시드를 이용한 도메인 관련 복합어 추출 기법

조성원[O] 최종필* 김민구*
아주대학교 정보통신 전문대학원[O], 아주대학교 정보 및 컴퓨터 공학부**
ajoujoa@ceai.ajou.ac.kr[O], { cjp, minkoo }@ajou.ac.kr**

Extracting Domain Related Multi-word Terms using Seeds

Sungwon Cho[O] , Jongpil Choi*,
Graduate School of Information and Communication, Ajou University,
Minkoo Kim*
College of Information & Computer Engineering, Ajou University

요 약

복합어 추출 기법은 최근 활발한 연구가 진행되고 있는 온톨로지 구축과 정보 검색에 중요한 기법으로 연구되어 왔다. 초기의 연구는 주로 언어학적인 필터 기법이나 통계적 기법을 사용하였지만, 최근 문맥 정보와 의미 사전 등을 이용하여 용어를 추출하는 방법으로 발전해 오고 있다. 또한 정보검색 분야와 온톨로지 분야에서도 모든 용어를 추출하는 방법보다 문서 집합의 도메인에 적합하다고 판단되는 용어들을 추출하는 방법이 그 성능을 향상시킬 수 있다. 본 논문에서는 통계학적 방법을 이용하여 도메인에 적합한 시드 용어의 추출을 하고, 그 시드 용어를 이용해 가중치를 정제하는 방법과 시드 용어로부터 관련된 용어를 추출해 나가는 방법을 적용하여 문서 집합의 도메인에 맞는 용어들을 추출하고자 한다.

1. 서 론

문서 정보의 양이 많아지면 많아질수록 정리되지 않은 정보의 양도 많아지기 마련이다. 따라서 사용자에게 맞는 정보를 찾기 위해 정보검색 기술이 발전하였고, 이러한 지식을 형식화 하고 정리하기 위한 노력으로 온톨로지를 구축하기 시작하였다[1]. 하지만 문서 집합에서 단어를 추출하는 것만으로는 검색은 할 수 있으나, 온톨로지를 구축하기 위한 완전한 지식을 얻어낼 수가 없다. 온톨로지는 그 도메인에 관련된 지식을 모아 놓은 것이기 때문에 정확한 뜻을 표현해야 한다. 단어 만으로 구축할 경우 복합어가 지니고 있는 정확한 의미를 표현할 수 없을 경우가 많다. 따라서 복합어를 추출하는 것이 온톨로지를 구축하는데 필요한 과정이다.

복합어를 추출하기 위해서는 일단 후보가 될 수 있는 단어 묶음을 언어학적인 필터를 이용해 골라 내고, 통계학적 방법으로 중요도를 산출한다. 본 논문은 모든 복합어의 추출보다는 문서 집합의 도메인에 관련된 용어들을 추출하기 위해, 도메인에 관련된 시드 용어를 수동으로 결정하는 방법을 택하였다. 그 이후에 도메인에 관련된 시드 용어 정보를 이용하여 다른 용어들의 가중치를 다시 매기는 방법과 시드 정보로부터 용어를 점차 늘려나가는 방법을 제안한다.

본 논문의 구성은 2장에서 관련 연구를 살펴보고, 3장에서 시드 용어를 뽑아내는 방법과 가중치를 다시 매기는 방법, 용어를 시드 용어로부터 늘려나가는 방법, 두 가지를 제시한다. 4장에서 실험 결과를 관련 연구와 비교 분석하고 5장에서 결론 및 향후 과제에 대해 서술한다.

2. 관련 연구

지금까지의 복합어 추출 분야에서는 언어학적인 필터를 이용하고 통계적인 정보를 더한 것을 기본으로 한 여러 가지 기법들이 소개되었다. 아래는 이러한 기법들 중 문맥 정보를 이용하는 방법과 결합률을 측정하는 방법에 대한 설명이다.

