

검색의도에 적합한 스니펫 추출

이현구⁰, 양윤영, 김은별, 차우준, 노윤영, 김은영, 최규현, 신동욱, 박찬훈, 강인호

NLP, 네이버

{hyeongu.lee, yunyeong.yang, silverstar.kim, woojune.cha, yunyoung.roh, keysilver.kim, gyuhyeon.choi, shin.dongwook, chanhoon.park, once.ihkang}@navercorp.com

Extract Snippets Suitable for Search Intent

Hyeon-gu Lee⁰, Yunyeong Yang, Eunbyul Kim, Woojune Cha, Yunyoung Roh, Eunyoung Kim, Gyuhyeon Choi, Dongwook Shin, Chanhoon Park, Inho Kang
NLP, Naver Corporation

요약

스니펫 추출은 정보검색에서 주요한 문서 정보를 짧은 문단 형태로 보여주는 것으로 사용자가 검색결과를 좀 더 효율적으로 확인할 수 있게 도와준다. 그러나 기존 스니펫은 어휘가 일치하는 문장을 찾아 보여주기엔 검색의도가 반영되기 어렵다. 또한 의미적 정답을 찾기 위해 질의응답 방법론이 응용되고 있지만 오픈 도메인 환경에서 품질이 낮은 문제가 있다. 본 논문은 이러한 문제를 해결하기 위해 스니펫 추출, 의도 부착, 검증 3단계로 스니펫을 추출하여 추출된 스니펫이 질의 의도에 적합하게 추출되도록 하는 방법을 제안한다. 실험 결과 전통적인 스니펫보다 만족도가 높은 것을 보였고, 스니펫 추출만 했을 때보다 의도 부착, 검증을 하였을 때 정확도가 0.3165만큼 향상되는 것을 보였다.

주제어: 스니펫 추출, 의도 부착, 스니펫 검증

1. 서론

온라인의 정보량이 증가함에 따라 정보를 쉽게 찾을 수 있도록 도와주는 정보검색(information retrieval) 기술은 꾸준히 발전되었다. 정보검색 기술은 질의와 의미적, 어휘적으로 유사한 문서를 찾아 보여주고, 나아가 문서에서 질의와 관련있는 문단을 스니펫(snippet) 형태로 보여준다. 그러나 스니펫은 주로 어휘가 일치하는 문장을 찾아 보여주기엔 질의 검색의도에 대한 정답이 아닐 수 있어 사용자가 문서를 직접 확인하여 원하는 정보를 얻어야 한다. 이러한 문제는 의미적 정답을 찾아주는 질의응답(question Answering) 방식의 문장 추출을 사용하여 어느정도 해소할 수 있지만 일반적인 검색 환경인 오픈 도메인(open-domain)에서는 성능이 낮아 추출된 스니펫의 품질 문제가 있다. 본 논문은 오픈 도메인에서 검색의도에 적합한 고품질의 스니펫을 추출하기 위해 질의응답 기술 기반의 스니펫 후보 문장 추출과 의도 및 질의 적합성을 판단하는 검증모델을 결합한 스니펫 추출 방법을 제안한다.

2. 관련 연구

기존 스니펫 추출 방식은 질의와 어휘적으로 일치하는 문장을 나열하여 문서 내 정보를 일부 보여주지만 검색 의도와 다른 문장을 보여줄 수 있다. 이를 위해 다양한 방법을 통해 의미적 정보를 파악하도록 연구되었다.[1-4] 최근 사전학습 언어모델(pre-trained language model)을 통해 자연어처리 여러 분야에서 높은 성능을 보임에 따라 질의응답, 문장 추출 또한 의미적 특성을 반영하여

좋은 성능을 보이고 있다.[5-6] 본 논문에서는 사전학습 언어모델을 기반으로 질의에 알맞는 스니펫을 추출하고, 추출된 스니펫에 질의의도를 부착 후 검증하여 의미적 특성을 반영할 뿐만 아니라 질의의도를 통해 보다 정교한 스니펫을 추출하는 방법을 제안한다.

