JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 4, pp. 765-771, Apr. 2022

Check for updates

반려동물 질병 질의응답 시스템을 위한 개체명 인식

오 한 석[1] · 이 현 아[2*]
[1]금오공과대학교 컴퓨터소프트웨어공학과 학사과정
[2*]금오공과대학교 컴퓨터소프트웨어공학과 교수

Named Entity Recognition for Pet Disease Q&A System

Han-Seok Oh[1] · Hyunah Lee[2*]

[1]Bachelor's Course, Dept. of Computer Software Engineering, Kumoh National Institute of Technology, 39177, Korea
[2*]Professor, Department of Computer Software Engineering, Kumoh National Institute of Technology, 39177, Korea

[요 약]

반려동물에 대해 높아진 관심으로 반려동물 인구가 크게 증가하고 있는 것에 비해, 반려동물이 아플 때 질병 관련 정보를 찾는 것은 여전히 쉽지 않다. 본 논문에서는 사용자가 입력하는 반려동물 질병 증상에 관련된 질병명을 답변으로 출력하는 질의응답 시스템을 제안한다. 제안 시스템에서는 질병명의 개체명 인식(Named Entity Recognition, NER)을 위해 BERT에 CRF층을 추가한 BERT-DIS-NER 모델을 만들고, 질병명의 특징을 반영할 수 있는 음절 단위 개체명 인식을 사용한다. 반려동물 질병 데이터 부족 문제를 해결하기 위해, 유사 문맥을 가질 것으로 예상되는 사람 질병 데이터를 이용하여 기본 모델을 학습하고, 반려동물 질병에 대한 데이터를 파인튜닝에 사용한다. 실험 결과에서 BERT-DIS-NER의 F1-score는 사람 질병 데이터만을 학습했을 경우 0.74, 반려동물 질병 데이터만을 학습했을 때 0.77, 사람 질병 데이터에 반려동물 질병 데이터를 파인튜닝할 경우 0.81의 결과를 보여, 제안 방식에 의한 성능 향상을 확인할 수 있었다.

[Abstract]

While the number of companion animals is rapidly increasing due to growing interest in pets, it is still not easy to find disease-related information when a companion animal is sick. In this paper, we propose a pet disease Q&A system that answers disease names related to the companion animal symptoms input by users. In our system, we create a BERT-DIS-NER model that adds a CRF layer to BERT for the disease named entity recognition and use syllable unit-based named entity recognition that can reflect the characteristics of disease names. In order to solve the problem of lack of animal disease data, a base model was trained using human disease data expected to have a similar context to pet disease, and data on the animal disease were used for fine-tuning. In experiments the F1-score of BERT-DIS-NER showed 0.74 trained with only human data, 0.77 trained with only animals, and 0.81 trained with human data and fine-tuned with animal data, so it was confirmed that the proposed method improves performance.

색인어 : 개체명 인식, 동물 질병 개체명, 인간 질병 개체명, 자연어처리, 기계학습

Keyword : Named Entity Recognition, Animal Disease Name, Human Disease Name, Natural Language Processing, Machine Learning

http://dx.doi.org/10.9728/dcs.2022.23.4.765

Received 24 March 2022; **Revised** 13 April 2022
Accepted 13 April 2022

***Corresponding Author; Hyunah Lee**

Tel: +82-54-478-7546
E-mail: halee@kumoh.ac.kr

Ⅰ. 서 론

반려동물에 대해 높아진 관심으로 반려동물 인구가 크게 증가하고 있는 것에 비해, 반려동물이 아플 때 질병 관련 정보를 찾는 것은 여전히 쉽지 않다. 반려동물 질병에 관한 기존 서비스는 수의사와의 상담 채팅, 사용자 Q&A 게시판, 질병명 검색, 간단한 예진 시스템 등의 방식이 있다. 수의사 채팅은 수의사의 상황에 따라 답변이 늦어질 수 있으며, 사용자 Q&A 게시판은 비전문가의 답변이므로 신뢰성이 낮다. 질병명 검색의 경우 정확한 질병명을 입력해야 하며, 예진 시스템의 경우 기정의된 증상목록만을 대상으로 동작하여 사용 편의성이 낮다.

본 논문에서는 사용자가 자연어로 자유롭게 입력하는 반려동물 질병 증상에 해당하는 질병명을 답변하는 반려동물 질병 질의응답 시스템과 이를 위한 개체명 인식 모델을 제안한다. 시스템은 질병 증상에 대한 입력 문장과 비슷한 문서를 반려동물 질병 질의응답 문서 집합에서 찾고, 개체명 인식모델을 이용하여 해당 문서 내의 질병명을 추출하여 결과로 출력한다. 질의응답 문서 집합은 수의사 답변[1]을 사용하여 신뢰도를 높인다. 사용자는 출력된 질병명과 이에 관련된 수의사 답변을 참고하여 반려동물의 상태에 따른 적절한 처치나 동물병원 방문 여부를 결정할 수 있다.

