학습데이터를 이용하여 생성한 규칙과 사전을 이용한 명사 추출기

장 동 현, 맹 성 현
충남대학교 컴퓨터과학과
{dhjang, shmyaeng}@cs.cnu.ac.kr

A Noun Extractor based on Dictionaries and Heuristic Rules Obtained from Training Data

Dong-Hyun Jang, Sung Hyon Myaeng
Department of Computer Science, Chungnam National University

요 약

텍스트로부터 명사를 추출하기 위해서 다양한 기법이 이용될 수 있는데, 본 논문에서는 학습 데이터를 이용하여 생성한 규칙과 사전을 이용하는 단순한 모델을 통해 명사를 효과적으로 추출할 수 있는 기법에 대하여 기술한다. 사용한 모델은 기본적으로 명사, 어미, 술어 사전을 사용하고 있으며 명사 추정은 학습 데이터를 통해 생성한 규칙을 통해 이루어진다. 제안한 방법은 복잡한 언어학적 분석 없이 명사 추정이 가능하며, 복합명사 사전을 이용하지 않고 복합명사를 추정할 수 있는 장점을 지니고 있다. 또한, 명사추정의 주 요소인 규칙이나 사전 등록어의 추가, 갱신 등이 용이하며, 필요한 경우에는 특정 분야의 텍스트 분석을 위한 새로운 사전의 추가가 가능하다. 제안한 방법을 이용해 "제 1 회 형태소 분석기 및 품사 태거 평가대회 (MATEC '99)" 의 명사 추출기 분야에 참가하였으며, 본 논문에서는 성능평가 결과를 제시하고 평가결과에 대한 분석을 기술하고 있다. 또한, 현재의 평가기준 중에서 적합하지 않은 부분을 규정하고 이를 기준으로 삼아 자체적으로 재평가한 평가결과를 제시하였다.

1. 서론

본 논문에서는 단순한 모델을 사용하여 효과적으로 명사 및 복합명사를 추출하는 방법을 기술한다. 제안한 모델의 주 사용 용도는 정보검색 분야이기 때문에 의미분석과 같은 복잡한 과정을 적용하지 않는다. 언어학적 측면에서 볼 때의 명사추출은 의미분석 과정을 적용하거나 형태소 분석기가 같은 언어적 도구를 이용하는 것이 바람직한 접근기법이라고 할 수 있다. 그러나, 정보검색 분야와 같이 대용량의 데이터로부터 중요 명사만을 추출할 경우에는 복잡한 언어지식을 요구하는 방법보다 영역에 관계없이 효율적으로 색인어를 추출하는 방법이 보다 실용적인 가치를 갖는다고 할 수 있다. 일반적인 형태소 분석기의 경우 가능한 모든 품사 후보를 생성하기 때문에 이로부터 명사를 추출하기 위해서는 태깅 시스템과 같은 또 다른 복잡한 도구를 사용해야 하는데, 이러한 도구를 구축하고 학습시키는데는 상대적으로 많은 노력과 시간이 요구된다.

제안한 시스템은 이러한 사항을 고려하여 정보검색 분야의 명사추정만을 주목적으로 하며, 최소한의 컴퓨터 자원을 사용하는 단순한 방법을 적용하고 있다. MATEC '99 의 평가 결과 참여한 시스템의 평균적인 성능을 보였으며, 특히 정보검색 분야의 색인에서 중요시되는 재현율에서 우수한 성능을 보였다.

본 논문의 구성은 2 장에서 명사 추정을 위해 사용하고 있는 접근방법을 자세히 설명하고, 제안한 방법으로 시스템을 구현하여 참가한 MATEC '99 대회의 평가 결과와 이에 대한 분석을 3 장에 기술한다. 4 장에서는 평가대회의 기준 중에서 본 시스템의 사용용도와 불일치 되는 부분을 기술하고 이 중 일부를 수정하여 재실험한 결과를 제시한다. 마지막으로 5 장에서는 결론 및 향후 연구사항을 기술한다.