3. 검색의도에 적합한 스니펫 추출

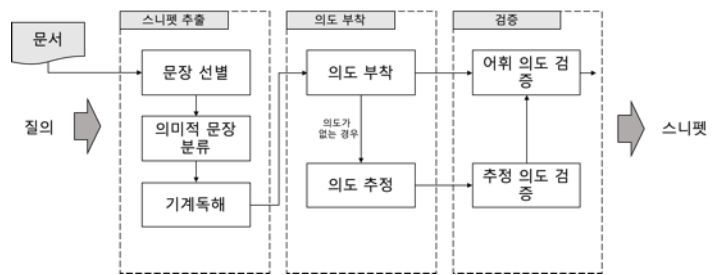


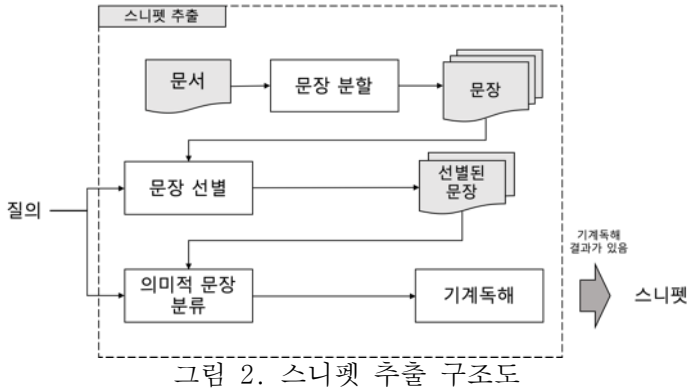
그림 1. 전체 구조도

그림 1은 제안 모델의 구조도이다. 제안 모델은 문서에서 스니펫을 추출하는 추출 모델, 질의를 통해 추출된 스니펫 후보의 의도를 부착하는 의도 부착, 질의에 대해 적절하게 추출됐는지 검증하는 검증 모델로 구성된다. 추출 모델은 문서에서 질의를 가장 잘 표현하는 문장을 추출하는게 목적이며, 주요 문장을 선별하는 문장 선별, 질의와 의미적 연관이 있는지 판단하는 문장 분류, 실제 정답이 있는지 판단하는 기계독해로 구성된다. 의도 부착은 검색의도 파악을 목적으로 하며, 질의의 어휘적 특성을 고려해 질의 의도를 판단하는 의도 부착, 생략된 의도나 모호한 의도를 찾아내기 위한 의도 추정으로 구성된다. 마지막

검증 모델은 질의의 의도가 스니펫에 제대로 포함됐는지 판단하는 의도 검증과 추정된 의도가 적합한지 판단하는 추정 의도 검증으로 구성된다.

분류 모델을 통해 질의와 연관됐는지를 분류한다. 그림 4는 문장 분류 모델을 나타낸다.

3.1. 스니펫 추출



본 논문에서 제안하는 스니펫 추출은 그림 2와 같다. 문장 선별은 문서에서 질의와 관련 없는 문장을 어휘 정보를 통해 제외하고, 의미적 문장 분류를 통해 질의와 의미적 연관이 있는 문장을 찾고, 마지막 기계독해를 통해 실제 정답이 포함됐는지를 파악한다.

3.1.1. 문장 선별

문장 선별은 질의 팀(term)이 하나 이상 포함된 120~300자 내외의 연속된 문장을 찾는다. 이는 질의에 대해 답을 할 수 있는 문장이라면 정답이 포함된 문장 앞, 뒤로 질의와 관련됐음을 알 수 있도록 표현하기 때문이다. 그림 3은 문장 선별을 보여주는 예시이다

Idx	문장
1	상반신의 이상으로 고혈압이 생긴 경우 손을 주물러 혈압을 낮출 수 있는데요,
2	이 방법은 하나의 혈이 아니라 2개의 혈을 이용하는데, 바로 대장경의 합곡혈 과 소장경의 후계혈 입니다.
3	그럼 합곡혈 위치 및 효능과 후계 혈자리 는 어디인지손등을 위로 하고 엄지손가락과 두번째 손가락을 벌린 다음 두번째 손가락뼈와 엄지손가락뼈를 따라 손목 쪽으로 밀어올라가면 엄지손가락뼈와 두번째 손가락 뼈가 합친 곳에 이르게 되는데, 이 합친 곳이 합곡혈 입니다.
4	목 혈관의 긴장을 풀고 심장을 잘 움직이게 해 혈압을 낮추는 효능이 있습니다.
5	대장경은 목을 통과하고 있는데, 목에는 인영이라는 혈이 있습니다.
6	이 혈은 혈관의 반사와 자율신경의 반사에 중요한 역할을 합니다.
7	합곡혈 은 인영과 같은 대장경으로 연결되어 있으므로, 합곡을 자극하면 경락의 반사로 인영을 자극했을 때와 같은 효과를 나타내어 목을 누르지 않아도 목 혈관의 긴장을 풀어줄 수 있는 것입니다.
...	...
12	이때 합곡혈 의 주변을 폭넓게 자극할 수 있도록 합곡혈 의 위치를 조금씩 옮겨가면서 누르면 자극 효과가 합곡혈 에 다른쪽 손의 엄지손가락을 대고 고정시킨 다음 세번째 손가락으로 후계혈을 자극합니다.
13	이 역시 위치 를 조금씩 옮겨가면서 후계혈 주변을 두가지 방법 모두 한번에 5분 정도 자극을 주며, 양 손을 합하여 하루 10분 정도 실시하면 온몸의 긴장이 풀리고 혈압이 떨어집니다.

