Lexicon-Assisted Training 을 통한 다국어 Text-to-SQL 성능 향상

장민성, 강성범, 김병창*, 박지홍* 박원준*, 홍상은*
㈜범일정보, *대구가톨릭대학교 컴퓨터소프트웨어학부

msjang@bumil.co.kr, ksb0806@bumil.co.kr, *bckim@cu.ac.kr, *rain1509@naver.com,
*one01021@cu.ac.kr, *hse09@cu.ac.kr

Improving Multilingual Text-to-SQL via Lexicon-Assisted Training

Jang Min Sung, Kang Sung Bum, Kim Byeong Chang, Park Ji Hong, Park Won Jun, Hong Sang Eun*
Bumil Informaition co., ltd, *Daegu Catholic University.

요 약

Text-to-SQL 은 자연어 질문을 실행 가능한 SQL 로 변환하여 데이터베이스 질의를 자동화하는 핵심 기술이다. 최근 모델 성능은 다양한 오픈 모델의 출현으로 크게 향상되었으나, 다국어 환경 또는 도메인 특화 환경에서는 데이터베이스 스키마의 정확한 해석이 어려워 성능 병목이 발생한다. 본 연구는 이러한 스키마 용어 해석 문제를 완화하기 위해, 스키마 요소와 다국어 동의어를 매핑한 구조화된 lexicon(dictionary)을 프롬프트에 직접 삽입하는 lexicon-assisted training 방법을 제안한다. 제안 방법은 자연어 질문과 데이터베이스 구조 간 의미 정렬을 강화하여 스키마 해석의 모호성을 줄이고, 결과적으로 질의 정확도를 개선하는 것을 목표로 한다. 4 개 백본 모델에 대해 lexicon 을 포함한 파인 튜닝을 수행하고, Spider 벤치마크에서 평가한 결과 lexicon-assisted 설정이 기존 대비 Execution Accuracy(EX)를 향상시키는 경향을 확인할 수 있다[1]. 본 방법은 새로운 스키마를 도입할 때 lexicon 만 갱신하면 되므로, 이식성과 유지보수 측면에서 높은 실용성을 지닌다.

Ⅰ. 서 론

자연어 기반 데이터 접근 수요가 증가하면서 사용자의 질문을 SQL 로 변환하는 Text-to-SQL 은 자연어 인터페이스의 핵심 구성요소로 자리 잡았다. 대표 벤치마크인 Spider 는 다중 테이블, 복잡 질의, 교차 도메인 일반화를 요구하며 Text-to-SQL 의 표준화된 형태로 평가할 수 있게 설계되었다[1]. 이후 스키마 인코딩과 링킹을 구조적으로 강화하는 연구가 발전했고, RAT-SQL 은 self-attention 을 통해 스키마 구조와 정렬 정보를 통합하여 처리해 성능을 크게 끌어올린 접근이다[2].

범용 생성 프레임워크의 발전도 Text-to-SQL 에 중요한 기반을 제공했다. T5 는 모든 NLP 과제를 text-to-text 형태로 통일하는 프레임워크를 제시하며 다양한 생성 과제에서 강력한 성능을 보여주었고[3], mT5 는 이를 다국어로 확장하여 101 개 언어의 사전학습으로 다국어 생성 성능을 뒷받침한다[4]. 이러한 흐름은 비영어권 질의에도 적용 가능성을 넓혔지만, 실제 서비스 환경에서의 성능 병목은 여전히 남아 있다.

다국어 환경 또는 전문 도메인에서는 스키마에 어휘(테이블/컬럼 명칭)를 정확히 해석하는 것이 어렵고, 스키마 요소의 잘못된 해석은 오류 SQL 생성으로 직결되어 모델의 신뢰성과 실용성을 크게 저해한다. 기존 접근은 대규모 주석 데이터 기반의 광범위 파인 튜닝 또는 대량 학습에 의존하는 경우가 많다.

본 연구는 정렬에 필요한 의미 단어를 명시적으로 제공하자는 관점에서, 테이블/컬럼과 한국어 표현의 대응관계를 사전으로 정리해 프롬프트에 포함시키는 학습 기법을 제안한다.

Ⅱ. 실험 및 결과

본 논문에서의 핵심은 테이블/컬럼과 한국어 동의어를 연결하는 구조화 lexicon 을 프롬프트에 삽입하여, 모델이 질문-스키마 의미 정렬을 보다 안정적으로 수행하도록 유도하는 것이다.