2. 기본 모델

본 장에서는 어절로부터 명사를 인식하고 분할하기 위해 분석되어질 텍스트와 관계된 코퍼스로부터 어휘의 패턴을 학습한 후 이를 이용하여 어휘를 분석하는 방법[1]을 기술한다. 제시한 기법에 대한 주요 원칙은 수작업으로 사전을 작성하는 작업을 최소화 시키고 복잡한 문법 규칙을 적용하는 형태소 분석기를 사용하지 않는 것이다[2].

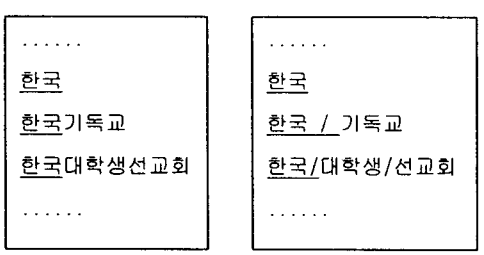
2.1 명사 추출 방법

명사 추출을 위한 처리는 두 단계, 즉 어휘의 사용 패턴을 학습해서 명사를 추출하여 트라이(trie) 자료 구조에 저장하는 학습단계와 트라이를 검색하고 어절로부터 명사를 추출하기 위해 조사나 어미, 술어로 구성된 사전을 검색하는 적용단계로 구성된다.

학습 단계의 주 목적은 명사 또는 복합명사 분리에 필요한 정보를 추출하여 사전 엔트리에 저장하고 트라이를 구축하는 것이다. 트라이는 태깅 된 코퍼스와 일반 코퍼스로부터 추출된 명사로 구성되며, 복합명사의 분할 위치와 코퍼스 내에 있는 명사들의 정보를 갖게 된다. 학습은 다음과 같은 단계로 이루어진다.

1. 텍스트 코퍼스로부터 단순 명사 또는 복합 명사인 단어의 리스트를 추출한다. 본 연구에서는 명사를 분석하는 것이 주 목적이므로 태깅 된 코퍼스로부터 명사만을 추출한다.

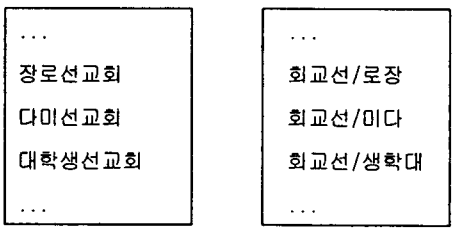
2. 추출한 명사 리스트로부터 공통 부분을 분리한다. 한 개 이상의 음절로 구성된 공통된 패턴은 명사 후보로서 고려되어지는데 단어의 시작부분부터 시작되는 정방향 리스트와 끝부분부터 시작되는 역방향 리스트를 구축한다. 예를 들어 [그림 1]의 (a)는 단어의 각 음절이 원래의 순서와 같은 정방향 리스트이다. «한국»은 공통된 부분이므로 명사 후보로 선택되게 되고, 이 알고리즘을 반복적으로 적용하면 «기독교», «대학», «대학생», «선교회»를 얻을 수 있다. 결과적으로 (b)와 같은 정보를 저장하게 된다.



[그림 1] 정방향 리스트

그러나 [그림 2]의 (a)와 같은 경우 «선교회»가 공통된 단어이지만 «장로»와 «다미»가 사전에 등록되어 있지 않으면 정방향으로만 공통된 단어를 찾을 경우 실패하게 된다. 이러한 문제를 해결하기 위하여 (b)와 같이 원래의 단어를 거꾸로

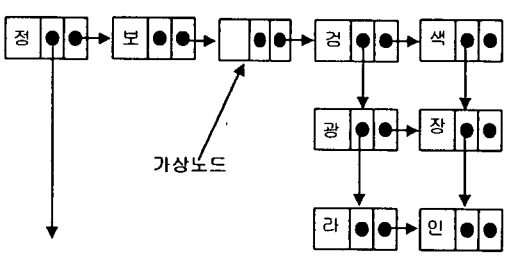
하여 역방향 리스트를 만든 후 공통된 단어를 추출할 수 있도록 한다. 즉, (b)로부터 공통된 패턴인 «회교선», 즉 정방향 단어인 «선교회»를 얻을 수 있다.



