Journal of KIIT. Vol. 24, No. 7, pp. 45-60, Jul. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.7.45

대규모 도로명주소 검색 환경에서 자모 기반 전처리 전략의 성능 비교 분석

박상현*, 도성룡**

A Comparative Performance Analysis of Jamo Preprocessing Strategies for Large-Scale Road Name Address Search

Sanghyeon Park*, Sungryong Do**

요 약

본 연구는 전수 도로명주소 데이터베이스 환경에서 한글 주소 질의에 대한 자모 기반 전처리 전략의 효과를 통제된 조건에서 정량적으로 비교·분석하였다. 이를 위해 전국 도로명주소 6,414,486건으로 검색 DB를 구축하고, 대전 지역 정답 주소 112,572건 기반의 562,860건 질의셋을 구성하였다. 전처리 전략은 기준선(B0), 자모·초성 보강(B1), 자모 기반 유사도 재정렬(B2), 규칙 기반 띄어쓰기 보정(B3)으로 구분하였다. 검색은 SQLite FTS5 기반 2단계 구조로 수행하고, Coverage@k, Recall@k, MRR, 처리 시간으로 성능을 평가하였다. 실험 결과 Coverage@500은 84.84%였고, B1은 Recall@10 83.63%, MRR 0.8215로 가장 우수하였다. 추가 분석 결과 후보 검색 실패는 초성 축약 및 복합 자모 오류에 집중되었다. 따라서 자모 기반 전처리는 한글 주소 검색 성능 향상에 기여하나, 전체 성능은 후보 검색 단계 품질에 의해 구조적으로 제약됨을 확인하였다.

Abstract

This study quantitatively compared Jamo-based preprocessing strategies for Korean address queries under controlled conditions in a full road-name address database. A search DB was built with 6,414,486 national road-name address records, and 562,860 queries were generated from 112,572 ground-truth addresses in Daejeon. The strategies were baseline (B0), jamo/initial-consonant augmentation (B1), jamo-based reranking (B2), and rule-based spacing correction (B3). Search used a two-stage SQLite FTS5 framework and was evaluated by Coverage@k, Recall@k, MRR, and processing time. Coverage@500 was 84.84%, and B1 performed best with Recall@10 of 83.63% and MRR of 0.8215. Further analysis showed that candidate retrieval failures were concentrated in initial-consonant abbreviations and combined jamo errors. The results show that Jamo-based preprocessing improves Korean address search, but overall performance is structurally constrained by candidate retrieval quality.

Keywords

korean address search, road name address, jamo preprocessing, query noise, information retrieval

* 고려사이버대학교 융합정보대학원 석사과정
- ORCID: https://orcid.org/0009-0001-2802-3752
** 고려사이버대학교 컴퓨터공학부 조교수(교신저자)
- ORCID: https://orcid.org/0000-0003-3394-8220

· Received: Apr. 09, 2026, Revised: May 17, 2026, Accepted: May 20, 2026
· Corresponding Author: Sungryong Do
Division of Computer Engineering, Korea Cyber University
106 Bukchon-ro, Jongno-gu, Seoul, 03051, Korea
Tel: +82-2-6361-1918 E-mail: sungryongdo@koreacu.ac.kr

Ⅰ. 서 론

주소 검색은 사용자가 입력한 자연어 기반 주소 문자열을 구조화된 주소 및 좌표 정보로 변환하여 주문·배송, 경로 탐색, 민원 처리 등 다양한 후속 업무에 활용하도록 지원하는 핵심 인프라 기능이다. 특히 도로명주소 체계가 정착된 이후, 대규모 주소 데이터베이스 환경에서 사용자의 입력 질의를 빠르고 정확하게 정답 주소에 매핑하는 기능은 공공·민간 서비스를 막론하고 필수 요소가 되었다. 그러나 한글 도로명주소 검색은 초성 중심 축약 입력, 자모 단위 오타, 띄어쓰기 변형이 빈번하다는 점에서 일반 문자열 검색과 다른 구조적 난점을 가진다. 실무에서는 자모 분해, 초성 보강, 자모 기반 유사도 재정렬, 규칙 기반 띄어쓰기 보정 등의 전처리 전략이 활용되어 왔으나, 전수 주소 검색 환경에서 각 전략의 효과와 비용을 동일 조건에서 비교한 연구는 제한적이다.

