재난안전 사회관심 분석을 위한 언어모델 활용 정보 네트워크 구축

A Language Model based Knowledge Network for Analyzing Disaster Safety related Social Interest

최동진** · 한소희* · 김경준*** · 배은솔****

Choi, Dong-Jin · Han, So-Hee · Kim, Kyung-Jun · Bae, Eun-Sol

요 약

본 논문은 대규모 텍스트 데이터에서 이슈를 발굴할 때 사용되는 기존의 정보 네트워크 또는 지식 그래프 구축 방법의 한계점을 지적하고, 문장 단위로 정보 네트워크를 구축하는 새로운 방법에 대해서 제안한다. 먼저 문장을 구성하는 단어와 캐릭터수의 분포를 측정하며 의성어와 같은 노이즈를 제거하기 위한 역치값을 설정하였다. 다음으로 BERT 기반 언어모델을 이용하여 모든 문장을 벡터화하고, 코사인 유사도를 이용하여 두 문장벡터에 대한 유사성을 측정하였다. 오분류된 유사도 결과를 최소화하기 위하여 명사형 단어의 의미적 연관성을 비교하는 알고리즘을 개발하였다. 제안된 유사문장 비교 알고리즘의 결과를 검토해 보면, 두 문장은 서술되는 형태가 다르지만 동일한 주제와 내용을 다루고 있는 것을 확인할 수 있었다. 본 논문에서 제안하는 방법은 단어 단위 지식 그래프 해석의 어려움을 극복할 수 있는 새로운 방법이다. 향후 이슈 및 트랜드 분석과 같은 미래연구 분야에 적용하면, 데이터 기반으로 특정 주제에 대한 사회적 관심을 수렴하고, 수요를 반영한 정책적 제언을 도출하는데 기여할 수 있을 것이다.

Keywords : 정보 네트워크, 지식 그래프, 유사문장 비교, 언어모델, 재난안전 사회관심 분석

1. 서 론

코로나19와 같은 신종 감염병 재난이 발생하거나 국지성 호우로 특정 지역이 침수되는 상황이 발생하게 되면, 사람들은 온라인이라는 공간을 이용하여 다양한 의견을 제시하고 논쟁을 펼쳐나간다. 이러한 의견들은 정부의 대책을 비판하거나 신속한 대응과 재발 방지방안을 주문하는 내용이 대부분이다. 재난관리업무를 담당하는 정책입안자와 이행자의 입장에서 보면, 많은 사람들이 제안하는 의견을 경청하는 것은 정책추진에 필요한 동력을 이끌어내는 시작점이라고도 말할 수 있다. 이러한 이유로 중앙정부와 지방자치단체는 하나의 정책을 펼쳐나감에 있어서, 전문가 자문과 더불어 다양한 대중의 의견을 수렴하려고 노력한다. 설문과 같은 여론조사를 수행하거나 빅데이터 분석 기술을 이용하여, 시급성이 높은 주제를 발굴하고 대응해나가고 있다. 이때, 빅데이터 분석에는 하나의 주제에 대한 주요 키워드를 순위화하고, 키워드의 빈도 변화와 함께 연관어, 감성분석 등이 사용되고 있다[Collobert, 2011]. 또한 하나의 문장이나 문서에 함께 출현하는 단어들을 서로 연관되어 있다고 가정하고, 단어를 노드로 표현하며 동시출현 단어를 간선으로 연결하는 네트워크 분석기법도 적용되고 있다[장요한, 2020]. 하지만 이러한 단어 단위의 텍스트 분석에는 해석이 어렵다는 한계가 있다. 이를 극복하기 위해서 문장의 유사도를 측정하는 연구가 활발히 진행되고 있는데, 문장은 수많은 단어의 조합으로 다양하게 표현될 수 있어 의미적 처리가 까다로운 부분이 있다. 최근 인공지능 알고리즘의 비약적인 발전으로 언어모델(language model)을 이용하여 문장의 형태와 의미적 정보를 수치화할 수 있는 기술성능이 꾸준히 향상되고 있으며 이러한 언어모델은 유사문장 비교(sentence classification) 분야에서도 활발히 적용되고 있다[Kim, 2014].

문장 단위로 구성된 정보 네트워크를 구축함에 있어서 유사 문장을 판단하는 과정은 정보 네트워크 구축에 가장 중요한 부분이다. 이에, 본 연구에서는 재난안전 분야에 대해서 사람들의 공통된 관심 주제를 분석함에 있어서 의미적으로 서로 유사한 문장을 비교하는 방법에 대해서 제안하고자 한다.