본 시스템에서는 반려동물 질병명을 자동으로 찾기 위해 개체명 인식을 사용한다. 구글에서 발표한 언어모델인 BERT[2] 모델에 CRF(Conditional Random Fields) 층을 추가하여 BIO(Begin, Inside, Outside)로 태깅된 질병명을 포함한 문장을 학습한다. 사람 질병 데이터를 이용하여 기본 모델을 학습한 후 동물 질병 데이터를 이용하여 파인튜닝을 진행한다. 질병명의 경우 대부분 고유명사 및 복합명사로 이루어져 형태소 분석기의 결과가 좋지 못하기 때문에 음절 기반의 학습을 진행한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 시스템의 기술인 개체명 인식에 관한 선행연구를 살펴보고, 3장에서는 실험에 관하여 설명하고, 4장에서는 반려동물 질의응답 시스템을 소개하고, 5장에서는 결론을 맺는다.

Ⅱ. 관련 연구

개체명 인식에 관한 연구는 영어권에서 최초로 언급되었으며[3-4], 이와 비슷한 접근 방식을 이용하여 한국어를 대상으로 하는 연구도 활발하게 진행되었다[5-11].

[5]에서는 한국어에 대한 규칙 기반 개체명 인식을 제안하였다. 규칙 기반 개체명 인식에서는 각 사전과 규칙을 사람이 직접 만들어야 하므로, 다른 분야로의 이식성이 낮고 각 분야에 해당하는 규칙을 매번 새로 작성해야 하는 문제점이 있다.

이후 제안된 통계 기반 개체명 인식 연구에서는 수작업으로 만들어진 자질을 기반으로 모델을 학습하며, 대표적으로 SVMs(Support Vector Machines), CRF(Conditional Random Fields), HMM(Hidden Markov Model), ME(Maximum Entropy) 등을 사용한다[6-9]. 통계 기반 개체명 인식의 경우 문장의 자질 추출을 위한 함수 등을 직접 정의해야 하여, 규칙 기반과 마찬가지로 새로운 분야에로의 확장성이 낮다.

이후 기술의 발전으로 딥러닝에 기반한 다양한 신경망 모델 및 언어모델이 등장했다, 이러한 모델들과 통계 기반 모델을 활용하여 기존 방식들의 문제점을 해결하는 연구가 진행되었다. [10]에서는 양방향 LSTM과 글자 단위의 특성을 고려하기 위해 CNN을 활용한 후 CRF를 사용한 bi-directional LSTM-CNN -CRF 기반의 모델에 미리 구축된 hand-craft 자질과 품사 태깅 정보 및 기구축 사전(lexicon) 정보를 추가로 활용하여 자질을 보강하여 성능 향상을 얻었다. BERT 기반 연구[11]에서는 BERT 모델에 CRF 층을 추가한 후 개체명 태깅에서 음절 단위의 토크나이징을 활용한다. [12]에서는 순환신경망 대신 합성곱 신경망을 이용한 음절 기반 한국어 개체명 인식 방법이 진행하여, 문장 외의 다른 자원을 사용하지 않으면서 형태소 분석 기반 방식이 사전에 등록되지 않은 고유명사를 오분석하는 문제를 해결한다.

국내의 질병 및 생물학 분야의 개체명 인식에 관한 연구[13-15]는 대부분 영어권의 대량의 질병 데이터를 이용하였다. [16]에서는 한국어 의료분야 개체명의 특성을 반영하기 위해 정규화 기반 개체명 인식 방식을 제안하였으나 수동으로 섬세하게 레이블링된 학습 데이터에 기반한다. [17]에서는 반자동으로 구축한 140개의 동물 질병명 사전 기반으로 자동 태깅한 데이터를 사용한 반려동물 질병 개체명 인식 연구를 진행하였으나, 학습 데이터 부족으로 인식할 수 있는 질병명의 개수가 적어 시스템의 확장성이 낮다.

본 논문에서는 기존 연구의 데이터 획득의 어려움을 해소하는 방법으로 공공데이터포털의 동물 질병명 목록으로 학습 데이터를 자동으로 태깅하고, 사람과 반려동물 질병의 문맥과 질병명의 유사성에 착안하여 사람의 질병 데이터로 기본 모델을 학습하는 방식을 제안한다.