본 연구는 전수 도로명주소 데이터베이스 환경에서 한글 주소 질의에 대한 자모 기반 전처리 전략의 효과를 통제된 조건에서 정량적으로 비교·분석하는 것을 목표로 한다. 이를 위해 전국 도로명주소 6,414,486건으로 검색 데이터베이스를 구축하고, 대전 지역 정답 주소 112,572건을 기반으로 원본 및 노이즈 질의를 포함한 562,860건의 질의셋을 구성하였다. 전처리 전략은 기준선(B0), 자모·초성 보강(B1), 자모 기반 유사도 재정렬(B2), 규칙 기반 띄어쓰기 보정(B3)으로 구분하였으며, SQLite Full-Text Search 5 (FTS5) 기반 2단계 검색 구조에서 Coverage@k, Recall@k, MRR(Mean Reciprocal Rank), 처리 시간 지표를 이용해 성능을 비교하였다.

본 연구는 다음 세 가지 연구 문제를 설정한다.

RQ1. 전처리 전략(B1~B3)은 기준선(B0)에 비해 주소 검색 정확도(Recall@k, MRR)를 개선하는가?

RQ2. 전처리 전략 간 성능 차이는 노이즈 유형에 따라 어떻게 달라지는가?

RQ3. 정확도 향상과 처리 시간 비용 간의 trade-off는 어떻게 나타나는가?

본 논문의 구성은 다음과 같다. Ⅱ장에서는 주소 검색 평가 지표와 한글 자모 구조, 전처리 설계 원칙을 정리한다. Ⅲ장에서는 관련 선행연구의 흐름과 한계, 본 연구의 차별성을 정리한다. Ⅳ장에서는 연구방법을, Ⅴ장에서는 실험 결과와 분석을 제시한다. 마지막으로 Ⅵ장에서는 결론과 시사점을 논의한다.

Ⅱ. 이론적 배경

2.1 순위 기반 정보검색 평가 지표

주소 검색 시스템은 하나의 정답만 반환하기보다 유사한 후보 주소를 순위 목록(Top-k) 형태로 제시하는 경우가 일반적이다. 사용자는 대개 상위 몇 개 후보만 확인하므로, 주소 매핑 품질은 단순 정오답 정확도만으로 충분히 설명되기 어렵고 “정답이 상위에 위치하는가”를 반영하는 순위 기반 지표가 필요하다. 본 연구는 도로명주소 검색을 Top-k 후보 추천 문제로 간주하고, 전처리 조건(B0~B3)에 따른 순위 품질을 Recall@k와 MRR로 비교한다.

Recall@k는 상위 k개 후보 안에 정답 주소가 포함되는 비율이다. Recall@1은 첫 번째 결과가 정답일 확률을 의미하며, Recall@3, Recall@5, Recall@10은 사용자가 제한된 결과 범위 내에서 정답을 찾을 수 있는지를 보여준다[1][2].

MRR은 각 질의에서 정답 순위의 역수(1/rank)를 평균한 지표로, 동일한 Recall@10을 보이더라도 정답이 상위권에 얼마나 집중되는지를 구분할 수 있다[1]-[4]. 따라서 본 연구는 Recall@k로 정답 포함 여부를, MRR로 정답의 상위 집중도를 함께 평가한다.

2.2 한글 자모 구조와 주소 입력 노이즈의 구조적 원인

한글은 초성·중성·종성의 조합으로 음절을 구성하는 자모 결합 문자 체계이다[5]. 따라서 주소 문자열을 음절 단위로만 비교할 경우 자모 수준에서 발생하는 입력 오류를 충분히 포착하기 어렵다. 이러한 구조적 특성은 주소 검색에서 반복적으로 나타나는 입력 노이즈의 원인을 설명하며, 자모 단위 전처리의 필요성을 뒷받침한다.

