**한국어 TextRank를 이용한 키워드와 핵심문장 추출**

**정진원, 서태원**

1. **서론**

이전에는 TextRank를 이용하여 영어 텍스트를 요약해보았다. 영어 텍스트에 대해 다양한 전처리 방식 및 유사도 공식을 적용하여 어떤 경우가 더 효율적인지, 각각 어떤 특징을 가지고 있는지 비교하였었다. 이번 프로젝트의 목표는 한글로 작성된 텍스트를 처리하는 것이기 때문에 한글 텍스트에 TextRank를 적용시켜 키워드 추출과 문장에 대한 요약적 추출을 수행하였고, 형태소 분석기(Mecab)의 사용 여부, TF-IDF의 사용 여부, 그리고 논문에 제시된 유사도 공식과 코사인 유사도 공식을 각각 적용시켰을 때에 따른 결과를 확인하였다. 또한, 영어 텍스트에 대해 적용했던 문장 추출 방식들이 한글에 대해서도 효과적인지 확인하였다.

1. **TextRank를 이용한 한국어 키워드 추출**

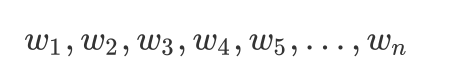
키워드 추출의 경우 co-occurrence라는 relation을 사용한다. 텍스트에 등장하는 단어들을 vertice로 두고, 두개의 단어가 N이라는 window 크기 안에서 같이 등장하면 두 vertice가 연결된 것으로 간주된다. 이 때 N의 크기는 대체적으로 2~10이다. 영어의 경우 window size=2이고, 명사와 형용사만을 사용할 때 키워드 추출의 성능이 가장 높으며, stopwords가 존재하지 않을 때 성능이 더 좋다[1]. 한국어 키워드 추출 또한 명사, 동사, 형용사의 어간의 품사만을 이용하여 단어 그래프를 생성하는 경우 모든 품사의 단어들을 이용할 때보다 성능이 더 우수했다[2]. 따라서 window size=2로 설정한 뒤 명사, 형용사, 동사를 이용하여 키워드를 추출하고, 추가적으로 형태소 분석기를 사용하지 않고 어절을 어근으로 두고 키워드를 추출하여 결과를 비교하였다.

**2.1 키워드 추출 과정** **및 예시**

키워드를 추출하기 위해 먼저 텍스트를 단어로 분리하였다. 이를 위해 다음과 같은 두가지 방법을 사용하였다. (1) Mecab 형태소 분석기를 이용한 전처리 과정을 거쳐 명사, 형용사, 동사를 단어로 선택하거나, (2) 전처리를 거치지 않은 문장의 어절들을 단어로 선택하였다.

예를들어, “주의력 결핍 과잉행동장애(ADHD)는 산만함, 과잉행동, 충동성을 특징으로 하는 질환입니다.”라는 텍스트가 있는경우, 이를 (1)을 이용하여 토큰화 하면 ['주의력', '결핍', '과잉', '행동', '장애', 'ADHD', '과잉', '행동', '충동', '특징', '질환’]이 되고 (2)를 이용하여 토큰화 하면 [‘주의력’, ‘결핍’, ‘과잉행동장애는’, ‘산만함’, ‘과잉행동’, ‘충동성을’, ‘특징으로하는’, ‘질환입니다’]가 된다.

