입력 데이터 형식 및 Positive/Negative에 따른 한국어 증상 기반 질병 예측 모델

김민정[1], 조인휘[2]
[1]한양대학교 컴퓨터·소프트웨어학과 석사과정
[2]한양대학교 컴퓨터·소프트웨어학과 교수
rlaalswjd405@hanyang.ac.kr, iwjoe@hanyang.ac.kr

Korean Symptom-Based Disease Prediction Model according to Input Data Format and Positive/Negative

Min-Jung Kim[1], In-Whee Joe[2]
[1]Dept. of Computer Software, Han-Yang University
[2]Dept. of Computer Software, Han-Yang University

요 약

본 논문은 Word2Vec를 이용하여 한국어 증상 기반 질병 예측 모델을 제시한다. 아산병원 질환 백과의 크롤링 데이터를 세 가지 형식으로 나누어, 모델에 알맞은 데이터 형식을 찾고 모델에 적용한다. 가장 모델에 맞는 데이터 형식은 증상별 질병과 질병별 증상을 합친 경우이다. 데이터의 양을 늘려 임베딩 스페이스를 넓혔고, 가장 중요한 증상과 질병의 유사도도 정확하게 출력되었다. 이는 유사도가 높은 질병과 증상들이 제대로 학습이 되었다는 것을 알 수 있다. 이렇게 만들어진 예측 모델에 positive 증상을 입력하면 유사도가 향상되고, negative에 입력하면 하락하는 결과를 확인했다. 따라서 환자의 증상을 positive에 넣으면, 그 증상을 가진 질병이 가까워지는 반면, 환자의 증상이 아닌 증상을 negative에 넣으면, 환자에게 맞지 않는 질병이 멀어진다. 그러므로 환자의 상태에 맞는 질병을 유추해, 의사나 환자가 증상에 대한 질병을 알고 싶을 때 또는 검색에 유용하게 사용할 수 있다. 더불어, 질병의 진료과 데이터를 추가하여, 환자에게 맞는 진료과를 찾는 데도 도움을 줄 수 있다.

1. 서론

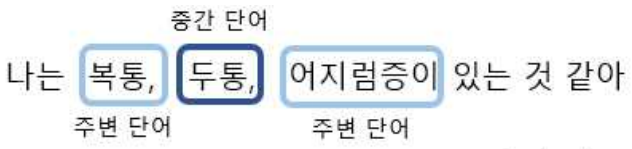
영어로 된 증상 기반 질병 예측 모델[1]은 찾기가 쉽지만, 한국어는 다른 언어들과 다르게 자음과 모음이 언제나 같이 발생한다는 규칙의 모아쓰기나 풀어쓰기와 같이 특정한 특징을 가지고 있다. 또 데이터의 양도 현저히 적어 연구에 어려움을 가지고 있다.

그래서 본 논문에서는 적은 데이터양을 늘리기 위해 데이터 형식을 다르게 하여 모델 학습에 맞는 데이터를 찾는다. 또, 학습한 질병과 증상 쌍을 토대로 환자가 가진 질병을 예측하는 모델을 제시한다.

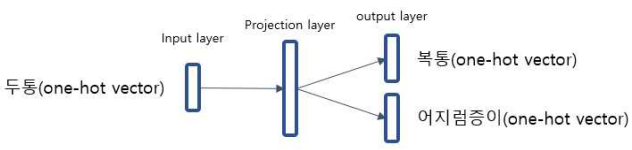
환자에게 맞는 증상과 맞지 않는 증상을 수집하고, 증상 기반 질병 예측 모델의 positive와 negative에 입력하여, 환자의 상태에 가장 잘 맞는 질병이 무엇인지 분류 및 예측할 수 있다.

2. Word2Vec

Word2Vec는 단어와 단어 사이의 의미에 대한 유사성을 확인하는, 워드 임베딩의 한 방법이다. 단어 간 관계를 파악하여 특정 단어나 다음 단어의 예측이 가능하다. 단어들의 유사성은 벡터 간 유사도인 코사인 유사도를 이용하여 측정한다. CBOW(Continuous Bag of Words)와 Skip-gram 두 가지 방식이 있으며, 본 논문에서는 Skip-gram 방식을 적용하였다.



(그림 1) window size.



(그림 2) Skip-gram.

