의료서비스를 위한 키워드와 문서의 연관성 향상을 위한 LSTM모델 설계

김준겸 · 서진범 · 조영복*

대전대학교

LSTM Model Design to Improve the Association of Keywords and Documents for Healthcare Services

June-gyeom Kim · Jin-beom Seo · Young-bok Cho*

Daejeon University

E-mail : lab8456@gmail.com

요 약

현재 다양한 검색엔진들이 사용되고 있다. 검색엔진은 메타태그 정보를 기본으로 크롤링, 색인생성, 검색 결과 출력의 3단계를 거치며, 사용자가 원하는 자료의 검색을 도와준다. 그러나 키워드를 기반으로 검색해서 얻은 방대한 문서가 관련이 없거나 적은 문서일 경우도 많다. 이러한 문제점 때문에 검색결과에서 내용을 파악하여 정확도를 분류를 해야 하는 번거로운 일이 발생하게 된다. 다양한 검색엔진을 통해 추출된 결과의 경우 검색엔진의 인덱스는 주기적으로 업데이트 되지만 가중치에 대한 기준과 업데이트 주기는 검색엔진마다 다르고 검색 순위 산정 기준이 서로 다르기 때문에 동일한 키워드를 검색어로 입력하고도 서로 다른 검색 순위를 보여주는 단점을 가지고 있다 따라서 본 논문에서는 기존 검색엔진 대신 사용자가 입력한 키워드와 문서의 연관성을 추출하여 사용자가 찾고자 하는 키워드를 입력했을 때 키워드와 문서의 연관성을 향상 시킬 수 있는 LSTM모델을 설계하고자 한다.

ABSTRACT

A variety of search engines are currently in use. The search engine supports the retrieval of data required by users through three stages: crawling, index generation, and output of search results based on meta-tag information. However, a large number of documents obtained by searching for keywords are often unrelated or scarce. Because of these problems, it takes time and effort to grasp the content from the search results and classify the accuracy. The index of search engines is updated periodically, but the criteria for weighted values and update periods are different from one search engine to another. Therefore, this paper uses the LSTM model, which extracts the relationship between keywords entered by the user and documents instead of the existing search engine, and improves the relationship between keywords and documents by entering keywords that the user wants to find.

키워드

양방향LSTM, 어휘적 중의성, 유사도 측정, 메타태그 방식

Ⅰ. 서 론

현재 자료 수집 등의 목적으로 구글, 네이버, 다음 DBPia등 우수한 검색 엔진들이 사용되고 있다. 검색 엔진은 링크와 텍스트 기반 인식을 기본으로 크롤러가 웹페이지에서 새로운 링크를 발견하면 웹 서버에 데이터 정보를 요청하고 웹 서버는 검색엔진에게 웹페이지 정보를 전송하는 메타태그 방식을 사용한다. 즉 크롤링, 색인생성, 검색결과

* corresponding author

출력 3단계를 거쳐 사용자가 원하는 자료의 검색 결과를 나타낸다. 그러나 한글검색에서 키워드를 기반으로 검색하는 경우 동음이의어를 포함한 결과로 대량의 검색결과를 보이나 검색의 정확도가 낮아질 수 있다. 또한 검색결과에서 내용을 파악하여 정확도를 분류함으로 정확도를 높여야 하는 번거로운 일이 발생하게 된다. 이러한 문제점들의 원인은 검색엔진을 통해 추출된 결과의 경우 검색엔진은 인덱스를 주기적으로 업데이트 하게 되는데 가중치에 대한 기준과 업데이트 주기는 검색엔진마다 다르고 검색 순위 산정 기준이 서로 다르기 때문이다.

따라서 본 논문에서는 검색어 입력이 동일한 키워드임에도 불구하고 서로 다른 검색 순위를 보여주는 검색엔진들 대신 사용자가 입력한 키워드와 문서의 연관성을 추출하여 사용자가 찾고자하는 키워드를 입력했을 때 키워드와 문서의 연관성이 가장높은 자료를 출력하고자 한다. 제안모델로 LSTM 모델을 이용하여 문서내의 어휘적 중의성을 해소해 연관성의 정확도를 향상시켜줌으로 기존의 검색엔진의 문제점들을 해결해보자 한다.

II. 관련연구

현재 자연어 처리 연구를 위해 딥러닝 처리 기술로 LSTM, RNN, Seq2Seq, BERT등 다양한 모델들이 사용되고 있다.