이후 window size에 따라 co-occur하는 단어들을 찾아 그래프를 생성한다. 그림과 같은 n개의 단어들로 텍스트가 구성된 경우



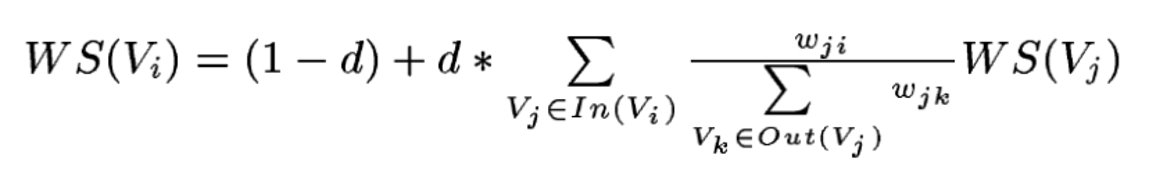
[w1, …, w\_k], [w2, …, w\_{k+1}], [w3, …, w\_{k+2}] ,… 가 window가 되며 window 안에 등장하는 모든 단어들의 쌍은 undirected edge를 가진것으로 취급된다. 앞서 토큰화한 문장 ['주의력', '결핍', '과잉', '행동', '장애', 'ADHD', '과잉', '행동', '충동', '특징', '질환’]에서 window size가 2인 경우의 window를 구해보면 다음과 같은 형식으로 나온다.

[‘주의력’, ‘결핍’, ‘과잉’], [‘결핍’, ‘과잉’, ‘행동’], [‘과잉’, ‘행동’, ‘장애’], [행동, 장애, ADHD], …, [‘충동’, ‘특징’, ‘질환’].

Window안의 단어 쌍들은 undirected edge를 가진 것으로 취급되기 때문에 다음과 같은 단어 쌍에 대해 edge를 연결한다. [(‘주의력', '결핍'), ('주의력', '과잉'), ('결핍', '과잉'), ('결핍', '행동'), ('과잉', '행동'), ('과잉', '장애'), ('행동', '장애'), ('행동', 'ADHD'), ('장애', 'ADHD'), ('장애', '과잉'), ('ADHD', '과잉'), ('ADHD', '행동'), ('과잉', '충동'), ('행동', '충동'), ('행동', '특징'), ('충동', '특징'), ('충동', '질환'), ('특징', '질환')]. 이를 그래프로 나타내면 다음과 같다.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 주의력 | 결핍 | 과잉 | 행동 | 장애 | ADHD | 충동 | 특징 | 질환 |
| 주의력 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 결핍 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 과잉 | 1 | 1 | 0 | 1 | 2 | 1 | 1 | 0 | 0 |
| 행동 | 0 | 1 | 1 | 0 | 1 | 2 | 1 | 1 | 0 |
| 장애 | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 |
| ADHD | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 |
| 충동 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| 특징 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 질환 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

이후 각 단어에 해당하는 vertex의 초기값을 1로 설정하고, 그래프와 초기값을 이용하여 그림 1의 TextRank 알고리즘을 단어들의 TextRank 값이 수렴할때 까지 iterative하게 수행한다. 단어의 중요도는 해당 단어와 연결된 단어들이 얼마나 중요한 단어이고, 얼마나 연관되어 있는지에 따라 결정된다.



최종적으로 높은 값을 가지는 vertex에 해당하는 단어들을 값에 따라 순차적으로 추출한 것이 키워드가 된다..

**2.2 결과 평가**

2.1에서 설명한 텍스트의 단어를 선택하는 두가지 방법을 이용하는 코드를 직접 작성하여 키워드 추출을 해보았다. 성능을 평가하기 위하여 TextRank를 이용하여 추출한 키워드와, 대표님께서 보내주신 예시를 비교하였다.

|  |
| --- |
| 주의력결핍과잉행동장애(ADHD)는 산만함,과잉행동,충동성을 특징으로하는 질환이며, 주요 증상으로는 집중과 집중유지의 어려움, 과도한 집중, 비조직화 건망증, 불안정함 혹은 끊임없는 활동, 충동성, 감정조절의 어려움등이 있으며 문진과 증세를 종합하여 진단하며 중추신경자극제인 메칠페니데이트계통의 약물치료와 인지행동치료방법등이 있습니다. |

|  |
| --- |
| **Keywords assigned by human annotator:**  ‘주의력결핍과잉행동장애’, ‘집중’, ‘충동성’  **Keywords assigned by TextRank(형태소 분석기를 사용한 경우):**  '행동', '집중', '충동', '치료', '어려움', '과잉', '중추', '신경', '자극제', '진단', '메', '계통', '종합', '데이트'  **Keywords assigned by TextRank(형태소 분석기를 사용하지 않은 경우):**  ‘약물치료와', '산만함', '메칠페니데이트계통의', '과잉행동', '중추신경자극제인', '충동성을', '진단하며', '특징으로하는', '질환이며', '주요', '증상으로는', '집중과' |