Ⅲ. 반려동물 질병 개체명 인식

제안하는 반려동물 질병 질의응답 시스템은 사용자가 입력한 문장과 유사한 질문을 반려동물 질병 질의응답 문서 집합에서 찾은 후 해당 문서의 수의사 답변에서 질병명을 추출하여 화면에 출력한다. 수의사 답변에서 질병명을 추출하기 위해서는 질병 개체명 인식 기술을 적용한다. 본 장에서는 제안하는 반려동물 질병 개체명 인식모델인 BERT-DIS-NER 모델의 학습 방법과 결과에 관하여 기술한다.

3-1 제안 방법

본 논문에서 적용한 모델은 선행 연구된 BERT-NER[11] 모델과 유사하게 BERT 모델에 CRF층을 추가하고, 입력 문장의 음절에 대한 임베딩과 태그를 CRF의 입력으로 넘긴다. BERT에서는 Tokenizer로 문장을 토큰화하고 이를 BERT의 입력값으로 넘기기 때문에 음절 단위로 토큰화할 수 있도록 Tokenizer를 학습해야 한다. 이를 위해 KoChar-ELECTRA[18]의 Vocab 파일을 이용하여 Sentencepiece Tokenizer를 학습시킨 후, 이를 이용하여 음절 단위 토크나이징을 수행한다.

수의사 답변을 포함한 동물 질병의 응답 문서는 개체명 인식을 위한 좋은 지식원이지만, 그 양이 많지 않아 데이터 부족 문제는 여전히 존재한다. 이를 해결하기 위하여 본 논문에서는 반려동물 질병 문서와 유사한 문맥과 질병명을 가질 것으로 예상되는 사람 질병 문서 데이터를 이용한다. 반려동물 질병 문서 데이터보다 큰 규모인 사람 질병 문서 데이터에서는 다양한 질병명에 대한 문맥을 충분히 학습할 수 있을 것으로 기대할 수 있다. 사람 질병 문서를 이용하여 개체명 인식을 위한 기본 모델(base model)을 학습시킨 후 반려동물 질병 문장 데이터를 이용하여 파인 튜닝(fine tuning)을 수행한다. 사람 질병명 중 하나인 '감기 바이러스'를 포함한 문장을 학습시키면 반려동물 질병명 중에서 이와 유사한 문맥을 가지는 '허피스 바이러스', '파보 바이러스' 등 학습하지 않은 질병명이 들어오더라도 질병으로 인식할 수 있다.

제안 시스템에서는 음절 기반 임베딩을 사용한다. 질병명은 대부분 고유명사나 복합명사로, 기존의 형태소 분석기를 사용하면 질병명을 제대로 분석하지 못하는 문제점이 있다[12]. 사람 질병명 중 하나인 '폐섬유화증'을 예로 들면 해당 질병명의 '증'은 질병명을 인식하는데 좋은 단서가 된다. 하지만 이를 형태소 분석기로 분석하면 "폐+ 섬유+ 화증"으로 잘못된 분석을 하거나 단어 전체를 미등록어로 인식한다. '만성피로 증후군'은 '만성 피로 증후군'이나 '만성피로증후군'으로 다르게 표기될 수 있는데, 각 단어를 형태소 분석기를 사용할 경우, "만성+ 피로+ 증후군", "만+ 성+ 피로+ 증후군"으로 다른 결과가 나온다. 앞의 예시와 마찬가지로 '만성' 또한 질병명을 인식하는데 좋은 단서지만 표기법에 따라 다른 분석 결과가 나오기 때문에, 모델학습을 위한 전처리 과정에서 사용하기에 적합하지 않다. 본 논문의 학습 및 검증에 사용되는 모든 데이터는 공백을 제거한 뒤 음절 단위로 태깅하여 모든 질병명을 일관되게 학습한다.

3-2 실험 데이터

BERT-DIS-NER 모델을 학습하기 위한 데이터로 네이버 지식iN 수의사 답변[1]과 전문의 답변[19]을 사용한다. 그림 1은 질의응답 문서의 예를 보인다. 사용자의 질문에 대한 수의사 답변에서 '무릎골탈구', '십자인대손상', '고관절질환'과 같은 증상에 관련된 질병명이 포함된다. 수집한 데이터에는 질병명 뒤에 한자가 같이 붙는 경우가 존재한다. Vocab에 한자와 특수문자는 포함되어 있지 않기 때문에 모두 제거한다.