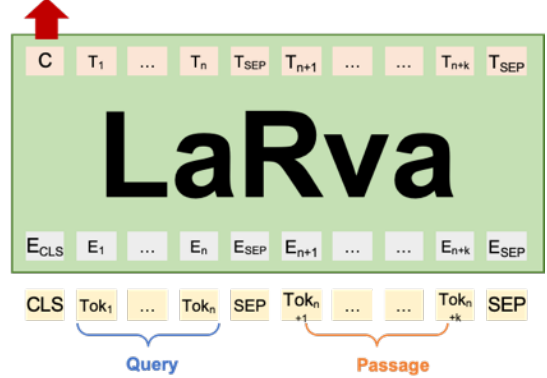
그림 3. 문장 선별의 예

그림 3에서 보는 것과 같이 질의 “합곡혈 위치”의 팀인 “합곡혈”과 “위치”가 포함된 2, 3, 7, 12, 13번째 문장을 선별한다.

3.1.2. 의미적 문장 분류

질의 어휘를 통해 선별된 후보 문장은 의미적 문장

문장의 스니펫 여부 (0/1)



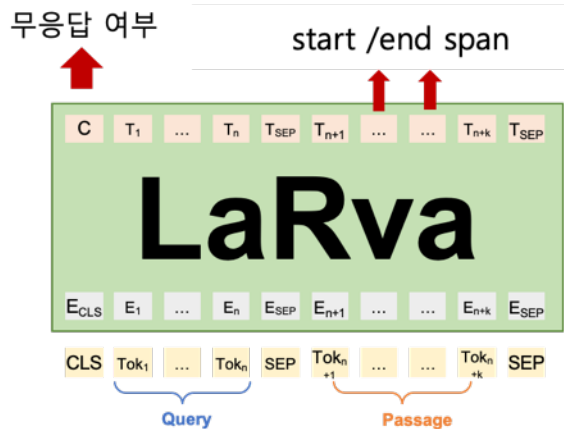
Idx	스니펫	label
3	그럼 합곡혈 위치 및 효능과 후계 혈자리 는 어디인지손등을 위로 하고 엄지손가락과 두번째 손가락을 벌린 다음 두번째 손가락뼈와 엄지손가락뼈를 따라 손목 쪽으로 밀어올라가면 엄지손가락뼈와 두번째 손가락 뼈가 합친 곳에 이르게 되는데, 이 합친 곳이 합곡혈 입니다.	1
12-13	이때 합곡혈 의 주변을 폭넓게 자극할 수 있도록 합곡혈 의 위치를 조금씩 옮겨가면서 누르면 자극 효과가 합곡혈 에 다른쪽 손의 엄지손가락을 대고 고정시킨 다음 세번째 손가락으로 후계혈을 자극합니다. 이 역시 위치 를 조금씩 옮겨가면서 후계혈 주변을 두가지 방법 모두 한번에 5분 정도 자극을 주며, 양 손을 합하여 하루 10분 정도 실시하면 온몸의 긴장이 풀리고 혈압이 떨어집니다.	0

그림 4. 문장 분류 모델

그림 4에서 보는 것과 같이 문장 분류 모델은 질의와 후보 문장을 입력하여 문장이 사용될지를 판단하는 이진 분류 모델이다. 질의와 후보 문장을 입력하고 [CLS]의 출력 부분에 가중치 행렬을 곱한 결과를 이진 분류한다.

3.1.3. 기계독해를 통한 정답 추출

3.1.2에서 분류된 문장에 정답이 포함되어 있는지를 판단하기 위해 기계독해를 적용한다. 정답이 포함되어있고 질의와 관련이 있다면 기계독해 모델을 통해 결과를 추출할 수 있기 때문이다. 그림 5는 선별된 문장에서 정답을 찾기 위한 기계독해 모델을 나타낸다.