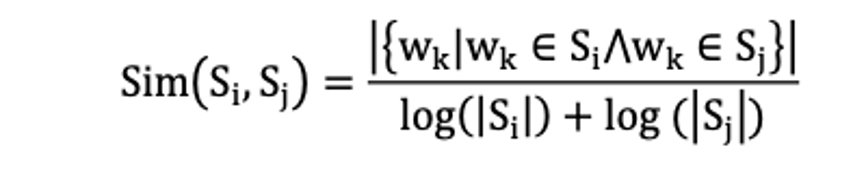
형태소 분석기를 사용한 키워드 추출의 경우 높은 textrank 값을 가지는 더 중요한 단어들에 사람이 선택한 키워드가 포함된 것을 확인할 수 있었다. 또한 ‘키워드'라는 의미에 더 부합하게 주요 단어를 조사나 어미가 빠진, 우리가 알고있는 단어의 형태로 추출하였다. 다만 ‘메칠페니데이트’ 와 같은 형태소 분석기에 저장되지 않은 단어의 경우 ‘메’, ‘데이트'와 같이 형태소가 잘못 분석되어 전혀 관계없는 단어가 키워드로 추출 되는 문제가 있었다. 형태소 분석기를 사용하지 않는 경우에도 사람이 선택한 키워드가 포함되어 있었지만, 더 낮은 중요도를 가지고 있어, 만약 추출된 키워드 중 상위 몇개만 사용한다면 문제가 생길 것으로 보인다. 결과적으로 키워드 추출의 경우 형태소 분석기를 사용하는 방식이 더 효과적이었다.

1. **TextRank를 이용한 한국어 핵심 문장 추출**

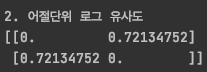
핵심 문장 추출의 경우 문장들을 vertice로 두고, 문장간의 유사도를 weight로 두어 문장간의 연결관계를 나타낸 그래프를 생성하여 그림1의 TextRank 알고리즘을 이용하여 주요한 문장을 추출한다. 문장은 전처리를 거쳐 특정 품사만을 이용하여 나타내거나, 어절 단위로 나타낼 수 있다. 한국어의 경우 어절을 어근으로 두었을 때, 요약 성능이 가장 높았다[3]. 두가지 방법으로 나타낸 문장들 간의 가중치를 논문에서 제안된 유사도 공식, 코사인 유사도 공식을 이용하여 계산하여 문장간의 연결 관계를 나타내는 그래프를 생성하였다. 코사인 유사도를 이용하는 경우 추가적으로 문장들을 단어의 빈도수로 벡터화하지 않고, TF-IDF 방식으로 벡터화 하여 코사인 유사도 공식을 적용시킨 뒤, 결과에 대한 비교를 하였다 [4].

**3.1 문장 추출 과정**

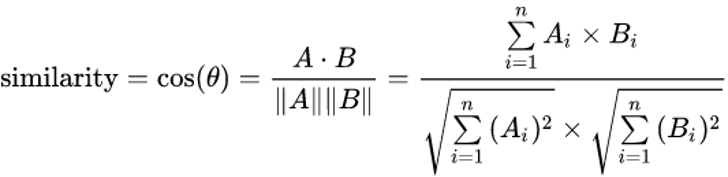
문장은 어절 전체를 하나의 어근으로 추출하거나 형태소 분석기를 이용하여 특정 품사만을 이용할 수 있다. 품사의 경우 2.1과 같은 방법으로 Mecab 형태소 분석기를 이용하였으며 [2]와 같이 명사, 동사, 형용사를 이용하였다. 전처리 과정을 마치면 TextRank를 이용하여 주요 문장 추출을 거치게 되는데, 이를 위해서는 문장간의 가중치를 계산한다.



