구문분석과 공기정보를 이용한 개념 기반 명사구 색인 방법

이 현아, 이 종혁, 이 근배
포항공과대학교 전자계산학과

Concept-Based Method for Noun Phrase Indexing Using Syntactic Analysis and Co-occurence Information

Hyun-A Lee, Jong-Hyeok Lee, Geunbae Lee
Dept. of Computer Science, POSTECH

요 약 문

한국어에서의 명사구 색인을 위한 기존의 방법들은 주로 간단한 규칙을 이용하여 왔고 그 결과 문장에 존재하는 모든 명사구를 추출하지 못했다. 이를 해결하기 위하여 본 논문에서는 개념 기반 명사구 색인 방법을 제안한다. 하나의 문장은 하나 이상의 개념으로 이루어져 있으므로, 명사구 추출은 개념을 고려하여 이루어져야 바람직하다. 문장은 구문적으로 하나 이상의 내포문으로 이루어져 있다. 일반적으로 내포문 단위 내의 용어들이 나타내는 각각의 개념들은 서로 높은 연관성을 가진다. 그러므로 문장이 가지는 개념의 상이성을 내포문의 개념 상이성으로 축소할 수 있다. 문장을 내포문 단위로 분할하기 위하여 의존 문법을 기반한 구문분석과 공기정보를 이용한다. 특히 공기정보는 원거리 의존관계(long distance dependency)를 결정하여 한 내포문에 속함을 밝혀내는 데 도움을 준다. 이러한 내포문 내의 의존관계를 이용하여 명사구를 추출한다.

1. 서론

문서표현(text representation)의 한 부분으로 구(phrase)의 사용은 정보검색 분야에서 일찍부터 연구되어져 왔다[9]. 문서의 구조를 반영하는 구를 색인어로 사용함으로써 색인어의 특정성(specificity)을 향상시키고 결과적으로는 문서표현의 질을 높일 수 있을 것이다. 한국어 문서에서는 문서의 내용을 나타내는 것이 주로 명사이고 따라서 명사구 단위의 색인이 필요하다. 한국어에서의 기존의 명사구 색인 방법은 주로 간단한 규칙을 이용하여 왔다[1,2]. 이렇게 특정 구문패턴을 이용하는 방법은 명사구에 대한 구문 패턴을 모두 찾아내지 못하므로 문서에 존재하는 모든 명사구를 추출해 내지 못하는 단점이 있다. 이를 해결하기 위하여 본 논문에서는 구문분석과 공기정보를 이용하는 개념 기반 명사구 색인 방법을 제안한다. 하나의 문장은 대부분 구문적으로 하나 이상의 내포문으로 이루어진다. 각각의 내포문은 큰 문장 속의 한 성분으로 역할한다[4]. 내포문 내의 용어들이 연관되어 큰 문장의 한 성분이 되는 것이다. 따라서 내포문 내의 용어들은 서로 높은 연관성을 가지므로 각각의 용어가 나타내는 개념의 연관성도 높다.

의미적으로 문장에는 하나 이상의 개념이 존재하므로 문장을 표현하는 색인은 개념을 고려하여 이루어져야 한다. 문장 내의 개념들은 서로 높은 연관성을 가지는 것끼리 그룹화되어 있다. 본 논문에서는 이러한 개념 그룹을 구문적 그룹인 내포문과 동일하게 봄으로써 문장 내의 개념 그룹의 상이성을 내포문의 개념 상이성으로 축소한다.

문장을 내포문 단위로 분할하기 위하여 의존 문법을 기반으로 한 구문분석과 조사의 격과 용언 간의 공기정보를 이용한다.

명사구는 분할된 내포문 내의 의존관계를 이용하여 추출된다.

2. 개념 기반 명사구 색인 방법

2.1 내포문 분할

내포는 한 문장이 다른 문장의 한 성분으로 포함되는 현상을 뜻한다. 어떤 문장이 다른 문장의 한 성분으로 안겨 있는 것을 내포문이라고 한다[3]. 이러한 내포문 내의 용어들은 문장에서 개념적으로 서로 높은 연관성을 가지면서 큰 문장의 한 성분이 된다.