2.1 문맥 정보에 기반한 연구

문맥 정보를 이용한 방법 중 대표적인 것은 NC-Value를 이용하는 방법이다. 문맥 정보에 기반한 방법이지만 통계적 정보인 C-Value와 함께 적용된다. C-Value는 출현 빈도수에 크게 좌우된다. 또한 다른 복합어의 부분집합인 경우 그 중요도를 낮추어 준다. 이는 상대적으로 다른 용어를 포함하는 복합어의 가중치를 높여주는 결과를 낳는다[3]. C-Value를 구한 이후에 모든 문맥에 대해 문맥으로서의 가치를 구하여, 그 합을 문맥 정보로 이용한다. 문맥으로서의 가치는 그 문맥이 전체 문서 집합에서 얼마나 많은 용어들과 함께 나왔는지의 정도를 측정하여 적용한다[2]. 따라서 그 의미는 매우 일반적이다.

반면 그 단어가 목표 복합어 와 얼마나 사전적으로 가까운지를 UMLS와 같은 시맨틱 네트워크를 이용해 적용한 SNC-Value가 있다[4]. 여기서의 문맥으로서의 가치는 목표 복합어와의 사전적 연관성이라고 할 수 있다.

이러한 방법들은 도메인과의 연관성보다는, 일반적으로 문서 집합 내에 존재할 만한 용어인지를 가늠할 수 있는 정도이다.

* 본 논문은 과학기술부의 국가지정연구실 사업(과제명:차세대 인터넷을 위한 지능형 온톨로지 자동생성 시스템 개발, 과제번호:M10302000087-03J0000-04400) 지원으로 수행되었음

2.2 연관률에 기반한 연구

연관률은 복합어를 단어들의 결합으로 보고, 두 개 또는 그 이상의 단어가 결합을 할만 한 확률을 구하는 것이다[5]. 이 방법은 복합어 후보로서 단어 쌍 (A,B) 들을 구성한다. 각각의 단어 쌍에 대해, 네 개의 척도를 구하여 그 값의 조합으로 순위를 매긴다. 네 개의 척도는 A가 출현한 빈도수, B가 출현한 빈도수, A B 모두 출현한 빈도수, 두 단어 모두 출현하지 않은 빈도수이다. 이를 가지고 적절한 조합을 이루어 순위를 결정한다. 그 조합이 의미하는 바는 A, B 두 단어가 결합하여 출현할 확률이라고 할 수 있다. 단어 쌍에 대한 측정이 끝나고 나면 더 많은 단어로 이루어진 복합어에도 적용을 한다. 이 방법 역시 용어를 추출하는 데는 쓰일 수 있으나 문서 집합에서의 중요도를 구한 것은 아니므로 도메인과의 적합성과는 거리가 멀다.

3. 도메인에 관련된 복합어 추출 기법

복합어를 포함한 용어를 추출하는 방법으로 통계학적 방법이 사용될 수는 있다[2]. 그러나 문서 집합에 자주 나온다는 통계학적 근거로 도메인에 적합하다는 판단을 내릴 수 없다. 따라서 본 논문에서 택한 방법은 도메인에 관련된 시드 용어를 이용하여 도메인에 관련된 용어를 추출하는 방법을 제안한다.

3.1 NC-Value를 이용한 시드 용어 추출

시드를 뽑아내기 위해 NC-Value 방법을 이용해 나온 결과를 이용한다. 상위 용어 중에 사람의 판단 하에 문서 집합의 도메인과 관련이 높다고 생각되는 것들을 뽑아서 리스트로 저장한다. 여기서 쓰이는 NC-Value는 조금 변형된 것으로 복합어만을 뽑아내는 것이 아니라 단어도 같이 뽑도록 하였고, 문장의 평균 길이를 생각하여 문맥 정보의 범위는 앞뒤로 5개의 정보를 이용하였다. NC-Value는 다음과 같은 식으로 구한다.

(식1)

$$NC-Value(a) = 0.8C-Value(a) + 0.2\sum_{b \in C_a} f_a(b)\,weight(b)$$

a : 용어
C_a : a의 전체 문맥
b : 문맥에 포함된 하나의 단어
$f_a(b)$: b가 a와 함께 출현한 빈도
$weight(b)$: b의 문맥 정보로서의 가치

3.2 시드 용어를 이용한 문맥 정보의 정제 방법

이 방법은 기존의 NC-Value 방법을 변형한 방법으로 문맥 정보에 해당하는 단어의 가중치를 구할 때 주어진 시드 용어를 활용한다.