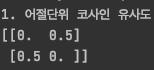
가중치 계산을 위해 그림3의 식을 이용하였다. 이는 문장 간 공통적으로 존재하는 단어들을 이용한 식이며, 문장의 길이가 길어질수록 유사도 식의 값이 커지는 것을 막기 위해 문장의 길이 |Si|에 log를 취해 나눠주는 방식으로 정규화를 해준 것이다. 예를들어, “나는 서강대를 다니는 학생이다.”라는 문장과 “서강대를 다니는 학생은 적다.” 라는 두 문장을 어절로 분리했을 때 유사도를 구한다면 첫번째 문장은 [‘나는', ‘서강대를', ‘다니는', ‘학생이다']로 분리가 되고 두번째 문장은 [‘서강대를', ‘다니는', ‘학생은', ‘적다']로 분리가 된다. 겹치는 단어의 수는 ‘서강대를’과 ‘다니는'으'로 2개이며, 문장의 길이는 각각 4, 4가 된다. 공식을 적용하면 두 문장간의 유사도는 2 / log(4) + log(4) = 0.721이 된다. 코드를 작성하여 계산한 결과 같은 값이 잘 나오는 것을 확인 할 수 있다.



유사도 공식간의 성능 비교를 위하여 그림 4의 코사인 유사도 공식도 이용하였다[1, 2].



코사인 유사도 공식의 경우 [0,1] 사이의 값을 가지며 문장 벡터간 방향이 얼마나 유사한지를 나타내준다. 앞서 사용했던 두 문장에 사용된 단어들은 [‘나는', ‘서강대를', ‘다니는', ‘학생이다', ‘학생은', ‘적다']이고, 이를 이용하여 두 문장을 단어의 빈도수로 벡터화 하면 첫번째 문장은 [1, 1, 1, 1, 0, 0]이 되고, 두번째 문장은 [0, 1, 1, 0, 1, 1]이 된다. 두 문장 간의 코사인 유사도를 계산하면 2 / (2 \* 2) = 0.5가 된다. 코드를 작성하여 계산한 결과 같은 값이 잘 나오는 것을 확인 할 수 있다.



추가적으로 영어의 경우 [4]에서 TF-IDF와 코사인 유사도를 결합한 경우 기존의 TextRank보다 성능이 좋다고 하여, 한국어에서도 동일한 결과를 가지는지 확인하였다. 이후 생성된 그래프와 문장들의 초기 가중치를 1로 두어 그림 1의 TextRank 알고리즘을 TextRank 값들이 수렴할 때 까지 수행시켜, 전체 문장 중 TextRank 값이 상위 10%에 해당하는 문장들을 추출하였다.

**3.2 결과 평가**

총 여섯가지 방법으로 핵심 문장을 추출하여 예시로 주어진 핵심 문장과 얼마나 유사한지 비교하였다. 여섯가지 방법은 어절 단위로 문장을 토큰화하여 (1) 문장간 가중치를 코사인 유사도를 이용하여 계산, (2) 문장간 가중치를 그림 3의 유사도를 이용하여 계산, (3) 코사인 유사도와 TF-IDF 방식을 이용한 방법으로 문장간 가중치를 계산하는 방법과 전처리를 거쳐 특정 품사에 해당하는 단어들로 이루어진 문장들의 가중치를 (1), (2), (3)의 방법을 이용하여 계산하는 방법이다. 입력 텍스트의 10%를 추출하였으며 총 77개의 예시 텍스트에 대해 사람이 선택한 핵심 문장과 6가지 방법으로 추출한 핵심 문장이 얼마나 비슷한지 확인해봤다.

다음은 7개의 결과 중 세개의 결과이다.