Q: 강아지가 다리를 절뚝거려요
강아지가 갑자기 다리를 절뚝거리네요. 몇년간 아토피때문에 다니던 병원에서 의사선생님이 다리를 이리저리 만져보더니 괜찮다고 그냥 가서 몇칠두고보라고했지만 걱정스러워서 엑스레이찍었는데 이상이 없다고 하네요 이삼일뒤에 안되겠어서 병원에서 소염제가섞여있는 약이라고해서 6일먹였는데도 절뚝거려요 (이하 생략)
A: 이*철 수의사님 답변
안녕하세요. 지식iN 동물의료상담 수의사 은*동물병원장 버드*무 이*철입니다. 상황만으로 판단하기에는 제한적이고 어려운 부분이 있습니다. 일반적으로 후지파행은 슬개골탈구, 십자인대손상, 고관절질환 등의 원인으로 발생할 수 있습니다. 또한 발바닥 상처나 가시와 같은 뾰족한 이물에 의해 발생하기도 합니다. 무슬개골탈구의 경우 정도에 따라 1기~4기로 구분합니다. 1기와 2기는 겉으로 확인하기에는 어렵고 확연히 구분되지는 않습니다. 슬개골탈구는 진행함에 따라 파행증상이 심해지고 자세 이상과 퇴행성 관절염을 동반합니다. 증상이 지속된다면 다시 한번 병원에서 정확한 진단과 적절한 처치 받아보시길 권해드립니다.

* It is an example of Korean sentences, so it is written in Korean

그림 1. 반려동물 질병 질의응답 데이터 예시

Fig. 1. Example of pet disease Q&A data

개체명 인식 모델학습을 위한 데이터 집합을 획득하기 위해 사람과 반려동물에 대한 질병명 사전을 구축한 후 태깅하여 학습을 진행한다. 사람 질병명 사전은 질병관리청[20]과 국내 대형 병원 홈페이지[21][22]의 질병 목록을 사용하여 구축하였다. 한 가지 질병에 대해 여러 질병명 표기가 존재하면 모든 동의어를 사전에 포함하였다. 예를 들어 질병명 중 하나인 "쯔쯔가무시"의 경우, '쯔쯔가무시병', '쯔쯔가무시증', '쓰쓰가무시'.'쓰쓰가무시병', '쓰쓰가무시증'으로 표기될 수 있어, 모든 동의어를 질병명 사전에 추가한다. 최종적으로 획득한 사람 질병명 사전은 3,704개의 질병명을 포함한다. 사람 질병명 개체명 인식을 위한 학습데이터는 네이버 지식iN 의사 답변[19]에서 총 261,936문장을 수집하였으며, 질병명 사전으로 태깅했을 때 1,310개의 질병명이 총 423,664 번 나타났다.

반려동물 질병 사전은 공공데이터포탈의 '한국과학기술정보연구원_동물질병데이터'[23]를 이용하여 구축하였다. 이는 1,538개 행의 반려동물과 관련된 질병 데이터를 포함한다. 데이터의 형태가 문장 형태거나, 여러 개의 질병명을 묶어놓은 형태이기 때문에 수작업을 통해 질병명을 분리한 후 중복되는 질병명을 제거하여, 최종적으로 총 1,474개의 반려동물 질병명을 가지는 사전을 구축하였다. 구축한 사전에는 사람과 동물에게 공통으로 발병할 수 있는 질병명이 빠져있는 경우가 있어, 반려동물 질병 문서 태깅에서는 사람 질병명 사전과 반려동물 질병명 사전을 모두 이용하여 태깅하였다. 지식iN의 수의사 답변 문장 중 '빠른 시일 내로 병원에 방문하시기 바랍니다' 같은 질병명을 포함하지 않는 의미 없는 문장은 제거하여, 총 19,160문장에 대한 질병명 태깅을 수행하였으며, 총 650개의 질병이 태깅되었다. 태깅한 문장을 학습과 검증 집합으로 나누어 각각 15,337문장, 3,823문장으로 분리하였다. 학습 집합에는 질병명 482개가 22,249번, 검증 집합에는 질병명 404개가 10,086번 나타났다.

표 1. 사람 및 반려동물 질병 데이터
Table 1. Data of human and companion animal disease

	sentence		disease	
	train	test	train	test
human	261,936	-	1,310	-
pet	15,337	3,823	435	404

표 1은 구축된 학습 데이터와 검증 데이터의 총 문장 수와 태깅된 질병명 개수를 보인다. 구축된 사람질병사전과 동물질병사전의 질병명에 비해 실제 문서에서 나타난 질병명은 절반 이하의 수준으로 나타났다. 이는 인터넷에서 의사와 수의사가 답변에 사용하는 질병명이 제한적이기 때문으로 보인다. 실험 집합 구성에서는 BERT-DIS-NER이 학습하지 않은 질병명을 얼마나 인식하는지 평가하기 위해서 검증 셋의 질병명 404개 중 161개의 질병명은 학습되지 않은 질병명을 포함하게 구성한다.