본 논문에서 사용된 입력 프롬프트는 [표 1]과 같다. 이 템플릿은 모델이 질문 표현을 스키마 요소로 연결할 때, 추측에 의존하기보다, 프롬프트에 포함된 정보를 사용하도록 유도한다.

Section	Example
Instruction	1) You are a Text-to-SQL model. 2) Refer to [schema] and [dictionary] to generate an exact SQL query. 3) Use the dictionary to match Korean terms to table/column names.
Schema	Customers (customer_id, payment_method, customer_name, ...) ...
Dictionary	Customers:["고객"], payment method:["결제 방식"], customer name:["고객 이름", "이름"], ...
Question	"결제 방식으로 현금을 사용하는 고객들의 이름을 반환하시오."
Target SQL	SELECT customer_name FROM Customers WHERE payment_method = "Cash";

[표 1]. Lexicon-assisted Text-to-SQL 프롬프트 예시

실험은 4 개의 백본 모델(KoBART-base-v2, KoGPT2, mT5-base, HyperCLOVAX-SEED-Text-Instruct-0.5B)을 대상으로 하며, 내부 한국어 질문 세트로 각 백본을 파인 튜닝한 뒤, multilingual Spider benchmark 와 동일 한국어 세트에서 평가했다. 평가 질의는 multi-table JOIN, GROUP BY 등 복잡 SQL 을 포함한다.

평가지표는 EM(Exact Match)과 EX(Execution Accuracy)를 사용한다. EM 은 생성 SQL 과 정답 SQL 의 문자열(또는 정규화 후) 일치 여부, EX 는 DB 실행 결과 일치 여부로 해석할 수 있으며, Spider 계열 연구에서 널리 사용된다[1].

[표 2]는 lexicon 삽입 유무에 따른 EM/EX 점수 결과를 보여준다. 전반적으로 lexicon 을 포함한 설정이 모든 백본에서 EM 과 EX 를 동시에 개선하였다.

Lexicon 삽입은 입력 토큰의 길이를 증가시킨다. 입력이 길어지면 컨텍스트 제약으로 인해 학습, 추론 난이도 및 비용이 가할 수 있지만, 그럼에도 성능이 개선된다는 점에서 새로운 스키마를 도입할 때 lexicon 만 갱신하면 되므로, 이식성과 유지보수 측면에서 실용적인 장점을 갖는다.

실험에 사용된 모든 모델의 매개변수 수가 10 억 개 미만의 소형 모델이고, 다국어 text-to-SQL 모델의 기본 정확도가 여전히 상대적으로 낮지만, 사전 기반의 미세 조정은 성능 향상을 보여준다.

Model	Non-Lexicon EM	Non-Lexicon EX	Lexicon EM	Lexicon EX
KoBART	0.00	0.00	0.03	0.05
KoGPT2	0.00	0.00	0.02	0.03
mT5-base	0.04	0.06	0.15	0.25
HyperCLOVAX	0.05	0.06	0.16	0.18

[표 2]. EM,EX 점수 비교

Ⅲ. 결론

본 연구는 다국어/전문 도메인 환경에서 Text-to-SQL 성능을 저해하는 핵심 요인 중 하나인 스키마 용어 해석의 모호성을 완화하기 위해, 테이블/컬럼과 다국어/도메인 동의어를 구조화한 lexicon 을 프롬프트에 삽입하는 lexicon-assisted training 을 제안했다. 4 개 소형 모델 비교에서, 제안 방식은 기준선 대비 EX 점수의 개선을 보였고, 특히 mT5-base 에서 가장 큰 향상이 관찰되었다. 또한 작은 규모 모델에서도 lexicon 정보 삽입이 다국어 SQL 생성의 모호성을 완화하고 성능 개선을 달성했으며, 더 큰 모델에 적용할 때 이득이 커질 것이라 예상한다.

ACKNOWLEDGMENT

본 결과물은 2025 년도 경상북도 지역혁신중심 대학지원체계(RISE)-(경상북도 K-IDEA Valley 프로젝트 현장실무형 고급인재양성)의 지원을 받아 수행된 결과입니다.

(2025-RISE-15-107)

참 고 문 헌

[1] Tao Yu et al., "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task," EMNLP 2018.

[2] Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, Matthew Richardson. "RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers." Proceedings of ACL 2020. 2020.

[3] Colin Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR 2020.

[4] Linting Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," NAACL 2021.