**예시 텍스트 1)**

|  |
| --- |
| 주의력 결핍 과잉행동장애(ADHD)는 산만함, 과잉행동, 충동성을 특징으로 하는 질환입니다. 이는 12세 이전 발병하고 만성 경과를 보이며, 여러 기능 영역에 지장을 초래합니다. 이 질환 환자 중에는 도덕적인 자제력 부족이나 반항심, 이기심으로 오해받아 괴로워하는 경우가 많습니다. 대략 3~4:1 정도로 남성에서 흔하게 발생합니다. 초등학생 중 13% 정도, 중고등학생 중 7% 정도가 이 질환을 지니고 있습니다. 성인기에 존재하는 산만함이나 충동성에 대해 별개의 시기에 발현한 성인 ADHD로 진단할 것인가, 이전 시기에 발현한 ADHD의 잔재 증상으로 이해할 것인가, 아니면 전혀 다른 별개의 질환에 의한 증상이 집중력 장애의 형태로 나타난 것인가에 대해서는 현재까지도 활발하게 논의하고 있습니다. |

**결과 1)**

|  |  |  |
| --- | --- | --- |
| 예시 핵심 문장 | | **주의력 결핍 과잉행동장애(ADHD)는 산만함, 과잉행동, 충동성을 특징으로 하는 질환입니다.** |
| 어절 단위 | 코사인 유사도 | 성인기에 존재하는 산만함이나 충동성에 대해 별개의 시기에 발현한 성인 ADHD로 진단할 것인가, 이전 시기에 발현한 ADHD의 잔재 증상으로 이해할 것인가, 아니면 전혀 다른 별개의 질환에 의한 증상이 집중력 장애의 형태로 나타난 것인가에 대해서는 현재까지도 활발하게 논의하고 있습니다. |
| 유사도 공식 | 성인기에 존재하는 산만함이나 충동성에 대해 별개의 시기에 발현한 성인 ADHD로 진단할 것인가, 이전 시기에 발현한 ADHD의 잔재 증상으로 이해할 것인가, 아니면 전혀 다른 별개의 질환에 의한 증상이 집중력 장애의 형태로 나타난 것인가에 대해서는 현재까지도 활발하게 논의하고 있습니다. |
| 코사인 TF-IDF | 성인기에 존재하는 산만함이나 충동성에 대해 별개의 시기에 발현한 성인 ADHD로 진단할 것인가, 이전 시기에 발현한 ADHD의 잔재 증상으로 이해할 것인가, 아니면 전혀 다른 별개의 질환에 의한 증상이 집중력 장애의 형태로 나타난 것인가에 대해서는 현재까지도 활발하게 논의하고 있습니다. |
| 전처리 | 코사인 유사도 | **주의력 결핍 과잉행동장애(ADHD)는 산만함, 과잉행동, 충동성을 특징으로 하는 질환입니다.** |
| 유사도 공식 | **주의력 결핍 과잉행동장애(ADHD)는 산만함, 과잉행동, 충동성을 특징으로 하는 질환입니다.** |
| 코사인 TF-IDF | **주의력 결핍 과잉행동장애(ADHD)는 산만함, 과잉행동, 충동성을 특징으로 하는 질환입니다.** |

**예시 텍스트 2)**

|  |
| --- |
| ADHD는 뇌 안에서 주의집중 능력을 조절하는 신경전달 물질(도파민, 노르에피네프린 등)이 불균형하여 발생합니다. 주의집중력과 행동을 통제하는 뇌 부위의 구조 및 기능 변화가 ADHD의 발생과 관련이 있습니다. 기타 원인으로는 뇌 손상, 뇌의 후천적 질병, 미숙아 등이 있습니다.  소아 ADHD의 유병률은 일반 인구의 6~9%입니다. 이 중 60~80%는 청소년기까지 계속됩니다. 50%, 즉 소아 ADHD 환자 2명 중 1명은 성인이 되어도 ADHD의 주요 증세나 전체 진단 기준을 충족시키는 증세를 유지합니다. |

