효율적인 색인어 추출을 위한 복합명사 분석 방법

장 동 현, 맹 성 현
충남대학교 컴퓨터과학과

A Korean Compound Noun Analysis Method for Effective Indexing

Dong Hyun Jang, Sung Hyun Myaeng
Dept. of Computer Science, Chungnam National University

요 약

정보 검색 기술은 적용 분야, 질의어, 데이터가 달라질 경우, 결과 또한 달라질 수 있음을 최근의 연구 결과로부터 알 수 있다. 사용되는 언어에 따라서도 고유한 문제가 제기될 수 있는데, 특히 한국어의 경우 복합명사는 명사끼리의 조합이 자유롭고 길이에 제한이 없기 때문에 이를 단위 명사로 분할하는 작업이 어렵다. 또한 영어와는 달리 복합명사가 문서 내에서 많은 부분을 차지하며 문서의 내용을 대표하는 경우가 많이 있기 때문에, 정보 검색 기술을 한국어에 적용하기 위해서는 수정, 보완하는 노력이 필요하다. 본 연구에서는 어휘에 관한 사전 및 코퍼스 정보를 트라이(trie)에 저장한 후 어휘들간의 공통 부분에 더미 노드(dummy node)를 삽입하여 복합명사를 단위 명사로 분할하는 기법을 제시하였다.

1. 서 론

다양한 코퍼스와 질의어를 이용하여 정보검색 시스템을 평가한 결과, 지금까지 잘 알려진 정보검색 기술이 항상 효과적이지는 않다는 사실이 연구 결과로 밝혀졌고[1] 사용되는 언어에 따라 새로운 색인 방법이 개발되어야 한다. 즉, 질의어 처리나 인덱싱(indexing) 기술을 포함한 대부분의 정보검색 기술이 새로운 언어에 적용되기 위해서는 그 언어의 특수성을 고려할 필요가 있다.

한국어의 경우 명사끼리의 조합이 자유롭고 길이에 제한이 없기 때문에 이를 단위 명사로 분할하는 작업이 어렵다. 또한 문서 내에서 복합 명사가 차지하는 중요도가 크기 때문에 정보검색에 있어서 이에 대한 정확한 분석이 요구된다. 예를 들어 복합명사 "대학생선교회"는 "대학생"과 "선교회"로 분리되어야만 한다. 그러나 명사의 한계가 명확하지 않기 때문에 "대학", "생선", "교회"로 분석할 수 있다. 이러한 경우 잘못된 색인어를 추출하게 되므로 사용자 요구와는 다른 문서의 내용을 제공할 수 있는 오류가 발생한다.

복합명사로부터 단위 명사를 분할하는 문제를 해결하기 위하여 모든 복합 명사를 포함하는 사전을 사용하는 것도 이론상으로 가능하다. 그러나 이러한 접근 방법은 미등록어 처리에 한계가 있고 거의 완벽한 사전을 요구하므로 이용하기가 힘들다. 일반적으로 한국어 처리에서 인덱스어 추출을 위해 사용하는 기법은 형태소 분석기를 사용하는 것인데, 이를 이용할 경우 복합어나 어휘 사전에 등록되어 있지 않은 미등록어의 경우 처리하기 어려운 문제가 있다. 본 논문에서는 이러한 문제점을 해결하기 위해 복잡한 형태소 분석기를 이용하지 않고도 효율적으로 복합명사를 분리하여 색인어를 추출할 수 있는 방안을 제시한다.

2. 관련 연구

한글 텍스트에서 미등록 된 복합명사 분석과 인식을 위해 복합명사 사전으로부터 추출한 통계적 정보를 사용해서 텍스트 내의 복합 명사를 분할하는 방법[2]이 있다. 이 방법은 X'이론에 근간으로 하여 의미적 중의성을 해결하고자 했다.

또한 명사간의 결합 선호도에 대한 통계 정보를 이용한 연구[3]도 있다. 그리고 복합명사

사이의 의미적 관계를 이용하거나 시소러스 정보를 이용하는 등의 의미적 정보를 사용[4]한 연구도 있었다.

본 연구에서는 코퍼스로부터 추출한 정보를 직접적으로 사용하고 분석대상이 되는 문서의 지역 정보를 이용하여 수작업을 통해 구축되는 사전이나 시소러스의 필요성을 최소화하는 방법을 제안한다.