각 문장 내에 질병이 질병 사전에 포함되는 경우 해당 질병명의 시작점을 B-DIS, 질병명의 중간 점을 I-DIS로, 질병이 아닌 부분의 경우 O로 태깅하였다. '만성 피로 증후군'과 같이 공백을 제거했을 때의 표기가 존재하는 경우가 있으며, 이 경우 하나의 질병명에 대한 형태소 분석 결과가 다를 수 있어 모든 데이터 공백을 제거하고 태깅한다. 예를 들어 입력 문장으로 '슬개골 탈구가 의심됩니다.'가 들어올 경우, 공백을 제거한 뒤 태깅한 문장은 "[CLS], _/O, 슬/B-DIS, 개/I-DIS, 골/I-DIS, 탈/I-DIS, 구/I_DIS, 가/O, 의/O, 심/O, 됩/O, 니/O, 다/O, ./O, [SEP]"가 된다. SentencecPiece Tokenizer를 이용하여 토큰화 할 경우 '_'가 문장의 맨 앞에 붙게 되는데, 이는 O로 태깅하였다.

3-3 실험 결과

실험에서는 총 3가지 모델에 대해 3종류의 학습 데이터를 적용하여 결과를 비교한다. 모델로는 기본 BERT 모델에 CRF를 추가한 모델(BERT-CRF)과 bi-GRU에 CRF를 결합한 모델(BERT-bi-GRU-CRF), bi-LSTM과 CRF를 결합한 모델(BERT-bi-LSTM-CRF)을 사용한다. CRF는 통계적 모델링 방법의 하나로, 패턴 인식과 기계학습과 같은 구조적 예측에 사용된다. 순서 라벨링 문제에서 CRF는 주어진 라벨들 사이의 인접성 정보를 기반으로 다음 라벨을 예측할 수 있는 모델로, 관련 연구에서 살펴본 바와 같이 개체명 인식에서 널리 사용된다. BERT는 2018년에 구글이 공개한 사전 훈련된 언어모델로, 트랜스포머를 이용하여 구현되었으며 대량의 텍스트 데이터를 학습한 모델이다. BERT의 기본 구조는 트랜스포머의 인코더 부분을 쌓아 올렸으며, 본 논문에서는 12개의 인코더를 쌓아서 만든 BERT-multilingual base model을 사용한다.

표 2. 모델 성능 평가 결과
Table 2. Result of model evaluation

Data Set	Pet Test		
Model	Human	Pet	Human+Pet
BERT-CRF	0.71	0.74	0.77
BERT-bi-GRU-CRF	0.73	0.76	0.78
BERT-bi-LSTM-CRF	0.74	0.77	0.81

학습 데이터로는 인간 질병 데이터만을 이용한 학습(Human)과 반려동물 질병 데이터만을 이용한 학습(Pet), 인간 질병 데이터를 학습한 모델에 반려동물 질병 데이터를 추가로 학습(Human+ Pet)하는 방법을 사용한다. 3가지 모델은 모두 같은 환경에서 32 batch로 학습한다. 평가를 위한 모델은 Human의 경우 1-3 epoch을 학습한 모델 중 가장 성능이 좋은 모델로 선정하며, Pet은 10-20 epoch 중 선정한다. Human+ Pet은 사람 질병 데이터만을 학습한 모델 중 가장 성능이 좋았던 모델에 반려동물 데이터를 추가로 학습하며, 10-20 epoch 사이의 모델 중 가장 성능이 좋았던 모델을 선정한다.

표 2는 각 모델의 F1-score를 보인다. F1-score의 계산은 [15]와 동일한 방법을 사용한다. 전체적으로 Human+ Pet이 가장 좋은 성능을 보여, 사람의 질병 문서로 기본 모델을 학습하여 성능 향상을 얻을 수 있음을 확인했으며, BERT-bi- LSTM -CRF가 전체적으로 좋은 성능을 보였다.

그림 2는 모델의 개체명 인식 결과 중 가장 성능이 좋은 세 모델의 결과 예시를 보인다. 입력 문장에서 '유선종양'과 '항문선종'은 학습 데이터에 포함되지 않은 질병명이다. BERT-CRF(Human+ PET)는 '항문선종'이 아닌 '문선종'을 질병명으로 인식한다. BERT-bi-LSTM-CRF(Pet)는 '항문선종'은 인식하지 못하며 '유선종양'의 경우 '종양'만을 인식한다. BERT-bi-LSTM -CRF(Human+ Pet)은 '유선종양'과 '항문선종'을 모두 질병으로 인식하지만 '난소종양'의 경우는 '종양'만을 인식한다.