[그림 2] 역방향 리스트

3. 코퍼스로부터 추출한 명사 리스트는 정방향과 역방향 두개의 트라이에 별도로 저장되며 질의어 뿐만 아니라 문서내의 어절을 처리하기 위한 사전으로서도 사용되어진다. 각 트라이는 [그림 3]과 같이 명사 다음에 가상노드(dummy node)를 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, 술어사전은 처리 될 어절에 포함되어 있는 조사나 어미, 술어 등을 제거하기 위해 사용되고, 이 때 최장 일치를 적용한다.



[그림 3] 가상노드가 삽입된 트라이

어절 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.

[표 1] 어휘 분석 형태

정방향	역방향	조사	적용 규칙
0	0	0	1
0	0	X	1
0	X	0	2
0	X	X	1
X	0	0	3
X	0	X	4
X	X	0	5
X	X	X	6

정방향 트라이, 역방향 트라이, 조사/어미 사전을 적용하는 순서에 따라 하나의 어휘에 대해 가능한 분석 결과는 [표 1]에서 보듯이 8가지이며 각 경우에 대한 결과를 분석해서 생성된 규칙의 예는 다음과 같다.

규칙 1. 정방향 매칭 결과 사용

규칙 2. 원시어절이 3음절 이하면 원시어절을 사용하고 그렇지 않은 경우에는 정방향 매칭 결과 사용

규칙 3. 역방향 매칭하고 1음절이 남으면 조사매칭 결과를 사용하고, 그렇지 않은 경우에는 역방향 매칭 결과 사용

규칙 4. 역방향 매칭후 1음절이 남으면 원시어절을 사용하고, 그렇지 않은 경우에는 역방향 매칭 결과 사용

규칙 5. 조사매칭 후 3음절 이하가 남으면 조사매칭 결과를 사용하고 그렇지 않은 경우에는 역방향 매칭을 적용

규칙 6. 원시어절을 그대로 사용

이상과 같이 사전과 학습 데이터로부터 생성한 규칙을 이용한 단순한 명사 추출 방법이 언어

학적 분석을 통한 방법에 비해 우수한 성능을 발휘하기란 어렵다. 그러나, 이러한 접근 방법이 갖는 가장 큰 특징은 시스템 자체가 단순하여 구현이 용이하며, 사전추가나 규칙 변경 등의 관리가 쉽다는 것이다. 또한, 단위명사나 사전 등록어 사이의 공통 부분을 기준으로 복합명사 추정을 할 수 있는 장점을 갖고 있다.

앞서 기술한 바와 같이 본 시스템은 언어학적 분석을 적용하지 않기 때문에 "쌈질"과 같은 준말을 원형 상태로 복원하여 추출할 수는 없다. 그러나, 명사 추출기의 주 사용 용도가 정보검색 분야로 사용자가 준말이 포함된 질의를 할 수 있기 때문에 이를 원형으로 복원하지 않은 것을 오분석으로만 판단하기에는 모호성이 존재한다. 다른 문제점으로는 술어형태의 어절에 대한 분석이 어렵다는 것이다. 술어의 경우 다양한 변형이 가능하기 때문에 오분석이 발생할 확률이 가장 크므로 앞으로 보완해야 할 부분이다.

3. MATEC '99 평가

본 장에서는 MATEC '99 평가 대회를 위해 사용한 시스템의 구현 환경 및 시스템 정보를 기술하고, 평가 결과를 보여주고 있다.

3.1 시스템 환경

MATEC '99 의 평가를 위해 사용한 시스템은 SUN Ultra Sparc 1 으로 CPU 처리 속도는 167MHz, 메모리는 128M 로, 이 환경에서 명사 추출기가 33,327 어절을 처리하는 데 약 18 초가 소요되었다. C 프로그래밍 언어로 작성된 소스 프로그램의 크기는 대략 950 라인이며 수행 프로그램의 크기는 111,096 바이트로 시스템 실행시 최소 필요 메모리의 크기는 약 5.9M 이다.

시스템에서 사용하고 있는 사전의 등록어 수는 다음과 같다.