첫째, 사용자는 모바일·현장 입력 환경에서 초성 중심의 축약 입력을 자주 수행한다. 예를 들어 “대전로 730”을 “ㄷㅈㄹ 730”과 같이 입력하는 경우, 사용자는 동일 주소로 매핑되기를 기대한다[5]. 둘째, 받침 누락이나 인접 음절로의 이동, ㅐ/ㅔ·ㅗ/ㅓ와 같은 유사 모음 혼동, 자판 인접키 오타 등 다수의 오류는 자모 단위에서 발생하며, 이러한 차이는 음절 단위 비교보다 자모 단위 편집거리로 더 직접적으로 정량화할 수 있다[6][7]. 셋째, 주소 문자열의 띄어쓰기는 구조적 규칙성을 가지지만 실제 입력에서는 공백이 임의로 삭제·삽입·이동되며, 특히 숫자와 도로명 접미사 주변에서 변형이 집중되는 경향이 있다[7][8]. 본 연구는 이러한 특성을 반영하여 초성 입력, 받침 오류, 유사 모음 치환, 띄어쓰기 변형, 단순 오탈자, 부분 입력 등을 노이즈 유형으로 정의하고, 동일한 질의 집합에서 전처리 조건(B0~B3)의 효과를 비교한다. 따라서 이러한 구조적 특성은 자모 기반 전처리가 주소 검색 성능에 실질적으로 기여할 가능성을 시사한다.

2.3 전처리의 개념과 설계 원칙

전처리는 입력 질의와 후보 주소를 검색 친화적으로 정규화·변환하거나, 1차 검색 결과의 후보 순위를 재조정하는 방식으로 구현된다. 본 연구에서 B1과 B3는 주로 입력·표현 전처리에, B2는 후처리·재랭킹에 해당한다.

전처리는 일반적으로 검색 품질 향상을 위한 전략이지만, 그 효과는 기준선 검색기의 강건성, 노이즈 분포, 운영 제약에 따라 달라진다. 자모 분해 및 초성 보강(B1)은 초성 입력과 자모 오류가 포함된 질의에서 매칭 기회를 넓힐 수 있으나, 일반 질의에서는 표현 변화로 인해 유사도 계산 결과에 영향을 줄 수 있다[6]. 편집거리 기반 재정렬(B2)은 자모 수준 차이를 정량화할 수 있지만, 추가 계산 비용을 수반하며 일부 경우에는 순위 불안정성을 유발할 수 있다[3][4]. 규칙 기반 띄어쓰기 보정(B3)은 공백 변형에 대응할 수 있으나, 예외적 표기를 충분히 반영하지 못하면 평균 기여가 제한될 수 있다[7][8]. 또한 전처리는 문자열 변환과 재랭킹 연산을 통해 평균 응답시간과 p95 응답시간을 증가시킬 수 있으며, 특히 B2는 SLA(Service Level Agreement) 관점의 부담이 크다[4][7].

요약하면 전처리는 보편적 개선책이라기보다, 노이즈 유형 분포와 기준선의 강건성, 지연 제약에 따라 조건부로 유효성이 달라지는 전략이다. 본 연구는 동일한 데이터셋과 노이즈 질의 집합을 고정한 통제 실험을 통해 B0~B3가 Recall@k, MRR, 평균 처리시간 및 p95 처리 시간에 미치는 영향을 함께 측정하고, 노이즈 유형별·운영 제약별 선택 기준을 도출하고자 한다.

III. 선행연구

3.1 선행연구의 흐름

도로명주소 및 유사 문자열 매핑과 관련된 선행연구는 크게 세 가지 흐름으로 정리할 수 있다.

첫째, 주소 정제·파싱 및 연계 연구는 비정형 주소 문자열을 표준화된 주소 구성 요소로 분해하거나, 문자열 유사도 기반으로 정제하는 문제를 다루었다. 편집거리 기반 문자열 유사도 알고리즘을 활용하여 비정제 주소를 표준 주소와 매칭하고 입력 오류를 정량적으로 반영하는 방법이 제안되었으며[6], 실제 시스템 환경에서 주소 정제를 통해 데이터 품질을 개선하는 접근도 이루어졌다[7]. 또한 학습 데이터 자동 구축을 기반으로 주소를 구성 요소 단위로 분해하는 파싱 모델이 제안되었고[8], 딥러닝 기반 주소 처리 미들웨어를 통해 행정데이터 간 연계를 지원하는 연구도 수행되었다[9]. 이러한 연구들은 주소를 구조화 가능한 데이터로 처리하는 데 기여하였으나, 검색 결과에서 정답의 순위 위치를 직접적으로 평가하는 접근은 제한적이었다.