서술어가 되는 용언은 그 종류에 따라서 완전한 문장을 위해 필요로 하는 격이 다르다. 이를 자리수 정보라고 한다[4]. [예문1]과 같이 '먹다, 던지다' 와 같은 용언은 목적어가 필수성분이다.

[예문1]

ㄱ. 아이들이 돌을 던졌다.

ㄴ. 코끼리는 풀을 먹는다.

그러나 [예문2]의 ㄴ에서 보듯이 앞뒤 문맥에 의해 용언의 필수격이 생략되어도 뜻이 통하는 문장이 존재할 수 있다.

[예문2]

ㄱ. 순이는 동생에게 장난감을 주었다.

ㄴ. 순이는 장난감을 주었다.

이렇듯 필수격의 생략이 가능하므로 문장에서 어떤 격을 필수성분으로 요구하는 용언의 수가 그 격을 가진 어절보다 많을 때 격과 용언 사이의 의존관계를 설정하기 어렵다. 이 때 조사의 격과 용언 간의 공기관계를 이용하여 의존관계를 결정하고자 한다.

기존의 구문분석 기법은 지배소가 될 수 있는 모든 어절과 의존관계를 찾기 때문에[8] 많은 파스트리가 생성되어 이를 그대로 색인 시스템에 적용할 경우 효율성이 아주 떨어지게 된다. 따라서 본 논문에서는 내포문 단위의 분할을 위한 새로운 알고리즘으로 shallow parsing을 수행한다.

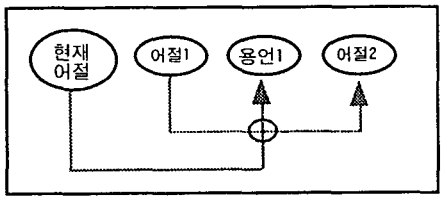
입력 문장을 내포문 단위로 분할하는 과정은 다음과 같다.

① 문장의 마지막 어절부터(right-to-left)[8] 의존문법을 기반한 shallow parsing을 시작한다. 각 어절에 대해 ②에서 ⑤까지 반복한다.

② 현재 어절이 용언이 아니면 오른쪽으로 인접한 어절과 의존관계를 찾는다.

③ 현재 어절이 용언이면 내포문의 지배소로 설정한다. 단, 관형어인 경우는 오른쪽 인접 어절들 중 의존관계가 있는 가장 가까운 어절과 의존관계를 설정한다.

④ ②에서 의존관계가 없으면 현재 어절의 오른쪽에 위치한 용언들과 각각 의존관계를 찾는다. 이 때 각 용언의 자리수 정보를 이용하여 현재 어절과 의존관계가 있는 용언을 하나 선택한다. 또한 투영성의 규칙(projection rule)을 지키기 위해 [그림1]에서와 같이 현재 어절과 인접한 어절의 지배소보다 왼쪽에 위치하는 용언과는 의존관계를 찾지 않는다.



[그림1] 투영성의 규칙

⑤ ④에서 현재 어절이 둘 이상의 용언과 의존관계가 있으면 조사의 격과 용언과의 공기정보를 이용하여 후보 용언 중 하나를 선택한다. 이 때 공기정보의 적용 알고리즘은 [그림2]와 같다.

```
if (I(c, y1) / val(y1)) > (I(c, y2) / val(y2))
then select(c, y1);
else if (I(c, y1) / val(y1)) < (I(c, y2) / val(y2))
then select(c, y2);
else if (I(c, y1) / val(y1)) = (I(c, y2) / val(y2))
then select(max(I(c, y1), I(c, y2)));
where
        I(c, y): 격 c와 용언 y의 상호정보,
        val(y): 용언 y의 자리수
```

[그림2] 공기정보 적용 알고리즘

공기하는 격과 용언의 상호 연관성을 객관적으로 산출하기 위해 정보 이론적 개념인 상호정보(mutual information)를 [그림3]과 같이 이용한다[10]. [그림2]에서 격과 용언의 상호정보를 용언의 자리수로 나누므로 자리수가 적은 용언이 더 높은 정확성을 가진다.

$$I(c, y) = \log_2 \frac{N \cdot f(c, y)}{f(c) \cdot f(y)}$$

where

N: 코퍼스의 크기,
f(c, y): 격 c와 용언 y가 인접하여 나타난 빈도,
f(c), f(y): frequency

[그림3] 상호정보

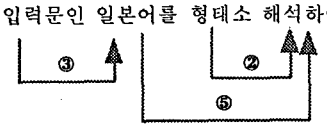
위의 내포문 분할 방법에 따른 예를 보면 다음과 같다.