- 명사사전: 5,854
- 어미사전: 1,210
- 술어사전: 1,931

3.2 평가결과 및 분석

구현한 시스템에 대한 MATEC '99 의 성능 평가결과는 [표 2]와 같다. 결과를 보면 재현율보다 정확률의 성능이 떨어지며, 특히 뉴스 같은 정제된 텍스트에 비해 담화체 형식인 소설분야의 성능이 낮은 결과를 보이고 있다. 이는 소설분야가 술어의 비율이 높고, 앞서 기술한 바와 같이 이러한 단어를 명사로 추정하기 때문에 나타나는 현상이다. 명사추출 분야의 평가에 참여한 14 개 팀의 평균 F-Measure[3]가 0.8274 이므로, 본 논문에서 기술한 단순한 방법으로 평균을 약간 능가하는 성능을 보일 수 있음을 알 수 있다.

[표 2] 성능평가 결과

분야	재현율	정확률
뉴 스	0.9	0.81
비소설	0.92	0.81
소 설	0.9	0.65
평 균	0.91	0.77
F-Measure		0.8342

시스템 분석 결과 중에서 오분석 발생하는 원인은 크게 세 종류로 분류할 수 있다.

첫째, 모호성을 갖는 단어의 경우 의미적 분석을 하지 않기 때문에 발생하는 오류이다. 예를 들어, "바람에", "보신" 등은 문장 정보를 이용하여 품사를 추정해야 정확한 결과를 얻을 수 있는데 사전과 규칙만을 적용하기 때문에 의미와는 상관없이 항상 명사로 추정된다. 그러나 "보신", "이란", "가장"과 같은 단어가 독립적으로 사용될 경

우에는 명사일 확률이 매우 적기 때문에 이와 같은 부류의 단어는 조사와 결합한 경우에만 명사로 추정하는 규칙을 적용하면 해결 가능한 문제이다. "이 의원"과 같은 유형은 "이"와 "의원"이 띄어쓰기가 되어있고 "이"가 조사 사전에 존재하기 때문에 명사로 추정되지 않는 경우이다.

둘째, 분석에 적용되는 규칙이 모든 경우를 만족하지 않기 때문에 발생하는 오류이다. 이와 같은 유형의 어절로는 "중심지이자", "종이었는데", 등이 있다. "중심지이자"의 오분석은 "중심지+이자"의 형태로 발생하는데 이 때 "이자"를 명사를 분석하는 경우가 발생한다. 이는 적용규칙에서 정방향 매칭이 조사 매칭보다 우선하기 때문에 발생하는 현상이며, "종이었는데"도 비슷한 경우로 "종이"만을 항상 명사로 추정하는 오류이다.

셋째, 성능 평가 기준이 본 시스템의 사용 용도와 다르기 때문에 발생하는 오류이다. 이러한 오류의 종류의 크게 두 종류인데 그 중 하나는 숫자가 포함된 명사형이다. 예를 들어, "3 김청산의", "5 공", "49 주년" 등이 이에 속한다. 일반적으로 숫자가 포함된 명사형은 정보검색에서 중요한 명사로 취급될 수 있는데 본 평가대회에서는 숫자를 제외한 형태를 올바르게 분석한 것으로 평가했다. 즉, "3 김청산의"는 "3 김청산"으로, "5 공"은 "5 공"으로 분석해야 의미가 있는데, 평가기준은 "김청산", "공"으로 각각 분석해야 하는 것으로 되어 있다. 이와 같은 기준을 적용한다면 "49 주년"의 경우도 "주년"을 명사로 추출하는 것을 정답으로 해야 하는데 이 경우에는 제외가 되어 있어 논란의 여지가 있다.

다른 종류의 오류로는 준말을 원형으로 복원하지 않아서 발생하는 오류이다. 즉, "쌈질", "얘길" 등이 이에 속하는데 원형으로 복원하는 것은 형태소분석기 평가에서의 기준으로 적합한 것이지 명사 추출 평가에서는 적합하지 않다고 생각된다. "쌈질" 같은 단어는 준말 자체가 사전에 명사로서 등록되어 있기 때문에 두 가지 형태의 명사, 즉, 준말 형태로 추출한 경우와 원형으로 복원해서 추출한 경우를 정답으로 해야 옳을 것이다. 그러나, "얘길"과 같이 조사와 결합된 준말형태는 원형으로 복원해서 명사를 추출해야 한다.

