문서 요약 시스템을 위한 수사 구조 트리 생성

정준호, 김미진, 이현주, 박미성, 이상조
경북대학교 컴퓨터공학과

Rhetorical Structure Tree Generation for Text Summarization System

Joon-Ho Jung, Mi-Jin Kim, Hyun-Ju Lee, Mi-Sung Park, Sang-Jo Lee
Department of Computer Engineering, Kyungpook National University

요약

본 논문에서는 수사 정보와 문장간 유사도를 이용하여 문서의 수사 구조 트리를 생성하는 방법을 제안하였다. 말뭉치에서 찾아낸 수사 정보를 토대로 분류하고, 이를 이용해서 문서 내의 수사 정보를 추출해서 가능한 모든 구조를 생성한다. 다음으로, 장간의 유사도를 사용해서 가장 높은 하나의 구조를 선택한다. 생성된 수사 구조를 사용하여 문서를 요약할 수 있는데, 수사 정보는 언어적 특성을 이용하는 것이므로 도메인에 독립적인 요약 시스템을 만들 수 있다.

1. 서론

장세시스템으로 인터넷에 있는 문서들을 검색할 때, 너무나 많은 문서가 검색되기 때문에 검색된 문서가 적합한 것인지 확인하기가 쉽지 않다. 만약 이러한 문제들에 대한 요약문을 가지고 있다면, 사용자가 원하는 문서를 찾는데 도움을 줄 수 있다.


본 논문에서는 수사 정보를 사용하여 수사 구조를 생성하는 방법을 제안한다. 수사 정보를 사용하여 요약문을 추출하는 방법은 문서에서 수사 구조를 추출하는 과정과 이를 이용해서 요약하는 과정으로 나누는데, 본 논문은 그 중 일부분만을 대상으로 한다. 문장간에 존재하는 병렬 관계, 부연 관계 등 여러 가지 관계가 있는데, 문서에서 추출한 수사 정보와 문장간의 유사도를 이용하여 이러한 관계를 추출한다. 통계적인 방법은 사용하는 도메인에 종속된 결과를 가져오지만, 수사 정보는 언어적 특성을 이용하는 것이므로 이러한 방법을 사용화면 도메인에 독립적인 시스템을 만들 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 수사 구조란 어떤 것인지 살펴보고, 3장에서는 수사 구조를 생성하는 방법을 제안하고, 마지막으로 4장에서는 결론을 내린다.

2. 수사 구조

수사 구조는 문서 내에서 각 문장 사이의 관계를 나타내는 이진 트리이다. 트리의 leaf node는 문장을 가리키고, 중간 node들은 문장 사이의 관계를 나타낸다. 이러한 관계는 문서의 장간 단위에서 추출한 수사적인 정보와 문장간의 유사도를 가지고 생성한다.

수사적인 정보는 말뭉치에서 추출하고, 이러한 정보를 바탕으로 비슷한 역할을 하는 것끼리 분류하여 문장사이에 어떠한 관계가 존재하는지 찾아낸다. 이 논문에서는 이미 예의 될 바 문서 관계를 분류하는 방법을 제시한다. 다음의 표는 이러한 관계의 예를 보여준다.

<table>
<thead>
<tr>
<th>관계</th>
<th>표현</th>
</tr>
</thead>
<tbody>
<tr>
<td>병렬</td>
<td>그리고, 또한</td>
</tr>
<tr>
<td>반대</td>
<td>그렇다, 하지만</td>
</tr>
<tr>
<td>결론</td>
<td>따라서</td>
</tr>
</tbody>
</table>

[표 1] 수사 관계의 예
다음의 예는 실제로 문서에서 수사구조가 어떻게 구성되어 있는지를 보여준다.

[예제 1]
정보(information)에 대한 개념적 정의는 일관적이지 않다. (s1) 정보는 그 용어가 쓰여지는 상황(context)에 따라 각기 다른 의미로 해석될 수 있기 때문이다. (s2) 정보의 원어는 중세 라틴어인 informatio로서 출발한다. (s3) 당시의 의미는 주어진 어떤 '정상적' 구조와 '교사' 등을 뜻했던 것으로 알려지고 있다. (s4) 또한 프랑스에서는 고전적인 원어로서 구조 법적인 차원에서 '어떤 질서에 대한 수급 및 처리'의 의미로 사용되었다. (s5)

[그림 1] [예제 1]의 수사 구조

위의 그림에서 각 leaf node는 위 예제(1)의 각 문장을 가리키고 중간노드는 관계를 나타내는데, "~ 때문에", "당시의", "또한" 등의 수사 정보와 5개 문장의 유사도를 사용해서 [그림 1]과 같은 수사 구조를 생성한다.