[예문3]

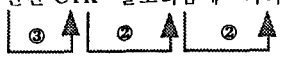
입력문인 일본어를 개선된 CYK 알고리즘에 따라 형태소 해석하여 접속이 가능한 모든 해석 결과를 얻는다.

[내포문 분할]

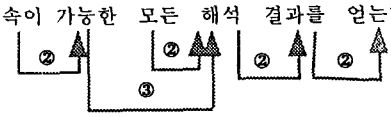
ㄱ. 입력문인 일본어를 형태소 해석하여



ㄴ. 개선된 CYK 알고리즘에 따라



ㄷ. 접속이 가능한 모든 해석 결과를 얻는다



이 때, 어절 '일본어를' 이 원문에서 용언 '따라' 와 '해석하여' 의 필수격으로 모두 가능하다. 이 중 하나를 선택하기 위하여 목적격 조사와 두 후보 용언 간의 공기정보를 이용하였고, 그 결과 '해석하여' 가 선택되어 <내포문 ㄱ>이 분할되었다. 여기서 숫자는 위에서 기술한 내포문 분할 과정 중 의존관계가 설정된 단계를 의미한다.

2.2. 내포문 내의 의존관계를 이용한 명사구의 추출

내포문 내의 의존관계를 이용한 명사구 추출은 명사구의 생성, 명사구의 삭제, 그리고 명사구의 통합(merge)의 세 단계로 이루어진다.

2.2.1 명사구의 생성

성분명사의 결합도가 높은 명사구일수록 성분명사가 각각 나타내는 개념의 연관성은 높아진다.

본 논문에서는 명사를 포함하는 어절 간의 의존관계 레이블을 통해 성분명사의 결합도를 나타낸다. 의존관계의 레이블의 결합도 순위는 [표1]과 같이 정한다.

1	L1: 조사가 생략된 명사와 인접한 명사
2	L2: 관형격조사 결합명사와 인접한 명사
3	L3: 목적격, 주격조사 결합명사와 서술형명사
4	L4: 관형화된 내포문의 명사와 피수식 내포문의 명사

[표1] 의존관계 레이블에 따른 결합도 순위

한 내포문 내에서 명사구의 생성 조건은 다음과 같다.

```
if (L(E1, E2) ∈ {L1, L2, L3, L4})
then PHRASE(N1, N2);
where
L(E1, E2): 어절 E1, E2의 의존관계 레이블,
N1, N2: 어절 E1, E2에 포함된 명사,
PHRASE(N1, N2): 두 명사를 구로 생성
```

이렇게 생성된 각 명사구는 두 개의 성분명사로 구성된다. [예문3]에서 다음과 같이 명사구가 생성된다.

[명사구 생성]

ㄱ. [L1]/형태소 해석/ [L3]/일본어 해석/ [L4]/입력문 일본어/

ㄴ. [L1]/CYK 알고리즘/ [L4]/개선 CYK/

ㄷ. [L1]/해석 결과/ [L4]/접속 해석/

2.2.2 명사구의 삭제

어떤 문장에서 수식어(modifier)의 역할만 하는 명사가 잘못된 의존관계에 의해 피수식어(modifee, head)가 되거나 그 반대가 되어 명사구로 형성될 수 있다. 이런 명사구를 삭제하기 위하여 본 논문에서는 다음과 같은 규칙을 도출한다.

```
if (LR(P1) > LR(P2) & LR(P2) < LR(P3) &
CN(P1) ≈ CN(P2) & CN(P2) ≈ CN(P3))
then REMOVE(P2);
where
LR(P): 명사구 P의 레이블 순위(label rank),
CN(P): 명사구 P의 성분명사(constituent noun),
≈: 두 명사구의 성분명사 중 하나가 일치,
REMOVE(P): 명사구 P을 삭제.
P1, P2, P3: 한 내포문에 속한 명사구
```

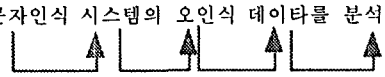
적용된 예를 보면 다음과 같다.

