문제가 있으므로 연구가 활발히 이루어지고 있지는 않다.

[1]은 구문분석기를 이용한 구 단위 색인 방법을 보이며 구 단위 색인의 성능을 향상시키기 위해서 의존규칙과 구문패턴 정보를 이용하는 단문분할기를 사용하여 구문분석기의 성능을 향상시킨다. 명사구 색인 실험결과, 재현율 61.57%, 정확률 63.48%의 성능을 보인다. 단문분할기를 사용함으로써 재현율과 정확율이 각각 20%와 70%정도의 성능향상을 보이지만 아직 단문분할기 자체의 오류가 있고 단문분할이 이루어지지 않는 문장에 대해서는 해결방법이 없다는 문제가 있다.

구 묶음 또는 명사구 인식에 대한 연구는 규칙기반의 방법과 통계적인 방법을 사용하는 것으로 나눌 수 있으며 두 가지를 모두 사용해서 각각의 단점을 보완하여 성능을 향상 시킬 수 잇다..

[2]는 간단한 구문패턴이나 경험적 규칙에 의존하는 방법이 한계가 있다고 보고 구문적 방법론으로서 동일 단문내의 단어들이 다른 단문 내의 단어들보다 높은 개념적 연관성을 보인다는 점에 주목하여 단문을 기반으로 한 명사구 색인 방법을 제안하였다. 의존관계 파싱과 용언-격조사 간의 공기강도 정보를 이용하여 단문분할을 수행하며 분석된 단문 내에서의 어절간 의존관계를 이용하여 구문적으로 유용한 명사구들을 추출한다. 완전 매칭의 경우, 58.65%의 재현율과 31.61%의 정확률을 보인다. 구문분석을 이용해서 단문분할을 수행하므로 구문분석기 자체의 성능에 문제가 있고, 단문 생성시에 오류가 일어날 경우, 연관된 여러 단문들에 동시에 오류가 일어나므로 성능이 많이 떨어지는 문제가 있다.

[3]은 규칙적인 어순을 발견하여 규칙을 이용한 구 묶음 과정을 보이는데 규칙만으로는 한계가 있으므로 말뭉치에서 어휘 정보를 추출하여 보완한다. 어휘 정보는 부사어 역할을 하는 명사에 관해서만 이용하였다. 실험결과, 구 묶음 정확률이 98.7%정도의 성능을 보이지만 실험 문장의 길이가 짧으므로 정확한 성능을 알 수 없다.

[4]는 격조사의 구문적인 특성을 이용하여, 수식어까지 포함한 명사구를 추출하는 방법을 제시한다. 구문 특성으로는 명사구의 처음과 끝 그리고 명사구 주변의 형태소를 이용하여 명사구의 수식 부분과 중심 명사를 문맥정보로 사용한다. 다양한 형태의 문맥 정보들은 최대 엔트로피 원리에 의해 하나의 확률 분포로 결합된다.

[5]는 기반 명사구(비재귀적인 단순 명사구)를 인식하는 방법으로 학습 말뭉치를 정답 말뭉치인 목표 말뭉치로 변화시키는 비통제 학습을 통해 인접한 주위 형태소들의 다양한 문법적 정보를 나타내는 규칙 템플릿을 이용하는 규칙들을 얻어낸다. 이런 방법에는 학습에 이용되는 목표 말뭉치의 역할이 매우 중요한데, 목표 말뭉치로 이용될 수 있는 트리 태그 부착 말뭉치의 양이 많지 않다는 문제를 가지고 있고 목표 말뭉치 자체의 오류 또한 커다란 영향을 미친다.

구문 분석기나 구 묶음 모델을 이용할 때, 발생하는 여러 문제 및 단점들을 해결하기 위해서 본 논문에서는 두 가지 방법을 절충하여 먼저 구 분할을 통해 처리하고자 하는 부분을 추출하고, 추출된 부분을 구문분석기에서 처리함으로써 속도와 색인 정확도를 모두 향상시킬 수 있었다. 다음 절에서 첫번째 단계인 구 분할 알고리즘에 대해서 알아보겠다.

3. 구 분할 알고리즘

구 분할에 이용된 규칙은 크게 두 가지로 나눌 수 있다. 첫째는 중심 어절의 앞부분을 추출하는 부분과 둘째는 중심 어절의 뒷부분을 추출하는 부분이다. 여기서 중심어절이란, 추출하고자 하는 명사구인 속격어구나 관형형 어구가 들어있는 어절을 말한다.

구 분할에 이용된 규칙은 다음과 같다.