Skip-gram[2]은 중간에 있는 단어를 이용하여 주변 단어들을 유추해 내는 방식이다. CBOW보다 예측 난도가 높기 때문에, 단어의 분산 표현이 뛰어나 더 좋은 성능을 가진다. 그림 1과 2는 Skip-gram의 예측 원리 예시이며, 그림 1은 예측에 필요한 주변 단어의 수를 결정하는 window size를 1로 설정하였을 때이다. 본 논문에서는 여러 경우로 실험해 보았을 때 가장 성능이 좋았던, 각 리스트의 평균 길이의 절반인 5로 window size를 설정하였다. 이는 리스트 내의 모든 단어를 Word2Vec가 학습할 수 있는 방법이다.

3. 질병 예측 모델

3.1. 데이터 형식

데이터는 아산병원 홈페이지의 질환 백과에서 질환명과 증상을 크롤링하여 얻었다. 크롤링 데이터를 이용하여, 입력 데이터의 형식을 다양하게 바꿔 성능을 높인다.

3.1.1. 질병과 증상

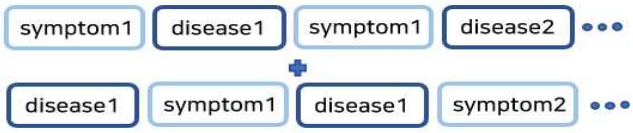
질병과 증상의 데이터 형식은 질병명과 그 질병에 대한 증상들을 차례대로 하나의 리스트에 넣은 것이다. 형태는 그림 3과 같고, 본 논문에서는 dis_sym_list라고 부른다. 예를 들면, [‘RS바이러스 감염증’, ‘열’, ‘청색증’, ‘기침’, ‘가래’, 코막힘‘, ’빈호흡‘, ’콧물‘]이다.

3.1.2. 증상별 질병

증상별 질병은 하나의 증상을 징검다리 형식으로 나열하고, 사이에 그 증상을 가지는 질병들을 추가하는 방식을 가진다. 예를 들면, [’재채기‘, ’만성비염‘, ’재채기‘, ’냉방병‘, ’재채기‘, ’상기도 감염‘, ’재채기‘, ’꽃가루 알레르기‘, ’재채기‘, ’알레르기 비염‘]이고 그림 4와 같은 형태를 하고 있으며, 본 논문에서는 dis_by_sym_list라 부른다.

(그림 3) 질병과 증상 데이터 형식.

(그림 4) 증상별 질병 데이터 형식.



(그림 5) 증상별 질병 및 질병별 증상 데이터 형식.

3.1.3. 증상별 질병과 질병별 증상

증상별 질병과 질병별 증상은 본 논문에서 dbs_sbd_list라고 부르며, 위의 증상별 질병에 질병별 증상을 합쳐 만든 형식이다. 질병별 증상은 하나의 질병을 징검다리 형식으로 나열하고, 사이에 그 질병이 가지는 증상들을 나열하는 방식이다. 본 논문에서는 이를 sym_list라 한다. 데이터의 예는 [‘결핵’, ‘객혈’, ‘결핵’, ‘열’, ‘결핵’, ‘기침’, ‘결핵’, ‘가래’, ‘결핵’, ‘체중감소’]으로, 그림 5와 같다.

3.2. 데이터 형식에 따른 결과값

Word2Vec 모델에 데이터 형식들을 적용해, 증상 '가래'를 입력하여 유사도[3]를 확인하였다.

표 1은 dis_sym_list를 적용한 결과이다. 증상과 유사한 증상들은 잘 랭크되었다. 하지만 질병과 증상이 골고루 학습되지 못해, 증상을 입력했을 때 관련 질병이 출력되지 못한다. 또 순위에 따른 유사도 차이가 0.001 이하로 매우 근소하다. 이는 임베딩 스페이스가 좁아 모든 단어가 서로 연관성이 높다고 계산었기 때문이다.