[예문4]

문자인식 시스템의 오인식 데이타를 분석한다.

[내포문 분할]

ㄱ. 문자인식 시스템의 오인식 데이타를 분석한다.



[명사구 생성]

ㄱ. [L1]/문자인식 시스템/ [L2]/시스템 오인식/
[L1]/오인식 데이타/ [L3]/데이타 분석/

[명사구 삭제]: /시스템 오인식/

ㄱ. [L1]/문자인식 시스템/ [L1]/오인식 데이타/
[L3]/데이타 분석/

삭제된 명사구인 '시스템 오인식'은 원래 피수식어인 '시스템'이 수식어가 되고, 수식어인 '오인식'이 피수식어가 되어 생성된 것이다.

2.2.3 명사구의 통합(merge)

두 개의 성분명사로 이루어진 명사구들이 새로운 명사구로 통합되어 더 자세한(specific) 개념을 나타낼 수 있다. 본 논문에서는 두 명사구가 통합될 수 있는 조건을 도출하여 이를 기반으로 명사구 통합을 수행한다.

두 명사구 P1, P2이 같은 내포문에 존재하고, 각 명사구의 성분명사가 다음과 같을 때,

P1 = {N1, N2} P2 = {N3, N4}

```
[조건1]
    if (N2 = N3)
    then MERGE1(P1, P2);
    결과: /N1 N2 N3/

[조건2]
    if (N2 = N4 & LR(P1) > LR(P2))
    then MERGE2(P1, P2);
    결과: /N3 N1 N2/
```

[조건2]는 결합도가 더 강한 성분명사를 인접시킴으로써 문장에 존재하는 개념간의 연관성을 더 잘 표현할 수 있도록 한다.

[예문3]에서의 통합 결과를 보면 다음과 같다.

[명사구의 통합]

ㄱ. /일본어 해석/ + /형태소 해석/
[조건2]/일본어 형태소 해석/
/입력문 일본어/ + /일본어 해석/
[조건2]/입력문 일본어 해석/

ㄴ. /개선 CYK/ + /CYK 알고리즘/
[조건1]/개선 CYK 알고리즘/

ㄷ. /집속 해석/ + /해석 결과/
[조건1]/집속 해석 결과/

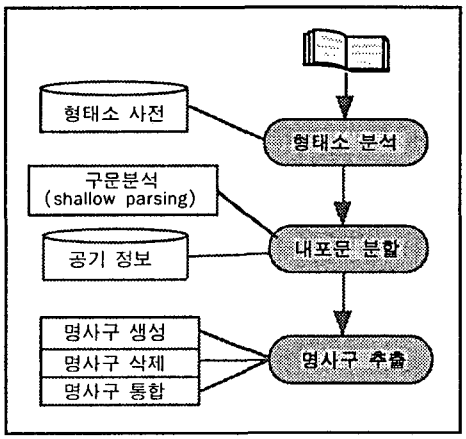
3. 구현 및 실험

본 논문에서 제안한 방법으로 구현된 색인 시스템의 전체 구성도는 [그림4]와 같다.

형태소 분석기에서는 형태소 분할 방법으로 tabular parsing 방법을 사용하였고, 원형 복원 방법은 사전기반 모델과 규칙기반 모델을 혼용한 방법을 사용하였으며, 형태소 간의 결합 타당성을 검사하는 방법으로 접속정보 모델을 사용하였다[6, 7]. 형태소 해석 단계에서 쓰이는 형태소 사전에는 상용단어 3만개와 대상문서의 영역을 고려하여 전산 관련 용어 1300단어가 등록되어 있다.

3.1 실험 방법

본 논문에서 제안한 명사구 색인 방법의 실험을 위해 색인 시스템의 대상문서로 한국통신에서 구축한 시험용 데이터 모음인 KTSET을[5] 사용하여 실험하였다.



