의사 논문 자료 분석

: self-attention LSTM, LSTM을 활용한 논문의 전문도(과 유사도) 정의

데이터

-분석할 논문 데이터 ( 29843 rows )

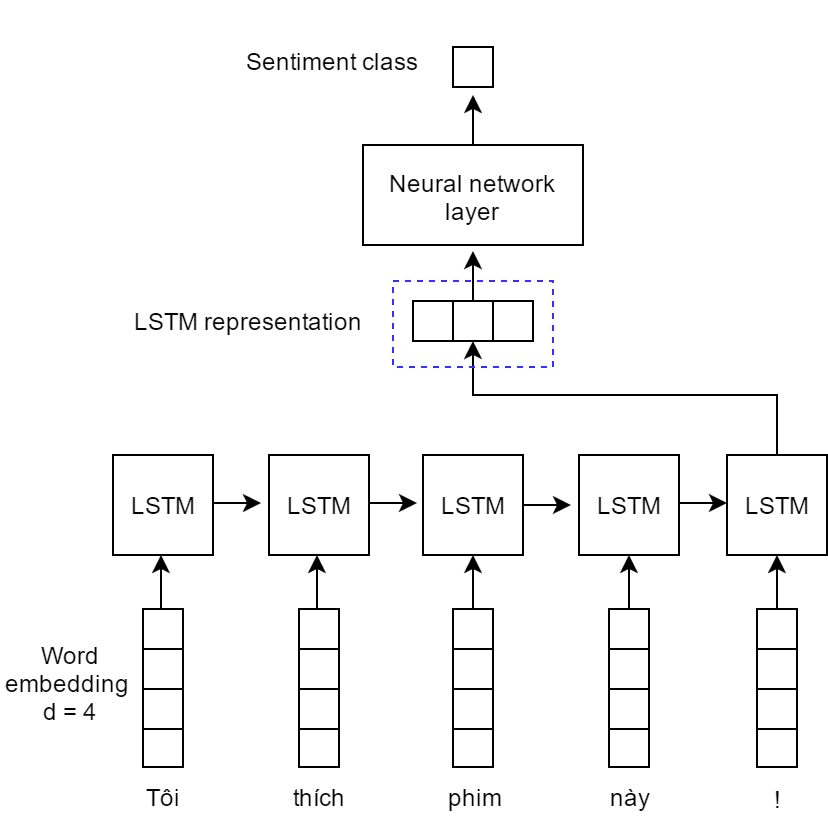
: 한국에 치과의사들의 논문 데이터, 크롤링하여 논문의 초록과 저자를 데이터 형식을 표준화

-학습용 논문, 저널 데이터 ( 24900 rows )

: 데이터 학습을 위한 각 과에 포함된 저널, 논문의 데이터를 크롤링하여 수집

분석 기법

-LSTM



: 필요한 정보를 얻기 위한 시간 격차에 문제를 가진 RNN 모델의 문제를 극복한,

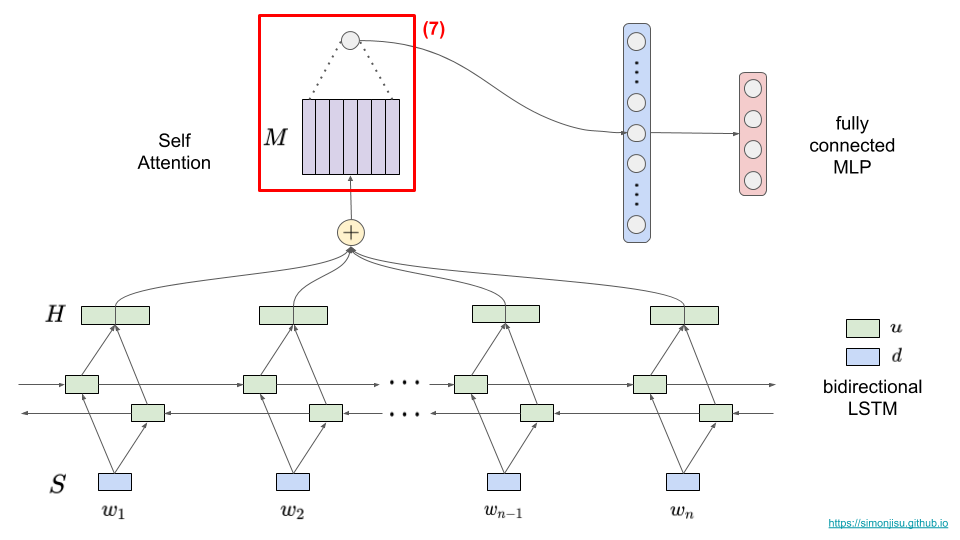
시간적 격차 정보를 기억하는 모델

LSTM은 cell state에 뭔가를 더하거나 없앨 수 있는 능력이 있는데, 이 능력은 gate라고 불리는 구조에 의해서 제어

Gate는 정보가 전달될 수 있는 추가적인 방법으로, sigmoid layer와 pointwise 곱셈으로 이루어져 있다.