Idx	스니펫	answer
3	그럼 합곡혈 위치 및 효능과 후계 혈자리 는 어디인지손등을 위로 하고 엄지손가락과 두번째 손가락을 벌린 다음 두번째 손가락뼈와 엄지손가락뼈를 따라 손목 쪽으로 밀어올라가면 엄지손가락뼈와 두번째 손가락 뼈가 합친 곳에 이르게 되는데, 이 합친 곳이 합곡혈 입니다.	엄지손가락뼈와 두번째 손가락 뼈가 합친 곳

그림 5. 기계독해 모델

기계독해 모델은 정답의 위치를 예측하는 부분과 정답이 있는지를 판단하는 무응답 분류로 구성된다. 정답 위치는 시작 위치, 끝 위치를 사용하며 [5]에서 사용한 방식과 동일하게 사용한다. 무응답 분류는 [CLS]의 출력 부분에 가중치 행렬을 곱한 결과를 이진 분류하여 구별하며 무응답의 경우 시작 위치, 끝 위치는 사용하지 않는다.

3.2. 질의 의도 부착

질의 의도 부착은 질의가 나타내고자 하는 의도를 분석하여 추출된 스니펫이 질의 의도에 적합한지를 판단하기 위한 선행작업이다. 질의 의도 부착은 사전에 정의된 질의 의도와 질의 의도에 해당하는 술어를 통해 명시적으로 언급된 의도를 부착한다. 의도를 나타내는 어휘가 생략됐으면 형태가 유사한 질의와 문서 정보를 통해 의도를 추정한다.

3.2.1. 어휘를 통한 명시적 의도 부착

의도를 알 수 있는 표현이 질의에 있는 경우 사전에 정의된 질의 의도 사전을 통해 의도를 부착한다. 그림 6은 질의 의도 사전을 통한 명시적 의도 부착의 예시이다.

질의 의도 사전	
의도	술어
definition	뜻, 의미, 정의, 개념, 이란
money	가격, 값, 금액, 비용 ...
...	...

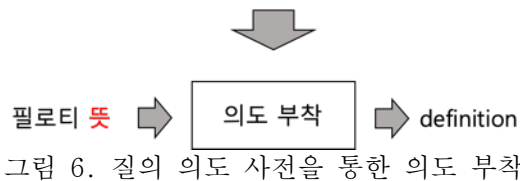


그림 6. 질의 의도 사전을 통한 의도 부착

그림 6에서 보는것과 같이 질의 의도 사전은 ‘의도-술어의 집합’으로 구성된다. 예시와 같이 “필로티 뜻”이란 질의는 “뜻”과 같이 의도를 알 수 있는 술어를 포함하고 있고 질의 의도 사전에서 뜻에 해당하는 “definition”을 의도로 부착한다.

3.2.2. 생략된 의도 추정

일부 질의는 의도가 생략되어 생략된 의도를 복원하거나 생략된 어휘로 인해 질의의 의도가 모호하여 명확한 의도를 추정해야한다. 본 논문은 이러한 의도 추정을 위하여 형태가 유사한 질의와 문서 정보를 활용한다. 먼저 형태가 유사한 질의는 “전체 질의 집합에서 의도가 생략된 질의 원형이 부분 문자열(substring)로 표현되는 질의”를 의미한다. 즉, 의도가 생략된 질의 앞뒤로 붙는 추가 어휘 정보를

활용한다. 이렇게 확보한 형태가 유사한 질의를 3.2.1에서 부착한 방법으로 의도를 부착한다. 만약 형태가 유사한 질의가 없거나 여전히 의도가 부착되지 않는 경우 문서 정보를 활용한다. 문서 정보는 문서의 제목을 사용한다. 문서의 제목에 질의 의도 사전에 포함된 의도를 추정할 수 있는 어휘 표현이 있을 경우 그 표현을 생략된 의도라 가정하여 부착한다.

3.3. 스니펫 검증

본 논문에서는 추출된 스니펫과 질의 의도가 적합한지 판단하기 위해서 스니펫 검증을 한다. 스니펫 검증은 의도 추정을 통해 부착된 의도가 추출된 스니펫과 적합한지 판단하는 추정 의도 검증과 특정 의도에 대하여 필수적인 어휘가 스니펫에 포함됐는지 판단하는 어휘 의도 검증으로 구성된다

3.3.1. 추정 의도 검증

의도를 추정하여 부착했을 경우 부착된 의도가 적합하게 확장되어 부착되었는지 검증해야한다. 스니펫 추출 때 사용된 질의 원형과 추정을 위해 사용된 형태가 유사한 질의가 의미적으로 다를 경우 스니펫이 적합하지 않게 된다. 이를 해결하기 위하여 3.1.2에서 사용한 의미적 문장 분류 모델을 통해 의미가 동일한지 검증한다. 기존 추출된 스니펫은 질의 원형을 통해 추출된 스니펫이고 형태가 유사한 질의가 질의 원형과 의미가 같다면 의미적 문장 분류 모델이 두 질의의 스니펫을 동일하게 분류한다는 가정에 기반한다. 그림 7은 추정 의도를 검증하는 과정을 나타낸다.