3. 코퍼스 패턴과 학습을 통한 명사 추출

본 장에서는 어휘로부터 명사를 인식하고 분할하기 위해 분석되어질 텍스트와 관계된 코퍼스로부터 어휘의 패턴을 학습한 후 이를 이용하여 어휘를 분석하는 방법을 기술한다. 제시한 기법에 대한 주요 원칙은 수작업으로 사전을 작성하는 작업을 최소화 시키고 사전과 복잡한 문법 규칙을 적용하는 형태소 분석기를 사용하지 않는 것이다.

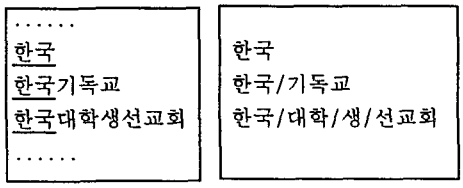
3.1 복합명사 분할 방법

복합명사 분석을 위한 처리는 두 단계, 즉 어휘의 사용 패턴을 학습해서 명사를 추출하여 트라이 자료 구조에 저장하는 학습단계와 트라이를 검색하고 어휘로부터 명사를 추출하기 위해 조사나 어미로 구성된 사전을 검색하는 적용단계로 구성된다.

학습 단계의 주 목적은 복합명사 분리에 필요한 정보를 추출하여 사전 엔트리를 저장하고 검색할 때 자주 사용하는 트라이를 구축하는 것이다. 트라이는 태깅 된 코퍼스와 일반 코퍼스로부터 추출된 명사로 구성되며, 복합명사의 분할 위치와 코퍼스 내에 있는 명사들의 정보를 갖게 된다[5]. 학습은 다음과 같은 단계로 이루어진다.

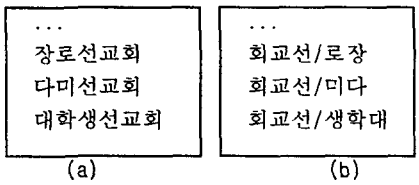
1. 텍스트 코퍼스로부터 단순 명사 또는 복합명사인 단어의 리스트를 추출한다. 본 연구에서는 복합명사를 구성하고 있는 단위 명사를 분석하는 것이 주 목적이므로 태깅 된 코퍼스로부터 명사만을 추출한다.

2. 추출한 명사 리스트로부터 공통 부분을 분리한다. 한 개 이상의 음절로 구성된 공통된 패턴은 명사 후보로서 고려되어지는데 단어의 시작 부분부터 시작되는 정방향 리스트와 끝부분부터 시작되는 역방향 리스트를 구축한다. 예를 들어 [그림 1]의 (a)는 단어의 각 음절이 원래의 순서와 같은 정방향 리스트이다. "한국"은 공통된 부분이므로 명사 후보로 선택되게 되고, 이 알고리즘을 반복적으로 적용하면 "기독교", "대학", "대학생", "선교회"를 얻을 수 있다. 결과적으로 (b)와 같은 정보를 저장하게 된다.



[그림 1] 정방향 리스트

그러나 [그림 2]의 (a)와 같은 경우 "선교회"가 공통된 단어이지만, 정방향으로만 공통된 단어를 찾을 경우 실패하게 된다. 따라서 (b)와 같이 원래의 단어를 거꾸로 하여 역방향 리스트를 만든 후 공통된 단어를 추출할 수 있도록 한다.

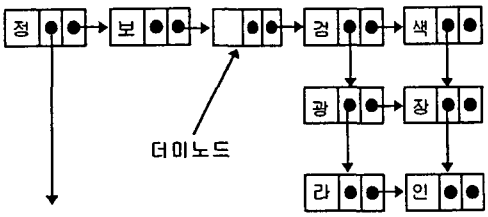


[그림 2] 역방향 리스트

(b)로부터 공통된 패턴인 "회교선", 즉 정방향 단어인 "선교회"를 얻을 수 있다.

3. 코퍼스로부터 추출한 명사 리스트는 정방향과 역방향 두개의 트라이에 별도로 저장되며 질의어 뿐만 아니라 문서내의 어휘를 처리하기 위한 사전으로서도 사용되어진다. 각 트라이는 [그림 3]과 같이 명사 다음에 더미 노드를 삽입함으로써 복합 명사를 구성하는 단위 명사의 위치에 대한 정보를 포함하게 된다.