|  |  |
| --- | --- |
| 좌) RNN 모형 | 우) LSTM 모형 |

-Self-Attention Lstm

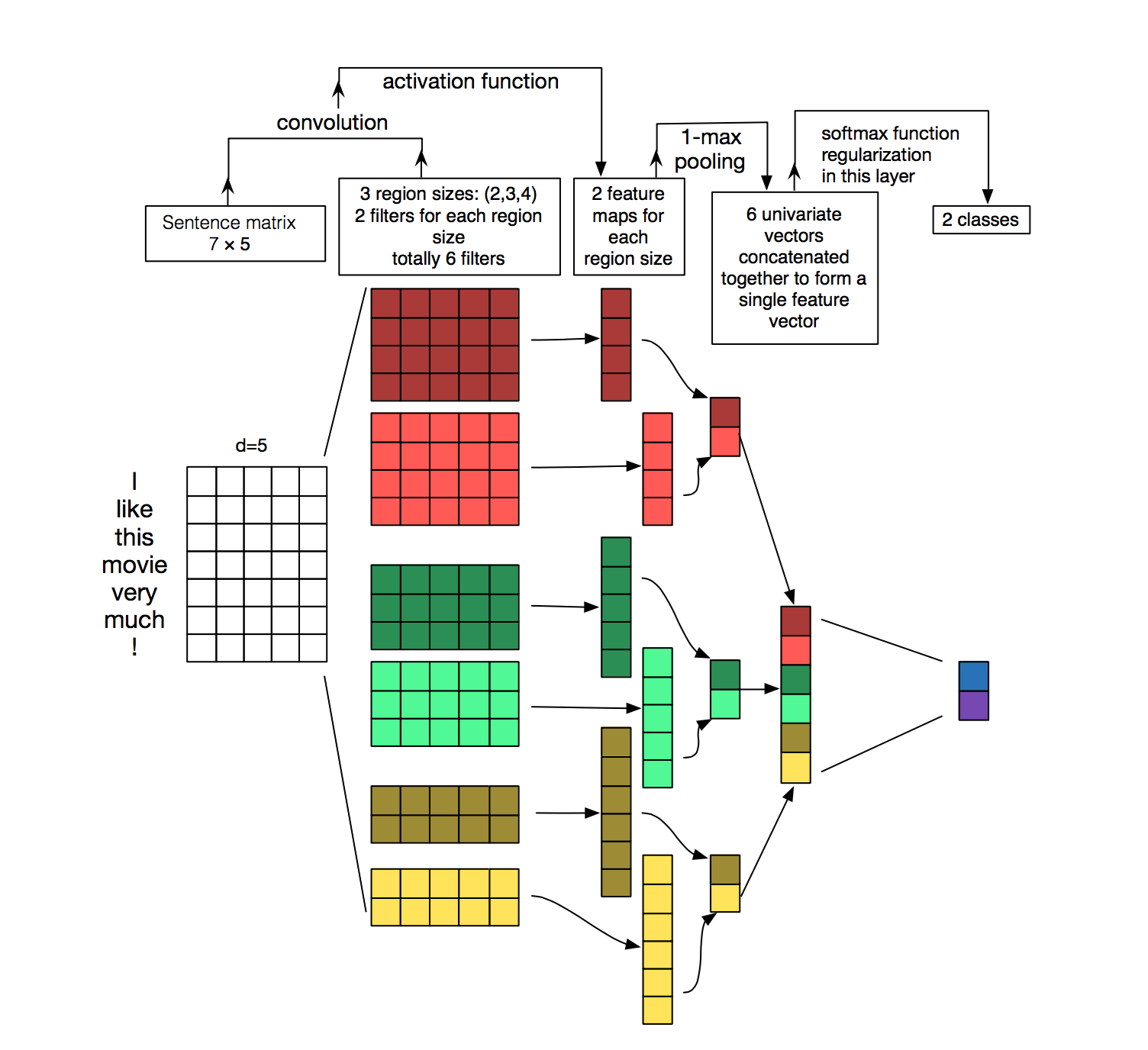


: 기존 lstm에 추가로 문장의 단어의 유사도에 대한 softmax를 추가하여

문장들 사이의 단어 혹은 구문이 될 확률분포를 통한 문장의 다양한 구성의 중요도를 가질 수 있다.

그로 인해 문장의 다양한 부분을 학습이 가능하다.

-CNN



: 신경망을 구성하는데 텍스트 임베딩의 필터링된 합성곱(CONV), 활성화 함수(ReLU), 맥스 풀링(POOL) 과정을 반복하여 피쳐 벡터를 형성하고 학습된 텍스트와 비교하여 유사도를 판별한다.