**결과 2)**

|  |  |  |
| --- | --- | --- |
| 예시 핵심 문장 | | ADHD는 **뇌 안에서 주의집중 능력**을 조절하는 신경전달 물질(도파민, 노르에피네프린 등)이 불균형하여 발생합니다. |
| 어절 단위 | 코사인 유사도 | 소아 ADHD의 유병률은 일반 인구의 6~9%입니다. |
| 유사도 공식 | 소아 ADHD의 유병률은 일반 인구의 6~9%입니다. |
| 코사인 TF-IDF | 소아 ADHD의 유병률은 일반 인구의 6~9%입니다. |
| 전처리 | 코사인 유사도 | 50%, 즉 소아 ADHD 환자 2명 중 1명은 성인이 되어도 ADHD의 주요 증세나 전체 진단 기준을 충족시키는 증세를 유지합니다. |
| 유사도 공식 | **주의집중력과** 행동을 통제하는 **뇌 부위**의 구조 및 기능 변화가 ADHD의 발생과 관련이 있습니다. |
| 코사인 TF-IDF | **주의집중력과** 행동을 통제하는 **뇌 부위**의 구조 및 기능 변화가 ADHD의 발생과 관련이 있습니다. |

**예시 텍스트 3)**

|  |
| --- |
| ADHD 증상은 환자의 연령대가 증가함에 따라 변화합니다. 과다 활동은 초기 청소년기로 접어들면서 유의미하게 감소합니다. 일부 환자는 충동성, 심한 감정 기복, 주의집중력에서 지속적인 결함을 보입니다. 이를 포함한 성인 ADHD의 증상은 다음과 같습니다.  ① 집중과 집중 유지의 어려움  - 아주 간단한 일임에도 일을 끝마치기 위해 고군분투합니다.  - 일을 끝마치지 못합니다.  - 세밀한 부분을 간과하는 실수가 잦습니다.  - 별로 상관없는 광경이나 소리 때문에 쉽게 산만해집니다.  - 한 가지 일을 하다가 어느새 다른 일을 하고 있습니다.    ② 과도한 집중  - 책, TV, 컴퓨터 등 흥분과 보상이 있는 일에는 몰입합니다.  - 과도한 집중으로 인해 다른 중요한 일과 시간 개념을 잊어버립니다.    ③ 비조직화와 건망증  - 정리 정돈을 잘하지 못합니다. 방, 책상, 차가 매우 어지럽습니다.  - 일의 예상 소요시간을 과소평가하는 경향이 있습니다.  - 만성적으로 지각합니다.  - 우선순위를 정하지 못하거나 계획적으로 행동하지 못합니다.  - 물건을 잃어버리거나 제자리에 놓지 않습니다.    ④ 불안정함 혹은 끊임없는 활동  - 가만히 앉아 있는 것을 어려워합니다.  - 자극적이고 흥분되는 일을 추구합니다.  - 동시에 여러 가지 일을 합니다.  - 쉽게 지루해합니다.    ⑤ 충동성  - 다른 사람의 대화에 자주 끼어듭니다.  - 자제를 잘 못합니다.  - 무례하거나 부적절한 생각을 그대로 내뱉습니다.  - 결과를 고려하지 않고 돌발적으로 행동합니다.  - 중독의 위험이 있습니다.    ⑥ 감정 조절의 어려움  - 자존감과 성취감이 낮습니다.  - 비판에 대해 과민 반응하며 쉽게 좌절합니다.  - 감정 기복이 심하고 조급합니다.  - 예민하고 폭발적으로 화를 냅니다. |