2.1 LSTM

LSTM은 주로 One-to-one 모델 또는 Many-to-one(n:1) 모델로 구분된다. One-to-one(1:1) 모델은 문서내 모든 단어에 대해 모델링 할 수 있는 특징이 있으며 Many-to-one 모델은 단어들 간의 문맥적 의미 관계를 보다 효과적으로 반영할 수 있다는 장점을 갖는다[1]. 그러나 두 모델은 전체 문서 시퀀스의 마지막 출력을 문서 벡터로 간주하는 특징이 있어 입력이 길어짐에 따라 초기에 입력된 패턴의 인식률이 급격히 저하되어 장문의 인코더로는 적합하지 않다는 문제점을 가지고 있다[2].

2.2 양방향 LSTM

양방향 LSTM은 기존 LSTM에 역방향으로 학습하는 레이어를 추가하였으며, 가변적인 시퀀스의 길이에 제약이 없고, 초기에 입력된 패턴의 인식률이 저하되는 LSTM의 문제점을 보완하여 제안 되었다. 양방향 LSTM에서는 좌(L)에서 우(R)로 다시 우(R)에서 좌(L)로 데이터를 처리하는 방법을 사용해 언어와 문서간의 유사도를 개선하는데 기여함으로 분류 성능을 개선하고 향상 시켰다[3]. 다음 <그림 1>은 양방향 LSTM 기본 모델로 X1~X3을 입력으로 Y1~Y3까지를 출력으로 나타낸 것이다.

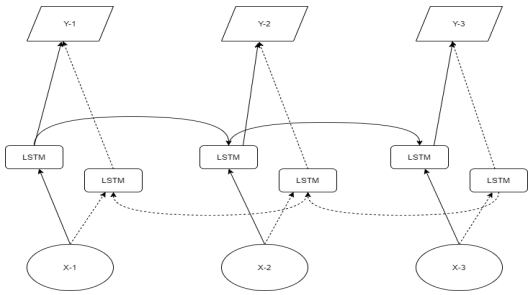

그림 1 양방향 LSTM 기본 모델

2.3 중의적 표현

중의적 표현방법은 문서내의 키워드와 사용자가 검색한 키워드의 연관성을 향상시키기 위해 동음이의어를 제거하는 방식을 사용하였다. 이런 중의적 표현 방식에서 중요한 동음이의어 제거를 위해 GloVe임베딩 층과 앙상블(양방향LSTM, LSTM, RNN)모델을 각각 조합하여 동음이의어를 제거하는 모델 연구가 진행 되었다[3]. 표 1은 중의적 표현방법에서 모델의 성능을 배치 크기에 따라 실행한 결과를 정리한 것이다.

표 1. 분류 비교 결과표

모델명	배치 크기 64			배치 크기 128		
	학습	검증	테스트	학습	검증	테스트
Glo-BLSTM	0.6426	0.6446	0.5917	0.6282	0.6321	0.5961
Glo-LSTM	0.6667	0.6263	0.5776	0.6423	0.6346	0.5907
Glo-RNN	0.3387	0.3997	0.3758	0.3472	0.3905	0.3752

동음이의어 제거를 위해 GloVe임베딩 층과 앙상블한 모델중 양방향LSTM 모델의 성능이 가장 우수한 것을 알 수 있다.

III. 키워드와 문서의 연관성 향상 모델

본 논문에서는 특정 키워드를 기반으로 검색하는 경우 방대한 양의 DB에서 어휘적 중의성이 해소된 결과를 제시 할 수 있는 딥러닝 기반의 자연어 처리모델을 제안하여 자료를 재 구분하는 과정을 최소화 함으로 보다 정확한 검색 결과를 도출하고자 한다. 그림 2는 본 논문에서 제시한 모델을 메타태그 방식의 검색엔진에 적용 시켰을 때의 흐름을 나타낸다.

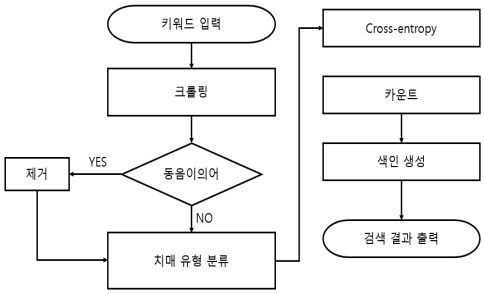

그림 2. 모델을 적용한 흐름도

위 그림 2에서와 같이 키워드에 따라 검색한 문서를 크롤링한 뒤 자연어 처리를 통해 동음이의어를 구분하며, 지정해둔 레이블에 따라 우선순위를 분류한 후 키워드를 카운트하여 유사도를 비교한다. 이는 사용자의 검색 의도를 보다 정확히 반영함으로 검색결과의 정확도를 향상시는 목적을 달성할수 있는 모델을 나타낸 것이다.