[그림4] 시스템 구성도

3.2 실험결과와 분석

본문에서 기술한 [예문3]에 대한 실행결과에서도 알 수 있듯이 문장 내에서 원거리 의존관계를 가지는 두 어절을 하나의 내포문으로 분할함으로써 기존의 규칙을 기반한 방법으로는 추출되지 못하는 주요 명사구인 **'일본어 형태소 해석'** 과 **' 일본어 해**

석' 이 본 논문에서 제안한 방법으로 추출되었다. 그러나 본문의 [표1]에서 제시된 의존관계 레이블로 명사구를 생성하는데 있어 [예문3]의 실행 결과에서 '접속 결과', '접속 해석 결과', '개선 CYK', '입력문 일본어'라는 부적절한 명사구가 생성되는 오류가 있다. 따라서 생성된 명사구들에 대한 필터링이 필요하다. 필터링 방법으로 성분명사의 상호 연관성을 측정하여 일정 임계값을 넘지 못하는 명사구는 삭제하는 방법이 있다. 본 논문에서 구현한 색인 시스템의 평가를 위해서는 본 시스템에서 추출한 색인어와 주제 전문가에 의해 추출된 색인어의 비교가 이루어져야 한다. 그러나 실제 이러한 비교, 평가가 어렵기 때문에 아직 정확한 평가를 하지 못하였다.

4. 결론

본 논문에서는 개념 기반 명사구 색인 방법을 제안하였다. 개념을 고려하기 위하여 구문분석과 공기정보를 이용하여 문장을 내포문 단위로 분할하고 각 내포문 내의 의존관계를 이용하여 명사구를 추출하였다. 이 때 공기정보는 조사의 격과 용언 간의 공기정보로서 원거리 의존관계를 가지는 두 어절이 한 내포문에 속함을 밝혀내는 데 도움을 준다. 본 논문에서 제안한 방법으로 색인작업을 한 결과, 규칙을 기반한 방법에서는 추출되지 못했던 명사구를 추출할 수 있었다. 앞으로 객관적 자료로 충분한 공기정보의 추출을 위해 좀더 많은 코퍼스가 필요하고 모든 용언에 대한 결합가 정보를 사전에 등록하는 작업이 이루어져야 할 것이다. 그리고 내포문 내의 의존관계를 이용한 명사구 추출에 있어서 좀더 많은 연구가 필요하다. 또한 부적절한 명사구의 제거를 위해 성분명사의 상호 연관성을 측정하는 기법으로서 성분명사의 공기정보를 이용하고 추출된 명사구에 대한 색인어 가중치로서 문서 내 출현빈도와 역문헌빈도를 고려하는 것이 필요하다.

5. 참고 문헌

[1] 김민정, 권혁철, "한국어 특성을 이용한 자동 색인 기법," 한국정보과학회 가을학술발표논문집, 19권, 2호, pp. 1005~1008, 1992

[2] 김판구, 조유근, "상호 정보에 기반한 한국어 텍스트의 복합어 자동색인," 한국정보과학회논문지, 제21권, 제7호, pp. 1333~1340, 1994

[3] 이주행, "현대국어문법론," 대한교과서주식회사, 1992

[4] 남기심, 고영근, "표준 국어문법론," 탑출판사, 1985

[5] 김재군, 김영환, 김성혁, 한국통신 S/W연구소 인공지능팀, 숙명여대 문헌정보학과, "한국어 정보검색연구를 위한 시험용 데이터 모음(KTSET) 개발," 제6회 한글 및 한국어정보처리 학술발표논문집, pp. 378~385, 1994

[6] 이은철, 이종혁, "계층적 기호 접속정보를 이용한 한국어 형태소 분석기의 구현," 제4회 한글 및 한국어정보처리 학술발표논문집, pp. 95~104, 1992

[7] 홍남희, 이은철, 이종혁, "한국어의 형태론적 품사분류," 포항공과대학교 지식 및 언어공학 연구실 Technical Report, TM-93-001, 1993

[8] Changhyun Kim, Jae-Hoon Kim, Jungyun Seo, Gil Chang Kim,"A Right-to-Left Chart Parsing with Headable Paths for Korean Dependency Grammar," CPCOL, Vol. 8, Supplement, pp. 105~118, 1994

[9] W. Bruce Croft, Howard R. Turtle, and David D. Lewis, "The Use of Phrases and Structured Queries in Information Retrieval," Proc. of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, Illinois USA, pp. 32~45, 1991

[10] Kenneth W. Church and Patrick Hanks, "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, Vol. 16, No. 1, pp. 22~29, 1990