3. 수사 구조의 생성

전체 시스템 구성도는 [그림 2]와 같다. 본 시스템은 문서에서 수사 정보의 추출과 문장간 유사도 계산, 문장간 구조 생성과 문단간 구조 생성, 그리고 최종 구조 생성 과정으로 이루어져 있다.

[그림 2] 전체 시스템 구성도

앞서 기술한 바와 같이, 문서 전체의 수사 구조를 만들기 위해서는 문장간, 문단간 수사 구조 생성의 두가지 과정을 거친다. 여기서 문장간 수사 구조와는 문단 안에 포함된 문장간의 구조이다. 각각의 구조를 생성한 후 문단간 수사 구조의 leaf node에 문장간 수사 구조를 대치시켜 전체 구조를 생성한다. 문장이나 문단간 구조를 만드는 과정은 매우 유사하므로, 이후 문장간 구조를 만드는 과정을 설명한다.

문장간 수사 구조 생성에는 수사정보와 문장간 유사도 사용한다. 그 과정은 다음과 같다.


[그림 3] 수사 구조 트리의 예

다음에는, 위에서 생성한 모든 수사 구조에 가중치를 분야하여, 그중 가장 가중치가 높은 것을 선택한다. 가중치는 문장간의 유사도를 이용한다. 문장간의 유사도가 높은 것을 지역하게 가할 수록, 가중치가 가장 높게 된다. 만약 가중치가 같거나 비슷하다면, 우축 노드의 길이가 좀 더 긴 것을 선택한다. 문장간의 유사도는 코사인 계수를 이용하여 다음과 같은 식으로 계산된다.

\[
\text{sim}(S_i, S_j) = \frac{\sum_{m \in S_i} W(t) W_j(t) (W(t) + W_j(t))}{\sqrt{\sum_{m \in S_i} W(t)^2} \sqrt{\sum_{m \in S_j} W_j(t)^2}}
\]

여기서 Si, Sj는 비교할 대상이 되는 문장이나 문단이고, W(t), W_j(t)는 각각 Si, Sj에서의 영역 t의 빈도이다. W(t)는 Si, Sj가 문장이나 문단이나 문 문단 내에서의 영역 t의 빈도 또는 문서 전체에서의 빈도가 된다. 여기서 W(t)값으로 나누어 주는 이유는 문서 내에서 일반적으로 많이 나온 영역은 문장간의 유사도 계산에서 가중치를 낮추어 주기 위함이다.
미지막으로, 생성된 수사 구조에서 수사 정보가 없는 문장에 대한 수사 관계를 설정해 본다. 첫째 단계에서 수사 정보가 없는 문장에 대해서는 관계를 설정하지 않는다. 수사 정보가 없는 문장은 앞 문장들과 같은 주제를 이어가는 경우(전환)와 다른 주제를 이어가는 경우(전환)의 두 가지 관계만을 생성한 다. 아직 결정되지 않은 노드의 좌측 문장들과 우측 문장들을 사 이의 유효도를 구하키 위해 결정을 남기면, 얻지 않으면 전 환의 관계로만 만들어 준다.

다음의 예는 실제로 이와 같이 수사 구조를 생성한 예인데, [그림 3]의 첫 번째 그림은 수작업으로 구조화를 본 것이고, 두 번째 그림은 위에서 제안한 방법을 이용하여 생성한 구 조이다.

[이제 2] (조선일보 사설 중에서)

4. 결론
본 논문에서는 문서의 수사 정보와 문장간 유사도를 이용한 수사 구조 트리를 생성하는 방법을 제시하였다. 항상 문서에서 어떤 곳, 어떤 척도, 어떤 메시지를 추출하고, 이를 이용해서 문서의 가능한 모든 수사 구조를 생성한 후, 유사도를 이용해서 하나의 구조를 선택하였다.

본 시스템에 생성한 수사 구조는 문서를 요약하는데 쓰일 수 있다. 또한 제안한 방법을 요약 시스템에 적용할 경우에는 각 문장에 따라 문장의 상대적인 중요도가 다른 점을 이용해 서 각 문장의 중요도를 계산한 후, 중요한 순서대로 문장을 추출해야 한다.

항후 연구 과제는 수사 정보의 중요도를 좀 더 세밀히 분류하고, 문장간 유사도의 의미와 흐름정보 등의 정보를 사용하여 수사 구조의 개선의 정확도를 높이는 연구가 필요하다.

참고문헌