본 논문에서는 의학분야에서 치매추적물질을 위한 플랫폼에서 치매 기전을 발견하기 위한 초기 대응방법으로 발병률이 높은 치매 유형별로 레이블을 구성하고 검색하고자 하는 키워드의 유사도를 높이고자 자이온프로세스 기전 플랫폼에 사용될 목적으로 제안 되었다. 제안 모델은 치매와 관련된 추적물질을 검색하기위한 모델로 유사도를 높이기 위해 학습레이블을 치매 발병률을 기준으로 분류하였다. 학습레이블은 치매 발병률이 가장 높은 알츠하이머치매, 파킨슨병 치매, 혈관성 치매, 초로기치매, 노인성 치매 순으로 구분하고 제안 모델에서 학습하고 각각의 레이블에 크로스엔트로피 알고리즘기반의 유사도를 향상시킴으로 정확도를 향상시켰다.

3.3 제안 모델의 구성도

다음은 제안 모델의 구성도를 그림 3과 같다. 그림 3은 검색 결과 크롤링된 내용에서 크롭된 문서에 자연어 처리를 적용 후 학습된 레이블과의 연관도를 비교하여 치매 유형을 분류하는 모델의 구성도를 나타낸 것이다. 모델 학습에 사용된 치매 분류 레이블을 이용해 학습된 자연어와 검색결과로 나타난 문서의 중복성을 제외시킴으로 사전 동음이의어에 대한 중의성을 해결하고 검색의 연관성을 높이고자 하였다. 겹치지 않은 문서를 제외시킴으로 동음이의어에 대한 중의성을 해결함으로 연관성을 높이고자 하였다. 제안 모델은 중의성 해결에 가장 우수한 성능을 보인 양방향 LSTM을 기반으로 구성되었으며 치매발명 기전을 검색 플랫폼을 위해 사용된다.

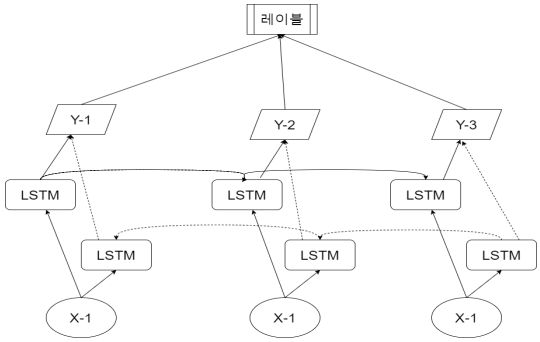

그림 3. 치매 분류 모델 구조

Ⅳ. 결 론

본 논문에서는 치매발병 기전 검색플랫폼에서 검색 키워드와 문서의 연관성 향상을 위해 동음이의어를 제거하는 방식을 제시하였다. 관련 논문을 기반으로 자연어 처리로 자주 사용되고 있는 LSTM 모델 대신 한글의 특성상 동음이의어에 대한 검색 정확도 문제를 해결하기 위해 양방향 LSTM을 기반으로 모델을 설계하였다. 본 논문은 치매와 관련된 동음이의어를 구분함으로 검색 키워드와 문서의 연관성을 최대한 높인 후 카운트를 하여 검색어와 관련도가 높은 순으로 결과를 출력함으로 검색의 정확도를 향상시켰다. 제안 모델은 향후 치매발병기전 검색 플랫폼을 통해 연구자들을 위한 모델로 발전할 수 있도록 지속적인 연구가 필요하다.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2018R1C1B5083789).

References

[1] D. H. Kim, J. H. Lee, “LSTM based Language Model for Topic-focused Sentence Generation”, *The Korean Society Of Computer And Information*, Vol. 24, No. 2, pp. 1-4, July. 2016.

[2] S. J. Kwon, J. A. Kim, S. W. Kang, J. Y. Seo “The Sentiment Classification of documents using LSTM Attention Encoder” *The Korean Institute of Information Scientists and Engineers,* pp. 762-764, June. 2016.

[3] H. Y. Ki, K. S. Shin, “Emotion Analysis Using a Bidirectional LSTM for Word Sense Disambiguation”, *The Korea Journal of BigData*, Vol. 5, No. 1, pp. 197-208, 2020.