둘째, 지오코딩·검증 및 품질관리 연구는 주소를 좌표와 연결하거나 데이터 품질을 유지하는 운영 관점의 문제를 다루었다. 도로명주소 기반 지오코딩 및 역지오코딩 기법을 통해 주소와 공간정보 간 연계를 강화하는 연구가 수행되었으며[10], 음성 입력 환경에서 발생하는 오류를 보정하여 주소 인식 정확도를 향상시키는 방법도 제안되었다[11]. 또한 글

로벌 주소 데이터의 품질을 평가하고 개선하는 프레임워크가 제시되었고[12], 주소 레이어 구축 자동화를 통해 대규모 환경에서의 운영 효율성을 향상시키는 연구도 이루어졌다[13]. 이러한 연구들은 주소 데이터의 활용성과 신뢰성을 높이는 데 기여하였으나, 입력 노이즈 유형별 검색 성능 차이나 전처리 전략의 순위 기반 기여를 직접적으로 비교하는 데에는 한계가 있었다.

셋째, 정보검색 기반 노이즈 처리 및 재순위화 연구는 질의 확장, 오타 보정, 재정렬 기법을 통해 검색 품질을 개선하는 데 초점을 두었다. 질의 확장을 통해 검색 재현율을 향상시키는 방법이 제안되었으며[14], 오타 노이즈를 포함한 질의 이해 성능을 개선하는 모델이 연구되었다[15]. 또한 dense retrieval 기반 검색 환경에서 재정렬의 효과를 분석하거나[4], 후보 재정렬 단계가 최종 검색 성능에 중요한 영향을 미친다는 점을 입증한 연구도 수행되었다[3]. 이러한 연구들은 질의 보정 및 재랭킹의 중요성을 강조하였으나, 자모 기반 전처리와 규칙 기반 전처리 전략을 동일한 주소 검색 환경에서 통제된 조건으로 비교하는 연구는 제한적이었다. 특히 전수 주소 데이터 환경에서 노이즈 유형별 성능과 처리 시간 비용을 함께 고려한 통합적 분석 사례는 부족한 실정이다.

3.2 선행연구의 한계와 개선 과제

선행연구를 종합하면 주소 검색·정제·매핑 관련 연구는 다양한 기법을 제안해 왔으나, 전처리 전략의 효과를 실무 의사결정에 활용 가능한 형태로 제시하는 데에는 몇 가지 공통 한계를 가진다. 첫째, 순위 기반 평가의 비중이 상대적으로 낮다. 많은 연구가 정제 성공 여부, 파싱 정확도, 분류 정확도 중심으로 성능을 제시하여, 실제 검색 환경에서 중요한 정답의 상위 노출 정도를 충분히 설명하지 못한다. 둘째, 전처리 기여분의 단계적 분해가 부족하다. 특정 기법 도입 전후를 비교하는 방식이 많아, 개별 전처리 요소가 어떤 방식으로 기여하는지를 독립적으로 해석하기 어렵다. 셋째, 오류 유형별 분석이 제한적이다. 초성 입력, 받침 오류, 유사 모음 치환, 띄어쓰기 변형, 단순 오탈자, 부분 입력 등은 서로 다른 표기 변형을 유발하지만, 전체 평균 중심의 결과 제시만으로는 어떤 오류군에 어떤 전략이 강한지 설명하기 어렵다. 넷째, 운영 비용 고려가 미흡하다. 전처리는 추가 연산을 동반하므로 응답시간 증가와 지연 악화를 유발할 수 있음에도, 다수 연구는 품질 개선을 강조하는 반면 비용은 부수적으로만 다루는 경향이 있다. 다섯째, 노이즈 생성 규칙, 평가셋 구성, 실험 파이프라인이 충분히 공개되지 않아 재현성이 부족한 경우가 많다. 따라서 본 연구는 전수 DB 환경, 순위 기반 지표, 노이즈 유형별 분석, 처리 시간, 재현 가능한 평가 구조까지 통합적으로 고려한 실험 설계를 통해 전처리 전략의 실질적 기여를 정량적으로 분해하여 제시한다. 표 1은 기존 주소 검색·정제 관련 연구의 주요 한계와 이에 대한 본 연구의 대응 방안을 정리한 것이다.