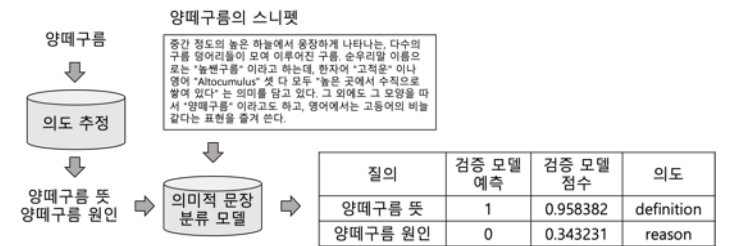


그림 7. 의도 검증 모델의 예

그림 7에서 보는것과 같이 질의 원형 “양떼구름”은 의도가 생략되어 있어 3.2.2의 과정을 통해 “양떼구름 뜻”과 “양떼구름 원인”을 형태가 유사한 질의로 사용하여 “definition”과 “reason”이란 의도로 부착하였다. 예시에서 보는것처럼 질의 원형 “양떼구름”에서 추출된 스니펫을 “양떼구름 뜻”과 “양떼구름 원인”을 입력으로 검증 모델(의미적 문장 분류 모델)에 입력한다. 검증 모델이 “양떼구름 뜻”은 예측한 결과가 긍정이고 검증 모델 점수도 높기에 유효한 추정이다. 반면 “양떼구름 원인”은 해당 스니펫에는 원인에 대한 내용이 포함되어 있지 않기에 검증 모델이 부정으로 예측하여서 해당 추정은 잘못된 것을 알 수

있다.

3.3.2. 어휘 의도 검증

질의 의도 중 일부는 반드시 포함되어야 하는 특정 어휘가 존재한다. 대표적으로 “~가격” 과 같이 money로 정의된 의도는 “0000원”, “000 달러” 등 금액을 나타내는 어휘가 포함되어야 하고 “~배우자”의 경우 사람의 이름을 나타내는 어휘가 포함되어야 한다. 본 논문에서는 추출된 스니펫을 정규화하여 표현단위(날짜, 시간, 가격, 길이 등)를 간단한 규칙으로 추출하고 개체명 인식기를 통해 의도 필수 어휘를 탐지한다. 그림 8은 어휘 의도 검증을 통해 스니펫이 유효한지 판단하는 예시이다.

질의 : 네이버 멤버십 가격
의도 : money

가격에 대한 표현

질의	스니펫
스니펫 후보 1	네이버플러스 멤버십은 매월 4,900원 (VAT포함)으로 다양한 혜택을 이용할 수 있습니다. 처음 한 달은 무료이고, 두 번째 달부터 월 4,900원 (VAT 포함)이 결제됩니다. 연간 이용권으로 가입하면 약 20% 저렴하게 이용할 수 있습니다.
스니펫 후보 2	네이버 멤버십은 매우 저렴한 가격입니다.

그림 8. 어휘 의도 검증의 예

그림 8에서 보는 것처럼 “네이버 멤버십 가격”의 경우 의도는 “가격” 이고 이 때 금액에 대한 표현이 필요하다. 첫번째 후보 스니펫은 “4,900원”이라는 표현이 있기에 가격이 포함되어 의도에 맞게 추출된 것을 알 수 있다. 반면 두번째 후보 스니펫은 가격과 관련된 문장은 맞지만 정답성 어휘가 포함되어 있지 않기에 결과로 부적합하다.

4. 실험 및 평가

4.1. 실험 데이터

본 논문에서 스니펫 추출을 위해 자체 구축한 데이터를 사용한다. 의미적 문장 분류 모델은 943,806개의 질의-문단 데이터를 구축하였고 81,236개의 긍정 데이터(positive data), 862,570개의 부정 데이터(negative data)로 구성되어있다. 해당 데이터를 중 9,343개(긍정 2,649개; 부정 6,694개)를 평가 데이터로 사용한다. 기계독해 데이터는 총 576,218개의 질의-정답 데이터를 구축하였으며 537,544개의 정답이 있는 데이터, 38,674개의 정답이 없는 데이터로 구성된다. 이 중 11,249개(정답이 있는 데이터 6,668개; 정답이 없는 데이터 4,581개)를 평가 데이터로 사용한다.