두 번째 단계인 적용 단계는 기본적으로 트라이를 사용하여 어절로부터 명사를 추출하는 단계인데, 조사/어미 사전은 처리 될 어휘에 포함되어 있는 조사나 어미를 제거하기 위해 사용되고, 이 때 최장 일치를 적용한다.



[그림 3] 더미 노드가 삽입된 트라이

어휘 분석시에는 정방향, 역방향, 조사/어미 사전 검색이 이루어지는데, 검색 순서에 따라서 다양한 결과를 얻을 수 있다. 예를 들어 "대학생선교회의"는 1)"대학생+선교회+의"와 2)"대학생선교+회의"로 분석될 수 있다. 첫 번째는 정방향 검색이 먼저 수행되어지거나 조사 검색과 역방향 검색이 수행된 경우이다. 두 번째는 정방향으로만 검색이 성공되었거나, 역방향 검색이 수행된 결과로 "선교"는 역방향 사전에 없기 때문에 분석되지 않았을 경우이다.

forward	backward	endings	규 칙
match	match	match	1
match	match	no match	1
match	no match	match	2
match	no match	no match	1
no match	match	match	3
no match	match	no match	4
no match	no match	match	5
no match	no match	no match	6

[표 1] 어휘 분석 형태

```
1. USE forward match result
2. IF NUM(original) <= 3
      USE original
   ELSE
      USE forward match result
3. IF NUM(remainder of backward match)==1
      USE ending match result
   ELSE
      USE backward match result
4. IF NUM(remainder of backward match)==1
      USE original
   ELSE
      USE backward match result
5. IF NUM(remainder of ending match) >= 4
      APPLY backward match
   ELSE
      USE ending match result
```

[그림 4] 휴리스틱 규칙

정방향 트라이, 역방향 트라이, 조사/어미 사전을 적용하는 순석에 따라 하나의 어휘에 대해 가능한 분석 결과는 [표 1]에서 보듯이 8가지이며 각 경우에 대한 결과를 분석해서 얻은 휴리스틱한 규칙은 [그림 4]와 같다.

3.2 실험 결과와 분석

제안한 기법의 타당성을 측정하기 위하여 학습 데이터와 테스트 데이터를 구분하여 실험을 했다.

학습, 즉 트라이를 구축하기 위해 사용된 데이터는 세 가지인데, 태깅된 코퍼스로부터 적어도 하나의 명사를 포함하고 있는 어절로 구성된 사전, KAIST에서 개발한 형태소 분석기에서 사용한 명사 사전, 그리고 이 두 가지를 혼합한 데이터이다.

테스트 데이터는 세 가지로, 그 중 두 가지는 태깅 된 코퍼스로부터 추출한 것인데 하나는 트라이 구축시 사용한 코퍼스 데이터이며 다른 하나는 코퍼스 중 트라이 구축시 사용하지 않았던 데이터이다. 다른 하나의 테스트 데이터는 코퍼스에는 포함되지 않은 분야로서 컴퓨터와 정보 과학 문서 데이터베이스로부터 추출한 데이터이다. 실험의 목적은 제안한 기법의 일반적인 효과 뿐만 아니라 사전 정보보다는 코퍼스 정보가 좀 더 유용하다는 것을 테스트하기 위한 것이다.

착안한 알고리즘을 색인을 목적으로 한 복합명사 분리에 사용하는 것이기 때문에 테스트에 있어서도 어휘에 적어도 하나의 명사가 있는 것만을 포함시켰다. 기본적으로 실험은 알고리즘이 얼마나 잘 명사를 추출하고 분리하는지 측정하며, 동사에 명사가 포함되어 있지 않으면 통계에 포함시키지 않았다. 결과는 [표 2]와 같으며 정확도를 백분율로 표현한 것이다.

학습 자료 / 실험 데이터	dictionary	corpus	dictionary + corpus
trained	93.12%	97.66%	96.19%
untrained	85.54%	87.75%	89.46%
kT-SET	89.04%	85.43%	87.55%

[표 2] 실험 결과

테이블에서 열(column)은 트라이 구축시 사용한 데이터이며, 행(row)은 테스트 데이터를 나타낸다. "trained"는 트라이 구축시 사용한

데이터이며, "KTSET"은 한국통신에서 구축한 컴퓨터와 정보 과학 분야의 문서 집합이다.