입력 문장:
수컷의 경우에는 행동장애와 **전립선염**, **전립선암**, **항문선종**과 같은 생식기 질환을 예방하고 암컷은 생리와 **유선종양**, **자궁축농증**, **난소종양**과 같은 생식기 질환을 예방할 수 있습니다.

BERT-CRF(Human+Pet):
['전립선염', '전립선암', '문선종', '유선종양', '자궁축농증', '난소종양']

BERT-bi-LSTM-CRF(Pet):
['전립선염', '전립선암', '종양', '자궁축농증', '난소종양']

BERT-bi-LSTM-CRF(Human+Pet):
['전립선염', '전립선암', '항문선종', '유선종양', '자궁축농증', '종양']

* It is an example of a Korean sentence, so it is written in Korean

그림 2. 개체명 인식 예시
Fig. 2. Example of named entity recognition result

미등록어에 대한 개체명 인식 성능을 검증하기 위해, 학습 데이터에 포함되지 않은 161개 질병명만을 포함하는 문장에 대한 별도의 평가를 진행하였다. 1,219개 문장에서 Human+ Pet에의 결과로 BERT-CRF가 0.73, BERT-bi-LSTM-CRF가 0.77, BERT-bi-GRU-CRF가 0.76의 결과를 보여, 학습되지 않은 질병 개체명에 대해서도 높은 성능을 보임을 확인할 수 있었다.

학습한 질병명임에도 제대로 인식하지 못하는 경우의 예로 '쯔쯔가무시병'은 학습에 사용된 질병명임에도 모든 모델이 '쯔쯔'만 질병명으로 인식하였다. 음절 기반 임베딩 사용으로 그림 2에서와 같이 질병명 일부분만을 인식하는 경우와 함께 단어 경계 인식 오류로 인한 문제들도 발견되었다. 한 글자 질병명인 '담'의 경우 문장 내에 존재하는 단어 '상담'의 '담'을 질병으로 태깅하고, 질병명 '한진'의 경우, 공백을 제거한 문장인 "정확한진단"에서 질병명을 태깅하는 오류가 발생했다.

Ⅳ. 질의응답 시스템

그림 3은 구축된 질의응답 시스템의 결과 예시를 보인다. 사용자의 질의에 대해 질병명과 함께 질문에 대한 수의사 답변을 관련도순으로 제시한다. 사용자 질의에 대한 상위 3개의 유사 문서를 cosine 유사도를 사용하여 추출한다.

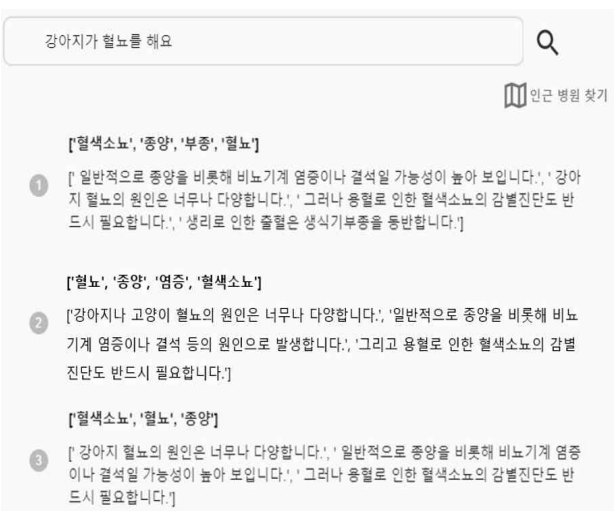
시스템에서는 질의응답에 대한 답변과 함께 그림 4와 같이 사용자 위치 인근 동물병원을 찾는 기능을 제공한다. 카카오 맵 API를 사용하여 장소 검색을 통해 해당 장소에 존재하는 병원의 위치를 제공하여, 동물병원 방문이 필요할 때 편의성을 제공한다.

Ⅴ. 결 론

본 논문에서는 반려동물 질병 개체명 인식에 기반한 반려동물 질병 질의응답 시스템을 제안하였다. 반려동물 질병 데이터 부족 문제를 해결하기 위하여 사람 질병 문서로 기본 모델을 구축하였으며, 자동으로 획득한 질병명 사전으로 질병명이 태깅된 데이터를 구축했다. 모든 모델에서 반려동물 질병 데이터만을 사용했을 때보다 사람 질병 데이터와 동물 질병 데이터를 사용했을 때 향상된 성능을 보였으며, 기본 BERT-CRF 모델보다 BERT-bi-LSTM-CRF 모델이 더 향상된 성능을 보였다.

