위키피디아에 기반한 단어 사이의 의미적 연결 관계 탐색

김주황[O], 홍민성[*], 이오준[*], 정재은[*]
[O]중앙대학교 컴퓨터공학부
[*]중앙대학교 컴퓨터공학과
e-mail:kjhs126@cau.ac.kr[O], {minsung87, concerto34, j3ung}@cau.ac.kr[*]

Discovering Semantic Relationships between Words by using Wikipedia

Ju-Hwang Kim[O], Min-sung Hong[*], O-Joun Lee[*], Jason J.Jung[*]
[O]Dept. of Computer Science and Engineering, Chung-Ang University
[*]School of Computer Science and Engineering, Chung-Ang University

요 약

논문에서는 위키피디아를 이용하여 단어 사이의 유사도와 내포된 연결 단어들에 대한 탐색 기법을 제안한다. 위키피디아에서 제공하는 API를 이용하여 두 단어 사이를 탐색함으로써, 기존 단어 사이의 유사도를 계산하는 방식보다 더 간단하고 폭 넓은 의미 집단을 포괄할 수 있다. 이는 그래프적 특성에 기반하며 그래프를 구성하는 방식으로써 동적 방식과 정적 방식으로 구성된다.

▸ Keyword : 의미 유사도(semantic similarity), 위키피디아(wikipedia), 그래프(graph), 워드 넷(word net)

I. 서론

현재까지 온톨로지나 워드넷(WordNet) 등 어휘 간 의미 유사도 기반의 다양한 연구들이 진행되었다. 이러한 연구들은 어휘목록 사이의 다양한 의미 관계를 기록하며 자동화된 본문 분석과 인공 지능 응용을 뒷받침해준다.

하지만 기존 연구들은 어휘들의 유의어 집단(synset)을 형성하는데 그치고 있다. 따라서 좀 더 폭 넓고 심도 있는 단어 간의 의미를 분석하기 위해, 단어와 단어 사이를 연결하는 단어들에 대한 탐색이 필요하다. 예를 들어, 존 F. 케네디가 암살된 연도는 1963년이며, 이는 우리나라에서 최초의 라면인 삼양라면이 판매된 연도이다. 이처럼 의미의 흐름을 분석하여 다른 방식으로 의미 집단을 형성하는 것이 가능하다.

본 논문에서는 위키피디아에 존재하는 두 단어 사이의 연결 관계를 찾는 것에 초점을 둔다. 이를 위해 그래프적 특징과 위키피디아의 API를 이용한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 기술한다. 3장에서는 본 연구의 상세한 내용을 서술하고, 4장에서는 결론과 향후 연구 방향에 대해 설명한다.

II. 관련 연구

기존 단어나 어휘사이의 유사도를 측정하는 연구들은 활발히 진행되어 왔다. Jay J. Jiang은 어휘 분류 체계와 통계적 기법을 활용하여 어휘 사이의 유사도 측정을 하였다 [1]. Danushka Bollegala은 검색 엔진을 이용하여 두 단어를 같이 검색함으로서 나오는 웹 페이지의 결과에 따라 단어 간 유사도를 측정하였다 [2]. 하지만 기존의 이러한 연구들은 단어나 어휘 단어 사이의 유사도를 측정하는 것에 그치고 있다.

따라서 단어 간 의미를 분석하기 위해, 단어와 단어 사이를 연결하는 단어들에 대한 탐색이 필요하다.

III. 단어 간 연결 탐색 기법

1. 그래프 구성

위키피디아 문서에는 개념적인 내용과 더불어 재검색 될 수 있는 단어들은 하이퍼링크로 연결되어 있다. 이러한 단어들은 검색된 단어와 유사성이 높은 단어라고 가정할 수 있다. 위키피디아 API에서 제공하는 링크 메타 데이터들을 이용하여 유사도 계산에 치중하거나 도메인에 얽매이지 않고 간단히 단어 네트워크를 확장시킬 수 있다. 따라서 검색된 단어와 연결된 단어들은 재검색이 가능하므로 재귀적으로 단어 네트워크 그래프를 구성해 나갈 수 있다.

그래프 구성은 동적인 방식과 정적인 방식으로 나뉜다. 동적인 방식은 HTTP를 이용한 REST 형식의 쿼리를 이용하여 검색된 단어와 링크된 단어들을 XML 형식으로 얻는 과정을 재귀적으로 호출하여 그래프를 형성한다. 정적인 방식은 위키피디아의 덤프파일 데이터베이스화하여 단어와 단어 사이의 연결 관계를 파악해 전체적인 그래프를 구성 한다.

2. 두 단어 사이의 연결 관계 탐색

2.1 동적 그래프

동적인 방식에서는 그래프의 전체적인 연결 관계를 파악하기 어렵기 때문에 BFS(Breadth-first search) 탐색 알고리즘을 이용한다. 그래프 탐색 과정에서 단계수를 지정하여 그래프의 크기를 제어한다. 일정 단계수를 넘어가게 되면 그래프 상으로는 연결되어 있지만 의미상으로는 서로 관련이 없다고 판단한다. 이에 대한 알고리즘은 표 1과 같다.

Table 1. 동적 그래프 생성 알고리즘

```
G : words of graph
L : the set of linked word
Q : queue
a, b : input words
c : depth of graph
w : selected word

Q <- a
While c < proper upper bound of c
  w <- Q
  If w == b
    path <- find path of graph from a to b
    return path
  End if
  L <- linked words of w using wiki API
  For i = 1 to i = number of L
          ∈ L
    If w_i is not visited
      G <- w_i
      Q <- w_i
    End If
  End For
End While
```

2.2 정적 그래프

전체 그래프를 데이터베이스화한 상태에서의 그래프 생성은 최단 경로 문제와 동일하다. 모든 간선의 가중치는 동일한 값으로 계산하며 잘 알려진 데이크스트라 알고리즘(Dijkstra algorithm)을 이용한다.

2.3 유사도 계산

그래프 탐색 결과로 연결된 간선의 수에 따라 두 단어 사이의 유사도를 측정할 수 있다. 유사도를 연결된 간선들의 합으로 표현함으로써, 두 단어 사이의 경로를 통해 빠르게 계산하는 것이 가능하다.

IV. 결론

위키피디아에서 제공하는 API를 이용하여 단어 간 유사도와 내포된 다른 단어들을 탐색하였다. 제안한 방법은 문서 내 하이퍼링크를 통해 그래프를 구성하여 간단하고 넓은 도메인을 포함할 수 있다는 장점이 있다.

향후에는 좀 더 정교한 유사도 계산 방법과 단어 사이의 여러 링크들을 포괄하여 의미 있는 연결 단어들을 추출에 대한 연구가 필요하다.

사사

이 논문은 2015년 미래창조부 및 정보통신기술진흥센터의 재원으로 서울어코드활성화사업의 지원을 받아 수행된 연구임 (R0613-15-1205). 또한, 이 논문은 2014년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2014R1A2A2A05007154).

참조논문

[1] Jay J. Jiang, David W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy" In Proceedings of International Conference Research on Computational Linguistics (ROCLING X), arXiv preprint cmp-lg/9709008, 1997.

[2] Bollegala, Danushka, Yutaka Matsuo, and Mitsuru Ishizuka. "Measuring semantic similarity between words using web search engines." Proceedings of the 16th international conference on World Wide Web, pp.757-766, 2007.