* 정회원 · 국립재난안전연구원 사회재난연구센터 연구원 hsh3562@korea.kr

** 국립재난안전연구원 사회재난연구센터 공업연구사 djchoi@korea.kr

*** 국립재난안전연구원 사회재난연구센터 시설연구관 kjkim96@korea.kr

**** 국립재난안전연구원 사회재난연구센터 연구원 esb613@korea.kr

2. 재난안전 사회관심 분석을 위한 정보 언어모델 활용 네트워크 구축 방법

문장이란 표준국어대사전에 따르면 생각이나 감정을 말과 글로 표현할 때 완결된 내용을 나타내는 최소의 단위로 정의되어 있는데, 이렇듯 어떠한 현상에 대하여 사람들이 느끼는 생각과 의견은 문장단위로 표현되고 전달된다. 문장은 주어와 서술어를 갖추고 있는 것이 원칙이지만, 때로는 '진짜?'와 같이 매우 짧게 표현되기도 한다. 이렇게 한 개 또는 두 개의 단어로 구성된 문장의 경우, 특정한 내용을 내포하고 있다고 보기 어렵다. 따라서 너무 짧은 문장의 경우, 노이즈로 간주하고 유사문장 비교집단에서 제외시킬 필요가 있다. 이에 본 논문에서는 문장을 구성하는 단어의 수와 음절 수의 수가 평균값 이상인 경우에만 유사문장 비교 대상으로 선정한다. 이렇게 전처리가 완료된 비교집단을 대상으로 두 단계에 걸쳐 유사문장을 비교하는 방법론을 그림 1과 같이 제안한다. 먼저 BERT 기반 사전 학습된 언어모델을 이용하여 문장을 벡터화하고, 벡터화된 두 문장간의 코사인 유사도를 측정한다. 이때 하나의 언어모델만을 이용하는 것이 아닌 서로 다른 두 개의 언어모델을 이용하여 문장을 벡터화하고 코사인 유사도 값을 계산한다. 두 개의 언어모델로 변환된 문장 벡터값에 대한 코사인 유사도를 측정하고 그 성능을 비교하기 위해서 다음 카페에서 코로나19로 검색한 결과 중 제목과 댓글 문장을 대상으로 유사문장 비교 실험을 진행하였다. "우한 폐렴 신종 코로나바이러스 예방법"이라는 문장에 대해서 언어모델을 이용하여 다른 문장들과 코사인 유사도를 측정한 결과를 살펴보면, 표 1과 같은 결과를 확인할 수 있었다. 비교대상 문장과 유사한 내용을 기술하는 문장들이 높은 코사인 유사도 값으로 측정되었지만, '신종 코로나바이러스 관련 공지'라는 문장처럼 비교문장과 연관성이 떨어지지만 높은 코사인 유사도가 측정되는 문제점을 확인할 수 있었다.

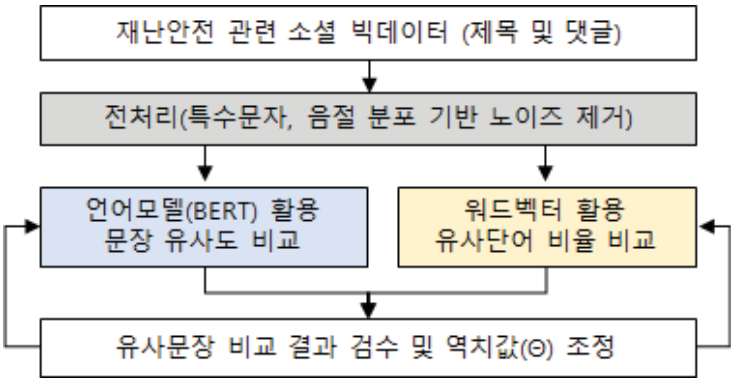

그림 1. 언어모델 활용 유사문장 비교 방법의 개념도

표 1. 제목 문장('우한 폐렴 신종 코로나바이러스 예방법')에 대한 유사문장 비교 결과 예

비교문장	유사도	
	K	E
우한 폐렴 신종 코로나바이러스 우리 모두 조심합니다	0.85	0.87
중국 우한 신종코로나 폐렴 바이러스에 대한 면역력 강화 식품	0.88	0.90
신종 코로나바이러스 실시간 상황	0.84	0.90
신종 코로나바이러스 관련 공지	0.84	0.89
태국에서 신종 코로나바이러스 감염 우한 폐렴 치료법 발견	0.87	0.90