<표 1> dis_sym_list 적용 결과표

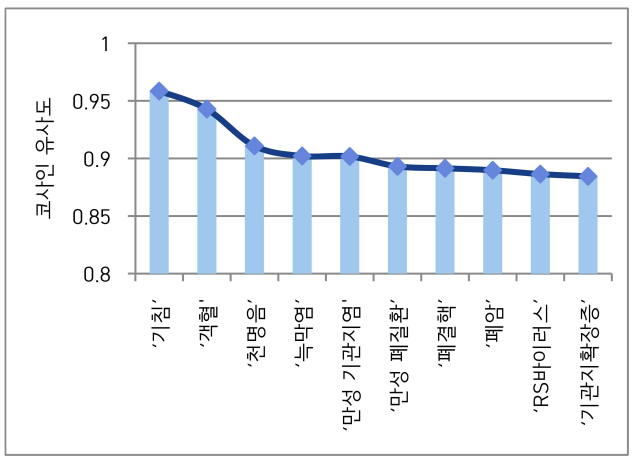
‘가래’와 유사한 단어	코사인 유사도
‘기침’	0.988510966
‘열’	0.987049400
‘빈혈’	0.987007200
‘복부 통증’	0.986711204
‘호흡곤란’	0.985920011
‘체중감소’	0.985832214
‘창백’	0.985738575
‘관절통’	0.985707879
‘피로감’	0.985691487
‘경련’	0.985426485

<표 2> dis_by_sym_list 적용 결과표

'가래'와 유사한 단어	코사인 유사도
'승모판 협착'	0.985757946
'만성 기관지염'	0.985002160
'미만성 간질성 폐질환'	0.984558224
'늑막염'	0.984535992
'결핵'	0.984464526
'만성 폐쇄성 폐질환'	0.984446465
'농흉'	0.984268248
'폐 혈전색전증'	0.984052538
'지역사회성 폐렴'	0.983957469
'재생불량빈혈'	0.983936429

<표 3> dbs_sbd_list 적용 결과표

'가래'와 유사한 단어	코사인 유사도
'기침'	0.958554208
'객혈'	0.942871809
'천명음'	0.910903155
'늑막염'	0.902159273
'만성 기관지염'	0.901782453
'만성 폐쇄성 폐질환'	0.893136084
'폐결핵'	0.891491472
'폐암'	0.889801084
'RS바이러스 감염증'	0.886446952
'기관지확장증'	0.884542942



(그림 6) dbs_sbd_list 적용 결과 그래프.

증상과 질병이 짝을 이뤄 학습할 수 있도록, dis_by_sym_list를 모델에 적용한 결과가 표 2이다. dis_sym_list 형식을 사용했을 때와는 다르게, 증상 '가래'와 연관성이 높은 질병들이 top10에 올랐다. 하지만 마찬가지로 첫 번째와 마지막의 유사도가 0.9857, 0.9839로, 0.0018 차이밖에 나지 않는다.

그러므로 임베딩 스페이스를 넓히기 위해 데이터의 양을 늘려준다. dis_by_sym_list에 sym_list를 더한 dbs_sbd_list를 모델에 적용하였다. 표 3과 그림 6처럼, 증상 '가래'와 유사도가 높은 증상과 질병들이 출력되었으며, 순위 간 차이도 벌어진 것을 볼 수 있다.

3.3. positive와 negative에 따른 결과값

모델을 이용해 유사도를 측정할 때, positive에 증상을 입력하면 그 증상이 가까워지고, negative에 입력을 추가하면 증상과 멀어지는[4,5] 결과값을 확인한다.

본 논문에서는 예시로 폐렴을 이용하였고 데이터 형식은 3-2에서 확인한 dbs_sbd_list를 적용했다. 폐렴의 증상에는 '흉수', '오한', '열', '가래', '가슴 통증', '호흡곤란', '피로감', '두통', '기침'이 있다.

3.3.1. positive에 증상 입력

폐렴의 증상을 positive에 넣어주면, 그림 7과 같이 폐렴을 비롯해 입력 증상들을 가진 질병들이 출력된다.

두 번째와 세 번째인 '메리오이드증'과 '코로나-19'에 없는 폐렴의 증상을 positive에 넣었을 경우, 입력값은 '오한', '가슴 통증', '흉수'이다. 결과는 그림 8처럼 폐렴은 그대로 첫 번째로 출력되고, '메리오이드증'은 0.9786에서 0.9725로, '코로나-19'는 0.9766에서 0.9632로 유사도가 낮아져 순위가 떨어졌다.

3.3.2. negative에 증상 추가 질병

negative에 '메리오이드증' 증상을 넣었을 경우, 그림 9와 같이 '메리오이드증'이 폐렴과 멀어져 top10에 순위 하지 못했다. 이때, positive에 '호흡곤란', '오한', '두통', '가슴 통증', '가래', '열', '기침', '피로감', '흉수'를 입력, negative에 '근육통'. '권태감', '객혈', '사망'을 입력해 주었다.