① 중심 어절 앞부분 설정(구 시작 위치)
- A. 순수복합명사[1]나 명사 병렬어구[2]만 고려
- B. 명사 병렬어구를 구성하는 단일명사가 세 개 이상일 경우는 병렬어구 처음까지 설정
- C. 명사 병렬어구를 구성하는 단일명사가 두 개일 경우
  - i. 명사 병렬어구 앞의 구조가 중심 어절과 같을 경우는 병렬어구 가운데에서 설정
    예: A 의 B 와 C 의중심 D 를 보다.
    → C 의 D
  - ii. 명사 병렬어구 앞의 구조가 중심 어절과 다를 경우는 병렬어구 처음까지 설정
    예:A 한 B 와 C 의중심 D 를 보다.
    → B 와 C 의 D

② 중심 어절 뒷부분 설정(구 마지막 위치)

[1] 단일 명사들이 연결된 형태, 띄어쓰기도 상관없이 고려(예: 요즘은 한국전자 통신연구원의 실적이... → '한국전자 통신연구원')

[2] And 나 Or 로 연결된 명사들. (예: 사과, 딸기 및 복숭아의 가격이... → '사과, 딸기 및 복숭아')

A. 주격, 목적격 조사를 가진 어절에서 설정
B. 명사 병렬어구를 구성하는 단일명사가 세 개 이상일 경우는 병렬어구 마지막까지 설정
C. 명사 병렬어구를 구성하는 단일명사가 두 개일 경우
   i. 명사 병렬어구 앞의 구조가 중심 어절과 같을 경우는 병렬어구 가운데에서 설정
      예: A 의중심 B 와 C 의 D 를 보다.
      → A 의 B
   ii. 명사 병렬어구 앞의 구조가 중심 어절과 다를 경우는 병렬어구 처음까지 설정
   iii. 예:A 한중심 B 와 C 의 D 를 보다.
      →A 한 B 와 C
D. 중심 어절이 관형형인 경우
   i. 관형형 어미나 속격조사를 가진 어절에서 설정
E. 중심어절이 속격인 경우
   i. 속격조사를 가진 어절에서 설정
   ii. 관형형에서 설정 금지
      ※ 예외: 관형형 어절이 '명사+지정사' 형태일 경우 설정
      → 예외의 예외: '명사+적+지정사'인 경우 관형형 설정 금지
F. 격조사를 가진 어절에서 설정(속격조사나 명사 병렬어구일 경우 제외)

구 분할 규칙 중 중심어절의 앞부분을 설정하는 경우는 중의성이 별로 없지만 뒷부분을 설정할 경우에는 중심어절이 먼 거리에 있는 어절과 의존관계를 가질 수도 있으므로 얼마정도의 오류가 발생한다. 하지만 그러한 복잡한 문장구조는 많이 발생하지 않으므로 명사구를 분할하는 실험에서 91%정도의 정확률을 얻을 수 있었다.

4. 실험

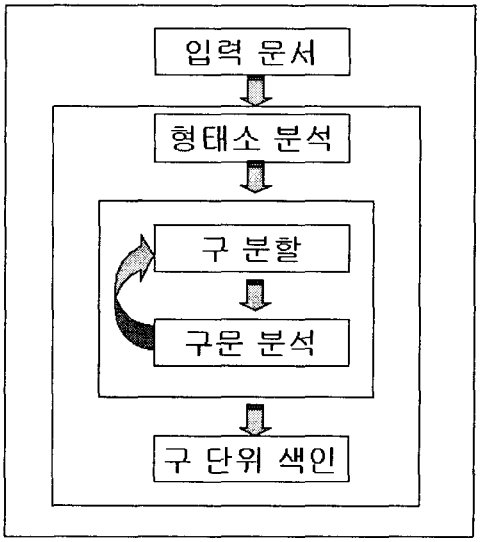
이번 실험에서는 [1]에서 사용한 시스템에 구문분석기의 전처리기로 구 분할기를 추가하여 시스템을 구성하였고 [1]의 실험결과와 비교하여 성능을 측정하였다. 사용된 시스템의 전체 구성도는 그림 1 과 같다.

그림 1 의 형태소 분석에서 구 단위 색인까지의 과정은 입력문서에 있는 모든 문장에 대해서 처리하며, 형태소 분석된 결과를 참조해서 구 분할 과정을 수행하고, 구 분할 과정에서 추출된 모든 구에 대해서 구문 분석을 실시하고, 구문 분석결과를 분석하여 색인 결과를 얻는다.

형태소 분석기와 구문 분석기는 [1]에서 사용한 SCAN 과 토미타의 GLR 구문 분석기를 그대로 사용한다.

실험은 국어정보베이스 II 와 조선일보의 이규태 코너에서 무작위로 선정한 74 문장을 대상으로 실험하였다.

성능을 비교하기 위한 비교 대상으로 첫째, Baseline 으로 사용한 것은 순수하게 구문분석기만을 사용하여 명사구 색인을 한 경우(Case1)와 둘째, 전처리기로 단문분할을 실행하고 단문 분할된 문장들에 대해서 명사구 색인을 한 경우(Case2)를 설정하였다.