다양한 전문 분야별 지식베이스를 제공하기 위해 개체명 분석 말뭉치와 분야별 개체명 사전 등이 구축되고 공개되었으나, 동물 질병과 같은 각 세부 분야를 위한 충분한 데이터와 관련 연구가 부족한 실정이다. 최근 MRC(machine reading comprehension)의 발전으로 질의응답 시스템의 성능이 크게 향상되고 있으나, 반려동물 질병의 경우 형태소 분석이나 BERT wordpiece로는 개체명 단위를 찾기 어려워 좋은 성능을 얻기 어렵다. 본 연구에서는 대규모 인간 질병명 사전으로 자동 태깅한 인간 질병 개체명 분석 말뭉치를 사용한 반려동물 질병에 대한 개체명 인식의 성능 향상을 얻어, 이종 데이터를 활용한 성능 향상을 확인하는 동시에 향후 동물 관련 문서에 대한 언어처리의 접근 방법을 제시했다.

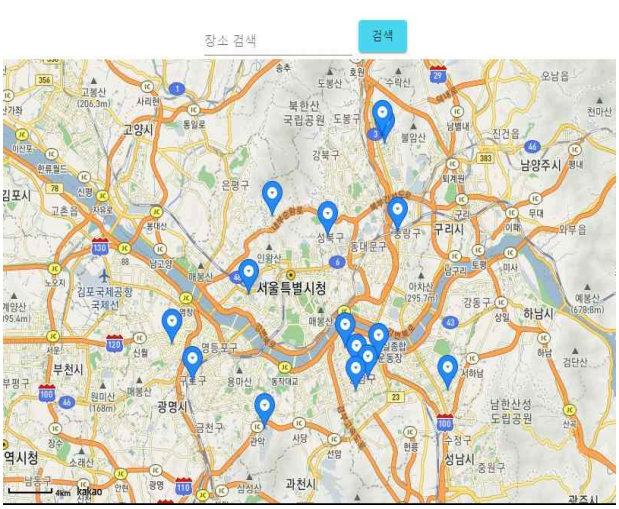
질병명의 경우 기존의 형태소 분석기를 사용하면 정확한 분석 결과를 얻기 어렵기 때문에 본 연구에서는 음절 단위로 개체명에 대한 학습을 진행하였다. 하지만 질병이 아닌 단어를 질병으로 태깅하거나 단어와 단어 사이의 경계를 정확하게 구분하지 못하는 오류가 다수 발견되어, 향후 연구에서는 음절과 BERT wordpiece, 형태소 분석 결과를 결합한 개체명 인식 방식을 수행할 예정이다.



* It is an example of a Korean sentence, so it is written in Korean

그림 3. 사용자 인터페이스 및 출력 결과 예시

Fig. 3. Example of user interface and output result



* It is an example of a Korean map, so it is written in Korean

그림 4. 인근 병원 출력 예시

Fig. 4. Example of output from a nearby hospital

감사의 글

이 연구는 금오공과대학교 대학 학술연구비로 지원되었음(202001930001)

참고문헌

[1] Naver Jisik-iN Expert Answer: Vet [Internet], Available: https://kin.naver.com/people/expert/index.naver?type=ANIMALDOCTOR

[2] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, August 2018. https://doi.org/10.48550/arXiv.1810.04805

[3] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson and M. Vilain, “MITRE: Description of the Alembic System Used for MUC-6”. Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, pp. 141-155, November 1995. https://doi.org/10. 3115/1072399.1072413

[4] A. Borthwick, J. Sterling, E. Agichtein and R. Grishman, “NYU: Description of the MENE Named Entity System as Used in MUC-7”. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, May 1998.

[5] K. H. Lee, J. H. Lee, M. S. Choi and G. C. Kim. “Study on Named Entity Recognition in Korean Text”. Proceedings of the 20th Annual Conference on Human and Cognitive Language Technology, Seoul, pp.292-299, October 2000.