학습에 사용된 사전학습 언어모델은 LaRva[7]를 사용하였으며 hidden size 768, 계층 수는 12이다.

4.2. 실험 평가

4.2.1. 스니펫 추출 평가

스니펫 추출은 각 모델의 성능을 알 수 있는 정량 평가, 오픈 도메인 상황에서 실제 만족도를 알 수 있는 정성 평가로 진행한다. 먼저 정량 평가는 스니펫 추출에서 사용된 의미적 문장 분류의 분류 성능과 분류된 문장에 정답이 포함됐는지 판단하기 위한 기계독해 성능을 평가한다. 의미적 문장 분류는 정확률(precision), 재현율(recall), F1-score로 평가하며, 기계독해는 완전 일치율(exact match), F1-score로 평가한다. 표 1은 의미적 문장 분류 모델의 성능을 나타낸다.

표 1. 의미적 문장 분류 성능

	Precision	Recall	F1-score
positive	0.6495	0.6961	0.6720
negative	0.6726	0.6244	0.6476

표 1에서 의미적 문장 분류 모델은 사용하는 문장인 positive, 사용을 하지 않는 문장인 negative에 대해 비슷한 성능을 보인다. 다만 의미적 문장 분류 성능이 낮을 것을 알 수 있는데 이는 평가에 사용된 문서가 다양한 형태를 가진 오픈 도메인 문서이고 시의성, 완결성, 가독성 등을 모두 고려한 데이터이기 때문에 난이도가 높다. 그림 9에서 사용된 데이터의 일부를 나타낸다.

질의	문장	정답	사유
자동차 소모품 교환시기	소모품은 제때에 반드시 교환 자동차의 소모품 교환주기를 넘기면 잔고장의 원인이 돼 차수명을 줄이는 주범이 됩니다. 정화플러그는 2만km마다, 엔진오일은 주행환경을 고려해 5,000~10,000km 주기로 교환합니다.	negative	자동차 부품의 일부만 표시
스마트캐디가격	eWheel V2 스마트 캐디 휠 주요 가격 · eWheel V2 : \$499 (한국배송료 +\$110) / 총 \$609 (약 72만2천원) ※ 기타 옵션은 후원시 선택 가능하며, 인디고고 클라우드 편딩은 즉시 결제 되므로 신중히 선택하십시오.	negative	해당 부분은 가격이 아닌 편딩 가격

그림 9. 의미적 문장 분류 평가 데이터의 예

그림 9에서 “자동차 소모품 교환시기”의 경우 문장에 일부 부품의 교환주기만 표시되어 negative으로 분류되었다. 질의가 “자동차 부품 점화플러그 교환시기”면 positive겠지만 “자동차 소모품 교환시기”에는 의미적으로 완전한 문장이 아니라 negative으로 분류된 것이다. “스마트캐디가격”은 문장만 봐서는 가격이 표시되어 있지만 해당 문장은 “편딩 가격” 이고 “실제 가격” 이 아니기에 오답이 포함된 문장이라 negative로 되어있다. 즉, 이렇게 어려운 문장이 포함된 데이터이기 때문에 모델이 풀기 어려워 기본 성능이 낮음을 알 수 있다.

표 2는 기계독해의 성능을 나타낸다.

표 2. 기계독해 성능

	Exact Match	F1-score
has answer	0.6045	0.8091
no answer	0.9127	0.9127
total	0.7502	0.8581

표 2에서 has answer는 정답이 포함된 경우의 성능이고 no answer는 정답이 없는 경우 성능이다. 본 논문에서 사용한

기계독해 모델은 정답이 없는 경우 0.9127로 높은 성능을 보인다. 반면 정답이 있는 경우 Exact Match가 0.6045로 높진 않지만 제안 방법에서 기계독해 모델은 문장 내 정답이 포함됐는지를 판단하는 것으로 F1-score와 같이 정답 부분이라도 잡아내는게 중요하기에 0.8091의 수준으로 정답을 탐지하게 된다.

다음으로 오픈 도메인 상황에서 스니펫 추출이 어느 정도의 커버리지(coverage)와 정확도(accuracy)를 보이는지 정성 평가한다. 오픈 도메인 상황의 평가를 위해 전체 질의 집합에서 임의의 4천개의 질의를 사용하였다. 질의에 대한 문서는 네이버 웹 검색결과를 사용하였으며 질의는 실제 검색 환경과 동일하게 질의응답성 질의 이외에 키워드성 질의를 포함하였다. 추출된 스니펫은 330개로 약 8.25% 정도의 질의만 스니펫이 추출되었다. 이 결과를 바탕으로 2가지 정성평가를 진행하였다. 먼저 스니펫이 추출된 330개 질의에 대하여 추출된 스니펫이 적합한지 평가한다. 이는 의도 부착 여부와 상관없이 스니펫 추출 모델의 성능을 의미한다. 표 3은 추출된 330개에 대해 평가한 결과이며 정답이 아닌 경우 아닌 이유를 나타낸다.