문맥 정보에는 문맥으로서의 중요성을 나타내는 용어의 가중치 값이 있다. 가중치 값은 전체 용어 중에 몇 개 의 용어와 함께 출현했는지를 나타내는 비율로서 문맥 정보로서의 가치를 나타내는 척도이다. 여기서 만약 시드 용어와 같이 나온 적이 있는 문맥일 경우 그 가중치를 더해 줄 필요가 있다. 왜냐하면 시드 용어와 같이 나온 적이 있다면 그 단어 자체가 중요한 의미를 지니지 않더라도, 시드 용어와 관련되어 있다는 판단을 할 수 있기 때문이다. 따라서 이러한 문맥과 함께 나온 용어는 시드 용어와 함께 도메인에 관련된 것이라고 판단할 수 있다. 시드는 사람이 직접 도메인에 관련된 것으로 뽑은 것이기 때문에, 이렇게 가중치를 부여하는 것은 도메인에 맞도록 복합어의 중요도를 결정하는 것이라고 할 수 있다. 가중치에 관한 식을 다음과 같이 수정한다.

(식2)

$$weight(b)=\frac{t(w)}{n} \rightarrow weight(b)=\frac{(t(w)-s)+(SP)ns}{n}$$

$t(w)$: b를 문맥 정보로 가지는 용어의 수
n : 고려되어지는 전체 용어의 수
s : b를 문맥 정보로 가지는 시드 용어의 수
SP : 시드 용어와 같이 나왔을 경우, n에 대한 특정 비율의 가중치

문맥으로 쓰이는 각 단어마다 위와 같은 가중치를 가지게 되는데 시드와 같이 나오는 횟수에 대해서는 그냥 1회의 출현으로 간주되는 것이 아니라, 고려되는 전체 복합어의 특정 비율에 해당하는 출현으로 가정한다. 이 SP 값은 실험적으로 측정해 정한 값으로 0.01, 즉 1% 값으로 한다.

3.3 시드 용어에서 부트스트랩을 이용한 용어 추출

본 논문에서는 3.1에서 구한 도메인에 관련된 시드 용어를 부트스트랩 방법에 적용하여 시드 용어와 관련 있는 복합어를 추출하고자 한다. 그 방법으로 먼저 추출된 시드 용어마다 전체 문맥 정보의 리스트를 구성하고 시드와의 관련도를 계산하기 위해 시드와 각 리스트와의 매트릭스를 (그림 1) 과 같이 만든다.

매트릭스에서 각 용어마다 모든 시드 용어와의 출현 빈도의 합을 구한다. 이것을 정렬하여 상위에 위치한 용어들을 추출한다. 이 과정이 여러 번 반복이 되는데, 이 때 전 단계에서 추출된 용어들은 다음 번 시도에서 시드 용어로서 쓰이게 된다. 이러한 몇 번의 과정을 거쳐 목표한 개수만큼의 용어가 추출되도록 한다. 이 방법은 도메인에 관련 된 용어만을 뽑아내기 위한 기법으로 초기 시드 용어와 크게 연관되지 않는 한 용어로 뽑히지 않기 때문에, 더욱 시드의 중요성이 강조된다. 다음은 각 용어마다 시드 용어들과의 관련도를 계산하는 식이다.

(식3)

$$Sum(a) = \sum_{s \in S_a} freq(a,s)$$

a : 용어
S_a : a를 문맥 정보로 가지는 시드의 집합
s : a를 문맥 정보로 가지는 하나의 시드
$freq(a,s)$: 용어 a와 시드 b가 함께 나온 횟수

	term1	term2	term3	term4	term5	...
seed1	0	0	0	10	3	...
seed2	0	1	1	2	0	...
seed3	0	4	2	5	0	...
seed4	0	0	4	2	1	...
.....	...	...	...	...	...	...

(그림 1)

4. 실험 및 결과

실험를 위한 텍스트 데이터로는 NIST에서 주관하는 TREC 컨퍼런스의 데이터 중 Text Research Collection Volume 2의 문서 집합을 이용하였다. 이 문서 집합은 AP 통신의 기사를 모아 놓은 것으로서 24만 개의 문서의 집합이다. 모든 문서에 대해 Brill Tagger를 이용해 품사 정보를 이용하였다.

문서 집합에 대해서 결합률을 이용해 추출한 리스트 [L1], NC-Value 방법에 의해 추출된 리스트 [L2], 그리고 문맥 정보의 정제 방법에 의해 추출된 리스트 [L3]와 부트스트랩 방법을 이용해 추출한 리스트 [L4]를 각각 비교해 보겠다. 각각 상위 10개의 리스트는 아래 [표1]에서 보는 바와 같다.