학습 데이터의 경우, 코퍼스로부터 구축한 트라이가 가장 좋은 성능을 보이며 사전으로부터 구축한 트라이는 가장 나쁜 결과를 보여주고 있다. 이러한 결과는 명사를 추출하고 분할할 때 발생하는 모호성의 많은 부분이 지역 정보에 의해 해결될 수 있다는 가정을 뒷받침한다. 코퍼스 사전에 일반 사전이 추가되었을 경우 성능이 감소되었다는 것도 특기할 만한 결과이다.

두 번째 행은 학습하지 않은 데이터를 테스트한 경우로 이 데이터는 트라이 구축시 사용했던 코퍼스의 일부분이다. 이 경우도 일반 사전을 사용한 경우보다 코퍼스 정보를 사용하여 더 좋은 결과를 얻을 수 있음을 알 수 있지만 일반 사전을 같이 사용한 경우 결과의 향상을 얻을 수 있음을 알 수 있다. 이러한 결과는 코퍼스에는 없던 새로운 정보를 사전이 제공했다고 볼 수 있다.

실험 데이터가 트라이의 것과는 완전히 다를 경우(3번째 행), 일반 사전을 사용하여 가장 좋은 결과를 얻었다. 흥미로운 사실은 코퍼스 정보가 추가되어질 때 정확도는 감소된다는 것이다.

실험 분석 결과를 정리하면 다음과 같다. 1)명사 추출과 분할시 동일 분야의 코퍼스 정보를 사용하여야 한다. 2)코퍼스 정보는 분석 대상 데이터가 같은 분야일 경우 유용하다. 3)코퍼스가 완전히 다른 분야일 경우 일반 범용 사전만큼 유용하지 않고 사전 정보가 추가 되어질 경우 성능이 감소될 수 있다. 4)일반 사전으로부터 얻은 정보는 같은 분야의 코퍼스로부터 얻은 신뢰할 만한 정보가 있을 경우 나쁜 영향을 줄 수 있다.

4. 결 론

본 논문에서는 한국어에 특수하게 나타나는 복합명사를 인식 하기 위하여 코퍼스 정보를 사용하는 방법을 제시하였다. 어휘의 정방향과 역방향 분석은 두개의 트라이에 저장이 되며 명사를 인식하고 분할하기 위한 실제 검색 시에는 조사 사전과 함께 검색이 된다. 제안한 방법의 신뢰성과 명사를 인식하고 분할 하는데 있어서는 같은 분야의 코퍼스 정보를 이용하는 것이 가장 중요할 것이라는 가정을 검증하기 위하여, 각각 세 가지의 트라이와 테스트 데이터를 사용하여 실험하였다.

실험 결과 분석될 어휘가 나타나는 문서의 정보를 이용하는 것이 가장 효과적이라는 것을 알 수 있었다. 분석될 어휘가 있는 문서의 분야와 같은 분야의 코퍼스가 사전을 이용하는 것보다 좋은 결과를 얻을 수 있었지만 지역 정보를 이용하는 것보다는 좋지 않았다. 그러나 분석 대상이 되는 문서의 분야가 달라질 경우는 사전을 이용하는 것이 코퍼스를 사용하는 것보다 좋은 결과를 보였다.

향후 연구로는 복합명사 분할시 중요한 역할을 하게 되는 트라이 구축시 확률 정보를 추가하여 좀 더 정확하게 복합명사를 인식하는 방법이 고려되고 있다. 또한 실제 분석될 어절이 있는 문서 자체의 지역정보를 사용할 수가 있는데 이에 대한 결과는 이미 부분적으로 검증이 되어있다.

[참고문헌]

[1] Myaeng, S. & Jang, D.(1996), "On Language Dependency in Indexing", Proc. of the Workshop on Information Retrieval with Oriental Languages, pp.17-23.

[2] 윤보현, 임희석, 임해창(1995), "통계 정보를 이용한 한국어 복합 명사의 분석 방법", 제 22 회 한국어정보과학회 봄 학술발표 논문집, pp.925-928.

[3] 김판구(1994), "한국어 형태소 분석을 위한 복합 명사의 인식 방법", 서울대학교 박사 학위논문.

[4] 최기선(1993), "한국어에서의 복합 명사구 인식에 대한 연구", 최종 연구보고서, 한국 전자통신연구소.

[5] 맹성현(1995), "대용량 통신처리에서의 다자간회의를 위한 멀티캐스팅 연구", 최종 연구보고서, 한국 전자통신연구소.