표 3. 스니펫이 추출된 경우 평가 결과

	데이터의 수	비율
정답	137	41.52%
스니펫 오류	59	17.87%
질의 오류	40	12.13%
문서 오류	94	28.48%

표 3에서 정답은 오픈 도메인 상황에서 임의의 질의를 입력하였을 때의 성능을 나타낸다. 스니펫 오류는 문서와 질의는 아무런 문제가 없지만 스니펫 추출 모델이 잘못된 스니펫을 추출한 경우이다. 질의 오류는 “구월동 1372-10” 처럼 스니펫 추출에 적합하지 않은 질의를 의미하며, 문서 오류는 대상으로 한 문서에 정답이 존재하지 않는 경우를 의미한다. 결과를 바탕으로 오픈 도메인 상황에서 0.4152의 정확도를 보인다. 추출 모델 단독 성능은 식 1과 같이 계산되며 0.6989의 정확도를 보이며 표 1의 성능과 유사함을 알 수 있다.

$$Accuracy = correct / (correct + snippet\ err) \quad (1)$$

다음으로 추출되지 않은 2,670개의 질의 중 모델 개선을 통해 추가로 추출할 수 있는 경우를 평가한다. 표 4는 2,670개의 데이터 중 200개를 샘플링하여 측정한 결과를 나타낸다.

표 4. 스니펫이 추출되지 않은 경우 평가 결과

	데이터의 수	비율
추출이 가능한 문서	26	13%
사용할 수 없는 질의	147	73.5%
정답이 없는 문서	26	13%
스니펫 타입이 다름	1	0.5%

표 4에서 추출이 가능한 문서는 정교한 스니펫 추출 모델이라면 추출을 할 수 있는 문서이지만 본 논문에서 사용된 모델은 추출하지 못한 케이스로 모델 개선을 통해 해결해야하는

문제이다. 사용할 수 없는 질의는 표 3의 질의 오류와 같은 경우, 정답이 없는 문서는 표 3의 문서 오류와 같은 경우이며 스니펫 추출 모델이 잘 걸러낸 경우를 의미한다. 스니펫 타입이 다른 문서 내에 정답에 해당하는 내용이 있지만 문장이 아닌 표, 그림과 같이 해당 모델이 추출할 수 없는 경우이다. 추출하지 못한 경우에서 약 13% 정도는 추가로 추출할 수 있는 경우로 2,670개 중 347개의 질의는 스니펫이 추출되어야 하지만 추출하지 못한 경우이다.

4.2.2. 의도 부착 및 검증 평가

의도 부착 및 검증은 오픈 도메인 상황에서 추출된 스니펫을 평가한다. 평가 대상은 4.2.1에서 추출된 330개의 질의이며 표 5는 각 모듈 별 처리되는 질의 수를 나타낸다.

표 5. 검증 모듈 별 데이터 분포

	의도 부착 수	검증 통과 수
어휘 의도 부착	40	39
의도 추정 (질의)	9	2
의도 추정 (문서)	154	41
미부착	127	0
합계	330	82

표 5에 의도 부착 수는 의도 부착 모델에서 의도가 부착되는 모듈 및 데이터의 수를 의미한다. 검증 통과 수는 의도가 부착된 결과를 검증 모델을 통해 검증 후의 데이터 수를 의미한다. 330개의 질의 중 40개는 질의에 의도를 나타내는 어휘가 표시되어 있어 의도가 부착되었고 9개는 형태가 유사한 질의를 통해 부착, 154개는 문서의 제목에 의도와 관련된 어휘가 나타난 것이다. 검증 통과 수의 비율이 어휘 의도 부착에 비해 의도 추정이 낮은 이유는 3.3.1의 추정 의도 검증 모델 때문이며 추정된 의도가 실제로 적합하지 않아 제외되었기 때문이다.

다음은 각 검증 모델을 적용하였을 때 성능을 평가한다. 표 6은 표 5에서 검증을 통과한 데이터들 82개를 정성평가한 결과이다.