'메리오이드증'을 제외하면서 negative에 입력했던 '근육통' 때문에, '근육통' 증상을 가지고 있는 '코로나-19'도 유사도가 0.9632에서 0.9152로 감소하여 함께 멀어졌다. 따라서 '근육통'을 positive로 옮기고, 다시 유사도를 측정했다. 그 결과 그림 10과 같이 '코로나-19'가 두 번째 순위로 상승해, '폐렴'과 다시 가까워진 것을 알 수 있다. 따라서 negative에 증상을 넣었을 때, 증상 간의 거리가 멀어지고 관련 질병 예측에 도움을 줘, 모델의 성능을 높인다.

```
('폐렴', 0.986514389)
('메리오이드증', 0.978627383)
('코로나-19', 0.976646542)
('늑막염', 0.975731909)
('히스토플라즈마증', 0.97013843)
('병원 감염성 폐렴', 0.966601967)
('폐렴간균에 의한 폐렴', 0.958607554)
('흑사병', 0.958358049)
('중증 급성 호흡기 증후군', 0.957951664)
('흡입성 폐렴', 0.955983638)
```

(그림7) 폐렴 증상을 positive에 입력했을 경우.

```
('폐렴', 0.979864656)
('늑막염', 0.975933372)
('히스토플라즈마증', 0.97368133)
('메리오이드증', 0.97250837)
('병원 감염성 폐렴', 0.971541843)
('폐렴간균에 의한 폐렴', 0.967944502)
('종격동염', 0.966376423)
('흡입성 폐렴', 0.963449299)
('코로나-19', 0.963248074)
('중증 급성 호흡기 증후군', 0.958933174)
```

(그림 8) 메리오이드증과 코로나-19에 없는 증상만 positive에 입력했을 경우.

```
('폐렴', 0.943209826)
('늑막염', 0.918830215)
('흑사병', 0.916348993)
('히스토플라즈마증', 0.915721893)
('코로나-19', 0.915261566)
('만성 폐쇄성 폐질환', 0.909665822)
('악성 중피종', 0.908742249)
('대동맥판막 협착', 0.90849626)
('지역사회성 폐렴', 0.906522631)
('천명음', 0.905910909)
```

(그림9) 메리오이드 증상을 negative에 입력했을 경우.

```
('폐렴', 0.965734601)
('코로나-19', 0.96438837)
('메리오이드증', 0.95087552)
('중증 급성 호흡기 증후군', 0.945546388)
('히스토플라즈마증', 0.943525493)
('병원 감염성 폐렴', 0.940618872)
('늑막염', 0.939854443)
('상기도 감염', 0.931923985)
('조류 인플루엔자', 0.928457915)
('발진열', 0.928378701)
```

(그림 10) 근육통을 negative에서 positive로 바꾸기.

4. 결론

아산병원 질환 백과의 크롤링 데이터를 학습한 질병 예측 모델의 학습 데이터 형식은 가장 알맞은 형식인 증상별 질병과 질병별 증상을 합친 형식을 사용하였다. 이는 유사도가 높은 질병과 증상들이 적절한 차이를 가져, 가장 적절한 데이터 형식이라 판단하였다.

알아낸 데이터 형식을 적용한 Word2Vec 모델의 positive에 환자가 가지고 있는 증상을 입력한다. 그러면 그 증상에 맞는 증상이나 질병과 가까워지고, 가지고 있지 않은 증상을 negative에 입력하면 멀어지는 결과를 확인하였다. 이를 통해 환자 개인의 현재 상태에 맞는 질병을 유추해 낼 수 있으며, 직접 질병을 찾아 검색에 사용하거나 병원을 찾을 때 도움을 줄 수 있다.

참고문헌

[1] Nnoaham, Kelechi E., et al. "Developing symptom-based predictive models of endometriosis as a clinical screening tool: results from a multicenter study.", Fertility and Sterility, Volume 98, Issue 3, Pages 692-701, 2012.

[2] McCormick Chris. “Word2vec tutorial-the skip-gram model.”, [online] Available: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/, 2016.

[3] Jatnika, Derry, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. “Word2Vec Model Analysis for Semantic Similarities in English Words.”, Procedia Computer Science, Volume 157, Pages 160-167, 2019.

[4] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems, Volume 26, 2013.

[5] Goldberg, Yoav, and Omer Levy. "word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method.", arXiv preprint arXiv:1402.3722 ,2014.