평가결과 분석에서 보듯이 본 논문에서 기술한 방법으로 명사를 추출할 경우의 가장 큰 단점이 문맥정보와 의미정보를 사용하지 않기 때문에 발생하는 오류이지만, 시스템의 복잡도가 낮다는 점과 평가에 참여한 팀의 평가결과 수치를 볼 때 충분히 사용가치가 있다고 하겠다.

4. 재평가 실험

성능평가 기준상의 모호성이 존재한다는 것은 3 장에서 숫자와 결합된 형태의 명사형과 준말에 대해서 기술한 바 있다. 이 외에도 영어는 평가대상에서 제외하고 한문은 명사로 추출해야 하는 것은 평가기준상의 문제점으로 생각된다. 순수하게 한글만 평가대상으로 하는 것이 평가기준으로 적합하다고 보며, 한문 같은 경우에는 한글로 변환한 후 추출하는 것이 오히려 더 타당할 것이다.

평가기준상의 문제점으로 기술한 모든 사항들을 고려하여 재평가 실험을 하는 것이 타당하지만 평가문서 집합 전체에 대해서 수작업으로 평가 답안을 수정하기란 어려운 작업이다. 따라서, 본 연구에서는 숫자가 포함된 단어를 평가대상에서 제외한 실험을 하였다. 즉, "3 김청산의", "5 공"과 같은 단어를 제외시켰다. 평가에 사용한 시스템의 경우, 이러한 단어에 대해서 평가기준에 맞추어 정답을 추출할 수 있지만 공정한 재평가를 위해서 포함시키지 않았는데 재 평가결과 평균 3%의 정확률이 향상되었다.

5. 결론

본 논문에서는 학습데이터를 이용하여 생성한 규칙과 사전을 이용하여 명사를 추출하는 방법을 설명하고 이를 시스템으로 구현하여 명사 추출기 평가대회의 평가 결과를 기술하였다. 제안한 방법은 단일 명사뿐만 아니라 복합 명사 추정이 가능하며 접근 기법이 단순하기 때문에 시스템 복잡도가 낮으며 유지관리가 용이하다. 평가결과 제안한 명사 추출 방법이 복잡한 언어학적 측면의 의미분석을 하지 않지만 평균적인 성능을 보임을 알 수 있었다. 또한, 평가기준 중에서 적합하지 않은 부분을 규정하고 이를 기준으로 삼아 자체적으로 재평가한 결과를 제시하였다. 그러나, 성능향상을 위해서는 사전 등록단어의 사용 용도에 따라서 사전을 좀 더 세분화하여 구축할 필요가 있으며, 술어의 정확한 분석이 요구된다.

현재는 평가대회의 정확한 평가기준이 마련되지 못한 점이 아쉽고 특히, 명사 추출기의 경우에는 시스템의 사용용도를 정한 후 이를 기준으로 평가지침을 마련하는 방안을 고려했으면 한다. 예를 들면 단순히 명사추출의 정확률, 재현율만 측정할 것이 아니라 추출결과를 검색에 직접 사용하여 성능분석을 하는 것이 바람직하다. 그리고, 명사분석기 뿐만 아니라 형태소분석기, 태거 등의 분야에 대한 국내 기술 향상을 위해서는 MATEC '99 와 같은 평가대회가 지속적으로 이루어지길 기대한다.

참고문헌

[1] Sung Hyon Myaeng & Dong-Hyun Jang, "On Language Dependency in Indexing", Proceedings of the Workshop on Information Retrieval with Oriental Languages, 1996.

[2] 장동현, 맹성현, "효율적인 색인어 추출을 위한 복합명사 분석방법", 제 8 회 한글 및 한국어 정보처리 학술대회, 1996.

[3] Gerard Salton & Michael J. McGill, Introduction to Modern Inforamtion Retrieval, McGraw-Hill, Inc., 1983.