L1	soviet erstwhil soviet paravan soviet alexandor soviet gennadij soviet galya soviet suthor soviet vartanan soviet yuvenali soviet pozhalovat soviet gr ast	L2	united states soviet union new york president bush britain london world war ii white house persian gulf new york stock
L3	united states soviet union new york president bush world war ii white house new york stock last year first time dow jones average	L4	britain president bush london persian gulf boston new jersey chemical weapons bush administration capital gains american troops

[표1]

[L1]에서는 단순히 나올 법한 용어들만 나열된 모습을 볼 수 있다. 동정자가 많기 때문에 전부 구소련에 관련된 용어들만 보인다. 따라서 이 리스트에서 도메인과 관련도는 없어 보인다. 반면 [L2]에서는 상당히 기사에서 많이 나올 법한 용어들이 눈에 띈다. 사람이 판단하였을 때 도메인과 어느 정도 관련이 있다고 판단할 수 있을 정도이 다. [L2]에서 [L3]와 [L4]를 추출하기 위해 시드 용어를 정하였다. 상위에서 시드 용어를 선별해 보니 미국과 관련된 용어가 많았다. 시드 용어는 모두 35개의 용어로 구성되었다. 이를 이용해서 리스트 [L3]와 [L4]를 추출해 내었다.

[L3]에서 눈에 띄는 것은 " britain" 과" london" 이 상위에서 밀렸다는 사실이다. 시드가 미국과 관련되어 있었기 때문이라고 볼 수 있다. [L4]에서는 거의 모든 용어들은 미국에 관련되어 있음을 확인할 수 있다. 또한 [L4]에서 일반적으로 나올 수 있는, 특별한 의미가 없는 용어가 훨씬 줄어든다는 사실도 확인할 수 있었다.

다음은 각 방법의 상위 100개의 리스트에 대한 정밀도(precision)를 측정한 것이다.

리스트	L1	L2	L3	L4
정밀도(precision)	0.07	0.55	0.61	0.68

[표2]

위의 표를 봤을 때 가장 도메인에 관련된 것을 찾은 경우는 [L4]임을 알 수 있다. 반면 [L1]은 거의 도메인에 관련되지 못한 것으로 나타난다.

5. 결론 및 향후 과제

본 연구에서는 시드 용어를 사용하여 문맥 정보를 정제하는 방법과 부트스트랩을 이용해 용어를 추출하는 방법을 제안하였다. 문맥 정보를 정제해 주는 방법은 도메인에 관련되지 않은 복합어를 제거해 줄 수 있는 장점이 있고, 부트스트랩 방법은 일반적인 단어를 극히 줄이고 거의 시드에 관련된 복합어만 뽑아낼 수 있다는 장점이 있다는 것을 알았다.

앞으로의 과제는 온톨로지를 구성하는 데 복합어의 추출이 더욱 잘 활용될 수 있는 방향으로 나아가는 것이다. 온톨로지는 AP문서 집합과 같이 통합적인 도메인 보다는, 도메인의 집중도가 높은 문서 집합에서 구축될 것이므로 특정 도메인에 대한 실험이 더 필요하다.

6. 참고 문헌

[1] A. Maedche, S. Staab: Semi-Automatic Engineering of Ontologies from Text, 2000

[2] Katerina T. Frantzi, Sophia Ananiadou, Junichi Tsujii: The C-Value/NC-Value Method of Automatic Recognition for Multi-word Terms, ECDL, 1998

[3] Frantzi, K.T: Incorporating Context Information for the Exraction of Terms, ACL, 1997

[4] Diana Maynard, Sophia Anaiadou: Identifying Contextual Information for Multi-word Term Extraction, 1999

[5] Kenneth Ward Church, Patrick Hanks: Word Association Norms, Mutual Information, and Lexicography (1989)

[6] Hiroshi Masuichi, Raymond Flournoy, S. Kaufmann, S. Peters: A bootstrapping Method for Extracting Bilingual Text Pairs, 2000

[7] Diana Maynard, Sophia Ananiadou: Terminological Acquaintance : the Importance of Contextual Information in Terminology, 2000