이에 본 연구에서는 오분류된 유사도 결과를 최소화하기 위하여 언어모델 기반 유사문장 비교결과에 추가적으로 명사형 단어의 의미적 연관성을 비교하는 방법을 제안한다. 예를 들어 그림2와 같이 '마늘이 코로나19를 예방한다'라는 문장이 있을 때, 마늘이라는 단어 위치에 다른 비슷한 단어가 사용되면 코로나19를 예방한다는 문장의 맥락을 유지하면서 동일한 주제에 대해서 새로운 정보가 표현되는 형태를 찾아 나가는 방법이다. 마늘과 유사한 단어를 찾아내기 위해서 본 논문은 skipgram 알고리즘으로 다음 카페 게시글의 제목과 댓글에 대하여 각각의 워드벡터를 자체 구축하였다. 그 결과, 명사 '마늘'과 가장 유사한 단어는 표 2에서 보여주는 것과 같이 '생강, 고추, 다시마, 장아찌, 양파, 된장'과 같은 단어로 학습되었다. 이렇게 구축된 워드벡터를 이용하여 아래 수식1과 같이 한 문장을 구성하는 n개의 단어()에 대해서, 워드벡터 기반 유사단어 k개에 대해서 비교하며 주어진 두 문장을 구성하는 단어가 서로 관련성 여부를 비율로 산출한다. 본 논문에서 제안하는 방법으로 2020년 2월 코로나19로 검색된 다음 카페 게시글 총 8,096개 제목 문장에 대해서, 비교 가능한 모든 경우의 수(nC2) 32,768,560개에 대해서 언어모델

기반 유사문장을 측정한 결과, 263,987개의 유사문장이 발견되었다. 다음으로 워드벡터를 이용하여 오분류된 결과를 삭제한 결과, 총 173,869개의 유사문장이 추출되는 것을 확인할 수 있었다.

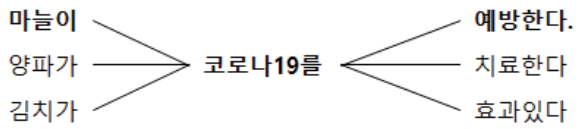

그림 2. 단어를 교체하며 표현되는 문장의 사례

$$\text{유사단어 비율} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} f(w_i, w_i \mid w_j)}{n} \quad (1)$$

표 2. 학습된 단어벡터를 이용한 유사단어 비교 결과

비교 단어	유사 단어 결과
마늘	생강, 고추, 다시마, 장아찌, 양파, 된장 등
코로나	감염증, 신종, COVID, 바이러스, 팬데믹 등
예방	접종, 방지, 백신, 수칙, 질병, 감염 등
치료	입원, 환자, 투석, 투여, 중환자실, 처방 등

3. 결 론

본 논문에서는 코로나19나 풍수해와 같은 재난상황에서 대중의 공통된 관심사를 분석하는 기술을 연구함에 있어서 문장 단위로 정보 네트워크를 구축하는 방법에 대한 내용을 다루고 있다. 정보 네트워크 구축에 있어서 가장 중요한 부분은 노드가 되는 문장과 서로 유사한 문장을 연결하는 작업이다. 즉, 유사 문장을 판단하는 알고리즘의 성능이 정보 네트워크 분석의 핵심이라고도 말할 수 있다. 최신의 인공지능 언어모델 알고리즘 중 하나인 BERT를 이용하여 문장을 벡터화하고 비교 대상이되는 문장들의 코사인 유사도값을 측정하고 비교한 결과, 두 문장이 서로 다른 내용을 기술하고 있음에도 불구하고, 유사도가 높게 측정되는 알고리즘의 한계가 발견되었다. 이에 본 논문에서는 공통된 주제를 기술하는 문장을 군집화하는 과정에서 오분류된 결과를 정제할 수 있는 방법을 제안하였다. 문장을 기술하는 명사형 단어들에 대해서 워드벡터를 이용하여 산출된 가장 유사한 단어들이 비교 대상이되는 문장에 사용되었는지 여부를 판단하며, 언어모델 성능을 최대한 활용함과 동시에 한계를 극복하는 방법이다. 제안하는 방법론의 타당성을 검토하기 위하여, 다음 카페 게시글 제목 문장에 대해서 본 논문에서 제안하는 방법론으로 유사문장을 비교한 실험을 수행하였다. 그 결과, “코로나바이러스 격리자에 긴급생활자금지원 검토 공무원은 유급휴가”라는 문장에 대해서 언어모델로 유사한 문장으로 판단된 문장 중, “접경지역에 코로나바이러스 자원보사자 500명 긴급투입”과 같이 오분류되는 문장들을 삭제할 수 있었다. 하지만 적정 유사단어의 비율에 따라 삭제되는 문장이 달라지기 때문에 최적의 역치값을 설정해야하는 과제가 남아있는 실정이다. 향후, 역치값에 따른 삭제문장의 분포와 삭제되는 문장의 내용을 분석하여, 본 논문에서 제안하는 방법론을 보완해나갈 계획이다.

감사의 글

본 연구는 국립재난안전연구원 주요과제(인공지능 기술 활용 재난안전 분야 인포데믹 피해 예방 연구, 2022-04-03)의 연구내용을 포함하고 있습니다.

참고문헌

장요한 (2020) 빅데이터를 이용한 국토 민생현안 모니터링 연구, 국토연구원, pp. 1-42.

Collobert, R. (2011)Natural Language Processing(Almost) from Scratch, Journal of Machine Learning Research, pp. 2493-2537.

Kim, Y. (2014) Convloutional Neural Networks for Sentence Classification, Proceddings of the EMNLP, pp. 1746-1751.