[그림 1] 시스템 구성도

[표 1] 색인 결과(평균 어절 수: 15)

	Case1	Case2	Case3
정답개수	66		
색인량	119	119	74
일치개수	44	44	57
재현율	66.67	66.67	86.36
정확률	36.97	36.97	77.03

[표 2] 색인 결과(평균 어절 수: 24)

	Case1	Case2	Case3
정답개수	82		
색인량	153	127	95
일치개수	52	51	66
재현율	63.41	62.2	80.49
정확률	33.99	40.16	69.47

[표 3] 색인 결과(평균 어절 수: 46)

	Case1	Case2	Case3
정답개수	55		
색인량	180	121	67
일치개수	31	37	50
재현율	56.36	67.27	90.91
정확률	17.22	30.58	74.63

표 1 은 보통길이의 문장으로 24 문장에 대해서 실

험하였다. 이 때 사용된 문장들은 단문분할이 발생하지 않아서 Case1, Case2 의 결과가 동일하다. 구 분할을 사용한 Case3 의 경우 1,2 에 비해서 좋은 결과를 얻었다.

표 2 는 중간정도 긴 문장으로 25 문장에 대해서 실험하였는데 단문분할이 발생한 Case2 가 Case1 보다 정확률이 좋았고 Case3 의 경우는 가장 좋은 결과를 얻었다.

표 3 은 매우 긴 문장으로 25 문장을 색인하였고 Case1 보다 Case2 가 상당히 결과가 우수하고 Case2 보다 Case3 가 훨씬 좋은 성능을 보이는 것을 확인할 수 있다.

표 1,2,3 의 실험결과를 비교해보면, 문장이 길어질수록 구문분석기만 사용한 경우는 성능이 많이 떨어지는 것을 알 수 있고 단문분할기를 이용할 경우, 성능이 떨어지는 것을 어느 정도는 막아주지만 그래도 성능저하가 있는 것을 확인할 수 있다. 하지만 구 분할을 사용한 Case3 의 경우는 문장의 길이에 상관없이 일정한 성능을 유지한다는 것을 알 수 있다.

추가적으로 색인 속도를 비교하기 위해서 HANTEC 2.0 의 HKIB94 1,000 문장을 색인하는 실험을 하였다. 표 4 는 그 결과를 보여준다.

[표 4] 색인 속도 측정

	Case1	Case3
시간	137 분	6 분 30 초

실험 결과, 구 분할을 이용할 경우에 확실한 속도 향상이 있음을 알 수 있다.

5. 결론

본 논문에서는 구 단위 색인을 위해 사용되는 구문분석기와 구 묶음 모델의 각각의 단점을 보완하고 장점을 이용할 수 있는 절충 방법으로 구문 분석기의 전처리기로 구 분할기를 사용하는 방법을 보였다.

구 분할 규칙은 추출하고자 하는 명사구를 추출할 수 있는 간단한 규칙을 사용하고 구문 분석기는 기존에 만들어진 모델을 그대로 사용하면서도 두 가지를 같이 사용함으로써 상당한 색인 성능 향상을 얻을 수 있다는 것을 알 수 있었다.

구문 분석기 자체를 수정하거나 구문 분석기 안에 구묶음 모델을 추가하여 사용한다면 많은 시간과 노력이 필요하고 호환 및 확장성에 제약이 있으므로 구문 분석기의 전처리기로 별도의 구 분할 모델을 만들어 사용하는 것도 하나의 대안이라고 생각된다.

본 연구는 구 분할기를 상당히 단순화 시켜서 어느 정도의 오류를 가진 상태에서 구문 분석을 수행하므로 아주 좋은 결과를 얻지는 못했다. 전처리기로서 구 분할기는 전체 색인 성능에 절대적인 영향을 줄 수 있으므로 추가적인 보완이 필요하고, 이번 실험에서는 규칙만 사용하였는데 공기 정보 등의 통계 정보를 추가로 이용한다면 더욱 좋은 결과를 얻을 수 있을 것으로 생각된다.

참고문헌

[1] Chung-Hee Lee, Hyun-Jin Kim, Myung-Gil Jang, "Improving speed and precision in phrase based indexing by using clausal segmentation", Proceeding of International Association of science and technology for development, 2002

[2] 이현아, 이종혁, 이근배, "단문 분할을 통한 명사구 색인 방법", 정보과학회논문지(B), 제 24 권, 제 3 호, pp.302-311, 1997

[3] 김미영, 강신재, 이종혁, "규칙과 어휘정보를 이용한 한국어 문장의 구묶음", 제 12 회 한글 및 한국어 정보처리 학술대회, pp.103-108, 2000

[4] 강인호, 전수영, 김길창, "최대 엔트로피 모델을 이용한 한국어 명사구 추출", 제 12 회 한글 및 한국어 정보처리 학술대회, pp.127-132, 2000

[5] 양재형, "규칙 기반 학습에 의한 한국어의 기반 명사구 인식", 정보과학회 논문지: 소프트웨어 및 응용, 제 27 권, 제 10 호, pp.1062-1071