[6] C. K. Lee and M. G. Jang. “Named Entity Recognition with Structural SVMs and Pegasos algorithm”, Korean Journal of Cognitive Science, Vol. 23, No. 4, pp. 655-667, October 2010

[7] C. K. Lee, Y. G. Hwang, H. J. Oh, S. J. Lim, J. Heo, C. H. Lee, H. J. Kim, J. H. Wang and M. G. Jang “Fine-Grained Named Entity Recognition using Conditional Random Fields for Question Answering”, AIRS 2006: Information Retrieval Technology, Singapore, pp. 581-587, October 2006, https://doi.org/10.1007/11880592_49

[8] Y. G. Hwang, B. H. Yun, “HMM-based Korean Named Entity Recognition”, Korea Information Processing Society, Vol. 10B, No. 2, pp. 229-236, April 2003. https://doi.org/10.3745/KIPSTB.2003.10B.2.229

[9] S. W. Kim, Korean Named Entity Recognition Using Two-level Maximum Entropy Model, M.S. dissertation, Yonsei University, Seoul, December 2004.

[10] D. Y. Lee, W. H. Yu and H. S. Lim, “Bi-directional LSTM-CNN-CRF for Korean Named Entity Recognition System with Feature Augmentation”, Journal of the Korea Convergence Society, Vol. 8, No. 12, pp. 55-62, December 2017. https://doi.org/10.15207/JKCS.2017.8.12.055

[11] S. H. Hwang, S. H. Shin, D. G. Choi, S. H. Kim, J. E. Kim, “Korean Named Entity Recognition using BERT”, Proceedings of the Korea Information Processing Society Conference, Jeju, pp. 20-822, November 2019. https://doi.org/10.3745/PKIPS.y2019m10a.820

[12] Y. S. You, H. R. Park, “Syllable-based Korean Named Entity Recognition Using Convolutional Neural Network”, Journal of Advanced Marine Engineering and Technology, Vol. 44, No. 1, pp. 68-74, January 2020. https://doi.org/10.5916/jamet.2020.44.1.68

[13] Z. Zhao, Z. Yang, L. Luo, L. Wang, Y. Zhang, H. Lin and J. Wang, “Disease named entity recognition from biomedical literature using a novel convolutional neural network”, BMC Med Genomics, Vol. 10, No. 73, pp. 76-83, December 2017. https://doi.org/10.1186/s12920-017-0316-8

[14] J. H. Lee, W. J. Yoon, S. D. Kim, D. H. Kim, S. K. Kim, C. H. So and J. W. Kang, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”, Bioinformatics, Vol. 36, No. 4, pp. 1234−1240, September 2019. https://doi.org/10.1093/bioinformatics/btz682

[15] M. S. Cho, J. U. Park, J. H. Ha, C. H. Park and S. H. Park, “Biomedical Named Entity Recognition using Multi-head Attention with Highway Network”. Journal of KIISE, Vol. 46, No. 6, pp. 544-553, June 2019. https://doi.org/10.5626/JOK.2019.46.6.544

[16] S.O. Na, Y.M. Kim and S.W. Kong, “Clinical Entity Recognition and Normalization for Medical Advice”, Proceedings of Korea Computer Congress(KCC), Jeju, pp. 343-345, June 2021.

[17] H. S. Oh, J. W. Seo, H. J. Kim, H. j. Kim, S. S. Kim and H. Y. Lee, “Pet Disease Q&A System Using Named Entity Recognition”, Proceedings of KIIT Conference, Jeju, pp 512-514, November 2021.

[18] KoCharELECTRA [Internet], Available: https://github.com/monologg/KoCharELECTRA

[19] Naver Jisik-iN, Expert Answer-Doctor, Available: https://kin.naver.com/people /expert/index.naver?type=DOCTOR

[20] Korea Centers for Disease Control and Prevention [Internet], Available: https://www.kdca.go.kr/

[21] Asan Medical Center [Internet], Available: http://www.amc.seoul.kr/asan/healthinfo/disease/diseaseSubmain.do

[22] SEOUL NATIONAL UNIVERSITY HOSPITAL [Internet],

Available: http://www.snuh.org/health/encyclo/search.do
[23] Korea Institute of Science and Technology Information_Animal disease data, Data Portal [Internet], Available: https://www.data.go.kr/data/15050442/ fileData.do

오한석(Han-Seok Oh)

2019～현　재 : 금오공과대학교 컴퓨터소프트웨어공학과 학사과정

※관심분야 : 자연어처리, 인공지능, 기계학습 등.

이현아(Hyunah Lee)

1996년 : 연세대학교 컴퓨터과학과 (학사)
1998년 : KAIST 전산학과 (석사)
2004년 : KAIST 전산학과 (박사)
2000년～2004년 : ㈜다음소프트 언어처리연구소
2004년～현　재 : 금오공과대학교 컴퓨터소프트웨어공학과 교수

※관심분야 : 자연언어처리, 텍스트데이터마이닝, 정보검색 등