**결과 3)**

|  |  |  |
| --- | --- | --- |
| 예시 핵심 문장 | | ADHD의 증상으로는 **집중과 집중 유지의 어려움**, **과도한 집중**,비조직화와 건망증,불안정함 혹은 끊임없는 활동,**충동성, 감정 조절의 어려움**등이 있습니다. |
| 어절 단위 | 코사인 유사도 | - 한 가지 일을 하다가 어느새 다른 일을 하고 있습니다.  - 일을 끝마치지 못합니다.  ① **집중과 집중 유지의 어려움**  - 아주 간단한 일임에도 일을 끝마치기 위해 고군분투합니다. |
| 유사도 공식 | - 한 가지 일을 하다가 어느새 다른 일을 하고 있습니다.  - 일을 끝마치지 못합니다.  ① **집중과 집중 유지의 어려움**  - 아주 간단한 일임에도 일을 끝마치기 위해 고군분투합니다. |
| 코사인 TF-IDF | - 한 가지 일을 하다가 어느새 다른 일을 하고 있습니다.  - 일을 끝마치지 못합니다.  ① **집중과 집중 유지의 어려움**  - 아주 간단한 일임에도 일을 끝마치기 위해 고군분투합니다. |
| 전처리 | 코사인 유사도 | ① **집중과 집중 유지의 어려움**  - 아주 간단한 일임에도 일을 끝마치기 위해 고군분투합니다.  **일부 환자는 충동성, 심한 감정 기복, 주의집중력에서 지속적인 결함**을 보입니다.  - **과도한 집중**으로 인해 다른 중요한 일과 시간 개념을 잊어버립니다. |
| 유사도 공식 | **일부 환자는 충동성, 심한 감정 기복, 주의집중력에서 지속적인 결함**을 보입니다.  - **과도한 집중**으로 인해 다른 중요한 일과 시간 개념을 잊어버립니다.  ② **과도한 집중**  - 책, TV, 컴퓨터 등 흥분과 보상이 있는 일에는 몰입합니다. |
| 코사인 TF-IDF | **일부 환자는 충동성, 심한 감정 기복, 주의집중력에서 지속적인 결함**을 보입니다.  - **과도한 집중**으로 인해 다른 중요한 일과 시간 개념을 잊어버립니다.  ① **집중과 집중 유지의 어려움**  - 아주 간단한 일임에도 일을 끝마치기 위해 고군분투합니다. |

결과적으로 큰 차이를 만든 것은 문장간의 가중치 계산 방법보다는, 형태소 분석기의 사용 여부였다. 어절 단위로 문장을 분리하여 핵심 문장을 추출하는 경우 세가지 그래프 구성 방식 모두 같은 문장을 추출하였다. 공식상으로 코사인 유사도 공식은 짧은 문장을 선호할 것이고, 논문에서 제안된 유사도 공식의 경우 긴 문장을 선호할 것으로 예측하였지만 그런 경향성은 확인 할 수 없었다. 하지만, 추출된 문장이 사람이 선택한 핵심 문장과 대부분 무관하였다. 형태소 분석기를 사용하여 전처리를 한 경우에는 문장간의 가중치 계산 방법에 따라 추출되는 문장에 차이가 존재하였는데, 코사인 TF-IDF 방식의 정확도가 가장 높았다. 또한 추출된 핵심 문장이 대부분 인간이 선택한 문장과 일치하거나 관련이 있었다. 앞으로는 형태소 분석기를 사용하지 않고 추출적 요약의 성능을 높이는 방법을 찾아야할 것 같다.

**참고문헌**

[1] Rada Mihalcea, Paul Tarau, “TextRank: Bringing Order into Texts”, Proceedings of the European Conference on Artificial Intelligence (ECAI 2004), Valencia, Spain, August 2004.

[2] [TextRank 를 이용한 키워드 추출과 핵심 문장 추출 (구현과 실험)](https://lovit.github.io/nlp/2019/04/30/textrank/)

[3] 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.

[4] F. Barrios, F. Lopez, L. Argerich, R. Wachenc hauzer, “Variations of the Similarity Function of TextRank for Automated Summarization,” Proceedings of the Argentine Symposium on Artificial Intelligence, pp. 65-72, 2016.