표 6. 검증 모델 정성 평가 결과(정확도)

	의도 부착 + 검증	검증
어휘 의도 부착	0.7435	0.8974
의도 추정 (질의)	1.000 (0.8000)*	1.000 (0.9200)*
의도 추정 (문서)	0.7073	0.8780
Micro Accuracy	0.7317	0.8902

* 은 다른 샘플을 통해 성능을 측정한 결과

‘의도 부착 + 검증’ 은 의도 부착과 검증 결과 모두를 맞췄을 때의 성능을 의미하고 ‘검증’ 은 부착된 의도가 정답일 때 검증 모델의 단독 성능을 나타낸다. ‘검증’ 결과를 바탕으로 스니펫에 의도가 올바르게 부착 될 경우 높은 검증 성능을 보였다. 결론적으로 스니펫 추출 결과의 품질이 의도 부착 및 검증을 통과했을 경우 전체 커버리지는 낮아지지만(137개->82개)

스니펫의 정확도는 0.7317로 스니펫만 추출한 0.4152보다 0.3165만큼 향상된 성능을 보였다. 논외로 ‘의도 추정 (질의)’의 경우 표본이 너무 작아 성능이 왜곡되어 ‘의도 추정 (질의)’과 동일한 방법으로 추출한 100개의 샘플을 통해 성능을 추가로 측정하였으며 ‘의도 부착 + 검증’의 경우 0.8000, ‘검증’의 경우 0.9200의 성능을 보였다.

4.2.3. 일반 스니펫과 비교

본 논문에서 추출한 스니펫이 질의 의도를 잘 표현하고, 의미적으로 추출됐는지를 보이기 위해 전통적인 어휘 일치 기반의 스니펫과 품질을 비교한다. 표 7은 네이버 웹검색 결과에 노출되는 스니펫과 제안 모델에서 추출한 스니펫을 1,000개의 샘플 질의로 비교하였을 때의 분포를 나타낸다.

표 7. 제안 스니펫과 일반 스니펫의 품질 비교

분류	데이터 수
제안 스니펫 > 일반 스니펫	591
제안 스니펫 = 일반 스니펫	398
제안 스니펫 < 일반 스니펫	11

표 7에서 보는 것과 같이 59.1%는 본 논문에서 제안한 스니펫 추출 방법으로 추출된 스니펫이 더욱 좋았으며 나쁜 경우는 11개로 1.1%로 매우 적어 본 논문에서 제안한 스니펫이 우수함을 보였다.

5. 결론

본 논문에서는 검색의도에 적합한 스니펫을 추출하는 방법을 제안하였다. 제안 방법은 스니펫 추출에서 문장 선별, 의미적 문장 분류, 기계독해 3단계를 통해 질의에 적합한 스니펫을 추출하였고 의도 분류를 통해 명시적으로 표현된 의도, 표현이 생략된 의도를 추정하였다. 마지막 검증에선 의미적 문장 분류 모델을 통해 의도 추정을 검증하고 스니펫에 반드시 포함되어야 할 어휘가 있는지 의도와 관련하여 검증하였다. 실험 결과 스니펫 추출 모델은 전통적인 스니펫보다 만족도가 높았으며 스니펫 추출만 했을 때보다 의도 부착, 검증을 적용하였을 때 0.3165만큼의 정확도 향상이 있어 본 논문에서 제안한 의도 분류 및 검증이 유효함을 보였다.

참고문헌

[1] 장재형, 김태연, 김재광, 이지형, "개체명을 활용한 스니펫 생성 방법.", *한국지능시스템학회 학술발표 논문집* 20.1, pp. 178-179, 2010.
 [2] 박선, 이연우, 조광문, 양후열, 이성로, "퍼지와 의미특징을 이용한 스니펫 추출 향상 방법.", *한국정보통신학회논문지* 16.11, pp. 34-41, 2012.
 [3] 김태환, 이진희, 김정선, "문장 중요도를 이용한 스니펫 생성 방법.", *한국정보과학회 학술발표논문집*, pp. 1024-1026, 2017.

[4] C. Zhang, X. Zhang and H. Wang, "A machine reading comprehension-based approach for featured snippet extraction.", *Proceedings of 2018 IEEE International Conference on Data Mining (ICDM)*, pp. 1416-1421, 2018.
 [5] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp. 4171-4186, 2019.
 [6] K. Clark, M. T. Luong, Q. V. Le and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", *Proceedings of the 8th International Conference on Learning Representations*, 2020.
 [7] 이동준, 김성동, "엄~청 큰 언어 모델 공장 가동기! (LaRva: Language Representation by Clova)", URL: <https://devview.kr/2019/schedule/291>, 2019.