**한국어 TextRank를 이용한 키워드와 핵심문장 추출 평가**

**정진원, 서태원**

1. **서론**

앞서 진행했던 프로젝트에서 TextRank를 이용하여 다양한 방식으로 한국어 텍스트에 대한 추출적 요약 및 키워드 추출을 위한 시스템을 구축하였다.. 핵심 문장 추출을 위해 사용했던 방식은 총 6가지로, 전처리를 거치지 않은 경우 하나의 어절을 어근으로 사용한 문장간의 유사도를 (1) 코사인 유사도를 이용하여 계산, (2) 논문에서 제안된 유사도 공식을 이용하여 계산, (3) 코사인 유사도와 TF-IDF 방식을 결합하여 계산하였으며, 전처리를 거쳐 특정 품사만을 이용하여 나타낸 문장간의 유사도를 (4) 코사인 유사도를 이용하여 계산, (5) 논문에서 제안된 유사도 공식을 이용하여 계산, (6) 코사인 유사도와 TF-IDF 방식을 결합하여 계산하였었다. 키워드 추출의 경우 (1)전처리가 된 텍스트의 단어들과, (2)전처리가 되지 않은 어절들을 co-occurrence 관계를 이용하여 단어간 관계를 나타내는 그래프를 생성, TextRank 알고리즘을 적용하여 키워드를 추출했었다. 핵심 문장 및 키워드 추출의 성능을 평가하기 위하여 핵심 문장과 키워드를 사람이 직접 고른 데이터셋을 구축하여, 문장 추출 시스템이 추출한것과 얼마나 일치하는지 비교하였다.

1. **핵심 문장 추출**

네이버 뉴스의 IT, 정치, 경제 부분의 기사와, 역사 교과서에서 발췌한 텍스트, 위키피디아 문서에서 수작업으로 5개의 주요 문장과 10개의 키워드를 추출하여 총 77개의 문서를 구축했다. 역사 교과서를 선택한 이유는 추후 FOCUSLY에서 자주 이용될 텍스트이기 때문이다. 구축한 문서의 정보는 아래와 같으며, 해당 수치는 소수점 첫째 자리에서 반올림하였다.

|  |  |
| --- | --- |
| 평균 문장의 수 | 1,096/77 = 14 |
| 평균 어절의 수 | 19638/77 = 255 |

문장 추출 시스템의 평가는 앞서 서론에서 설명한 6가지 방법으로 추출한 문장이 사람이 선택한 문장에서 포함된 비율로 계산하였다. 결과는 아래의 표와 같으며, 해당 수치는 소수점 셋째 자리에서 반올림하였다.

|  |  |  |
| --- | --- | --- |
|  | 유사도 공식 | 평균 |
| 어절 | (1)코사인 유사도 | 0.49 |
| (2)논문에서 제안된 유사도 공식 | 0.48 |
| (3)코사인 + TF-IDF | 0.5 |
| 전처리 | (4)코사인 유사도 | 0.54 |
| (5)논문에서 제안된 유사도 공식 | 0.52 |
| (6)코사인 + TF-IDF | 0.55 |

전처리를 사용한 경우의 결과가 어절을 어근으로 두었을 때보다 성능이 월등히 앞섰다. 이는 어절을 어근으로 두었을 때 성능이 좋았던 [1]의 결과와는 상반된다. 하지만 평가 방식이 다르므로 큰 의미는 가지지 못할 것이다. 또한, 두가지 경우 모두 코사인 유사도와 TF-IDF 방식을 결합하여 사용하였을때 가장 우수한 성능을 보였다.

1. **키워드 추출**

핵심 문장 추출에 이용된 텍스트에서 직접 추출한 10개의 키워드가 전처리된 텍스트에서 추출된 10개의 키워드와 전처리되지 않은 텍스트에서 추출된 10개의 키워드에 포함된 비율로 계산하였다. 결과는 아래의 표와 같으며, 소수점 셋째 자리에서 반올림하였다. Co-occurrence window의 크기는 2로 설정했다.

|  |  |
| --- | --- |
|  | 평균 |
| 전처리 X | 0.28 |
| 전처리 O | 0.39 |

키워드 추출의 경우도 마찬가지로 전처리를 한 경우의 성능이 더 좋았다. 전처리를 하지 않은 경우, 사람이 잘 선택하지 않는 ‘수', ‘때'와 같은 한글자 키워드가 선택 되는 경우가 많았고, 전처리를 거치지 않았기 때문에 ‘협업을', ‘협업이’와 같이 조사가 다른 단어를 다른 단어로 인식하여 중복된 키워드가 많아 발생한 차이로 보인다. 전처리를 한 경우에도 어느정도 한계가 존재하였는데, 여러 단어로 이루어진 ‘빅데이터’와 같은 키워드를 ‘빅'과 ‘데이터'로 분리하여 다른 의미를 가지는 키워드를 선택하는 경우가 존재하였다. 전처리를 하지 않아 발생한 문제들은, 키워드의 최소 길이를 설정하거나, 선택된 키워드에서 겹치는 글자가 몇글자 이상이면 하나는 제외하는 방식으로 성능 개선을 이룰 수 있을 것이다.

1. **향후 방향**

ROUGE 지표를 이용한 평가는 하지 못했는데, 이는 국립국어원 모두의 말뭉치에서, 문서요약 말뭉치를 이용하면 될 것으로 보인다.

**참고문헌**

[1] J.-P. Hong, J.-W. Cha, “Korean Important Sentence Extraction using TextRank Algorithms” Proceedings of the Korea Computer Congress, vol. 36(1C), pp. 311-314, 2009.