결과

: 총 24900개의 논문 데이터를 학습

한국 치과의사 29843개의 데이터, CNN 모델을 통한 과별 연관도 측정

(추가중) 데이터 가산 부분 추가

시행착오

--데이터 형태의 불일치

|  |
| --- |
|  |
| 데이터 형태의 불일치 논문과 국내 치과 의사  연결에 문제 발생 |

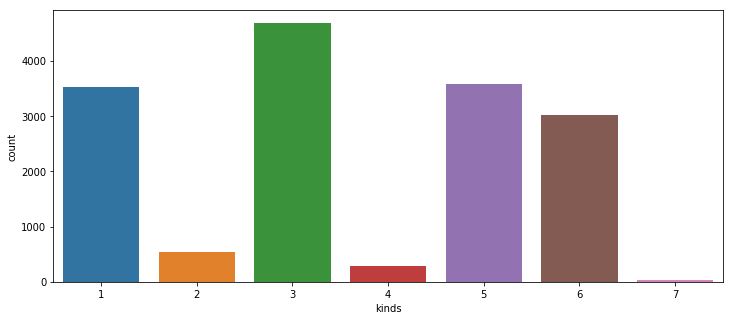
|  |
| --- |
|  |
| 논문 데이터와 의사 데이터를 연결하기 위해  의사의 정보 크롤링 논문과 의사의 속성으로 매칭을 실시 |

|  |  |
| --- | --- |
|  | docid : 의사의 개인을 고유 id 지정  reportid : 논문 고유 번호를 지정  의사와 연결된 논문을 매칭 |

-학습용 데이터의 불균형

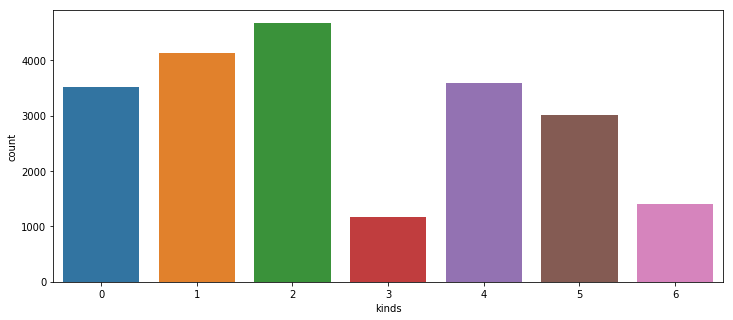
각 학과의 학습용 논문 데이터 각 학과의 논문 데이터가 불균형을

이루고 있어 학습하는 모델의 결과에 영향을 미치게 된다.



각 학과의 학습용 데이터의 부족한 부분을

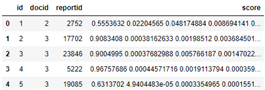
새로운 자료의 크롤링을 통하여 보충



-데이터 가중치

: 의사의 양력과 치과 관련 정보에 대한 수치화에 대한 문제, 각 의사 데이터(논문, 강의경력, 대외 활동)에

대한 적절한 값을 분석 또는 heuristics한 방법 으로 모색



분석 시 문제

-모델별 정확도 차이

: 동일 조건( 학습, 검증 데이터 : cleaned\_reports2 )하에서 다른 정확도와 테스트 결과가 나타난다.

|  |  |
| --- | --- |
| Model | Accuracy |
| LSTM | 78.12% |
| SelfAttentionLstm | 40.90% |
| CNN | 87.88% |

-Colab

: <cuda error device-side assert triggered>

정확한 해결 방법을 알지 못 하는 에러가 발생

<[https://colab.research.google.com](https://colab.research.google.com/)>

많은 작업 시에 발생 가능성 예상

-데이터 맞는 분석 기법

: 논문의 형태에 비롯해 구문체가 아닌 문어체의 형식으로 글의 문맥을 크게 고려하지 않는 데이터로 판단된다. 모델의 검증 결과 CNN이 가장 큰 검증 값을 가져오면

실제 데이터에 적용 시에도 가장 적합한 모델로 사료된다.

-하이퍼 파라미터에 대한 패착

: 각 모델에는 모든 부분이 계산을 이루어 지지않는 부분이 존재, 모델의 하이퍼 파라미터라고 하는 이 부분은 결과 값에 큰 영향을 미치지만

어떠한 근거에 기대어 값을 정하기가 어렵다는 단점을 가짐, 하여 이 부분은 수차례 직접적인 테스트를 거쳐서 가장 알맞은 수치를 적용해 가는 heuristics한 방법으로 접근하고 해결

-데이터 가중치

: 의사의 양력과 치과 관련 정보에 대한 수치화에 대한 문제

의사의 데이터에 편향과 편차가 있어 데이터의 변환을 시도하였지만 데이터의 원래의 속성을 잃어버리기 때문에

데이터의 손상을 방지하며 의사의 데이터를 빠짐 없이 적합 시키기 위해 새로운 가중치 값을 산출

|  |  |
| --- | --- |
|  |  |

좌) 의사 논문 데이터의 수치 변환전 우)의사 논문데이터의 LOG 변환후

전체 의사 논문 수치화된 논문 데이터를 의사가 작성한 논문의 수로 나누어 평균의 값을 취해주고 를 활용하여 의사 논문 개수에 대한 데이터를 가중치로 또 활용해준다.

Log값을 사용하면 논문 수에 대한 편차는 크지 않게 조절.

각 의사 데이터(논문, 강의경력, 대외 활동)에 대한 적절한 값을 분석 또는 heuristics한 방법으로 모색