표 1. 기존 연구의 한계와 본 연구의 대응
Table 1. Limitations of previous studies and how this study addresses them

Category	Limitations of existing studies	Proposed approach in this study
Evaluation data	Focus on limited datasets	Use of full national road name address dataset
Query type	Clean inputs or limited noise	Systematic generation of noisy queries (initial consonants, typos, spacing variations, etc.)
Comparison method	Before/after comparison of a single method	Stepwise comparison of preprocessing strategies (B0 - B3)
Evaluation method	Lack of ranking-based evaluation	Ranking-based evaluation using Recall@k and MRR
Error analysis	Limited analysis by error type	Performance analysis by noise type
Cost analysis	Limited consideration of computational cost	Inclusion of average and p95 processing time
Reproducib-ility	Limited reproducibility	Provision of noise query CSV and experimental pipeline

표 1에서 확인할 수 있듯이, 본 연구는 기존 연구의 한계를 보완하기 위해 전수 DB 환경, 순위 기반 평가, 오류 유형별 분석, 처리 시간 측정을 통합적으로 반영하였다.

3.3 차별성과 기여

본 연구는 이러한 한계를 보완하기 위하여 전수 도로명주소 데이터베이스 기반의 재현 가능한 평가 환경을 구축하고, 동일한 질의셋과 동일한 후보 검색 구조 아래 전처리 조건 B0~B3를 체계적으로 비교하였다. 특히 주소 검색을 Top-k 후보 추천 문제로 정식화하고, Recall@k와 MRR을 핵심 품질 지표로 사용함으로써 정답이 상위에 위치하는 정도를 직접 측정하였다. 또한 초성, 오타, 띄어쓰기 변형 등을 포함한 노이즈 질의를 체계적으로 생성하고, 노이즈 유형별 성능을 분해하여 전처리 전략의 효과를 비교하였다. 더불어 평균 처리 시간과 p95 처리 시간을 함께 평가함으로써 정확도와 비용 간의 trade-off를 실증적으로 검토하였다. 이러한 점에서 본 연구는 전처리 전략의 기여와 한계를 단순 평균 성능 비교를 넘어 운영 의사결정에 활용 가능한 형태로 제시한다는 차별성을 가진다.

Ⅳ. 연구방법

4.1 연구 설계 개요

본 연구는 전수 도로명주소 데이터베이스 기반 검색 환경에서 한글 주소 질의에 대한 전처리 전략의 효과를 비교·분석하기 위하여 데이터 구축, 전처리 설계, 검색 실행, 평가 및 분석의 네 단계로 연구 절차를 설계하였다. 먼저 전국 도로명주소 원본 데이터를 기반으로 검색 가능한 주소 데이터베이스를 구축하고, 평가용 정답 주소를 선정한 뒤 다양한 입력 오류를 반영한 노이즈 질의를 생성하였다. 다음으로 기준선(B0), 자모·초성 보강(B1), 자모 기반 유사도 재정렬(B2), 규칙 기반 띄어쓰기 보정(B3)의 네 가지 전처리 조건을 정의하였다. 이후 SQLite FTS5 기반 1차 후보 검색과 동일 후보 집합에 대한 2차 재정렬 구조를 적용하여 각 전처리 조건의 성능을 비교하였다. 마지막으로 Coverage@k, Recall@k, MRR 및 처리 시간 지표를 이용하여 전체 성능과 노이즈 유형별 차이를 분석하였다. 이러한 절차는 실제 주소 검색 흐름을 반영하면서도, 전처리 전략 간 차이를 동일 조건에서 비교할 수 있도록 설계하였다. 또한 입력 질의 생성부터 검색 실행, 결과 집계까지를 하나의 실험 파이프라인으로 구성하여 조건 간 비교 가능성을 높이고자 하였다. 그림 1은 본 연구의 전체 실험 절차와 분석 흐름을 나타낸 것이다.

본 연구의 설계는 후보 검색 단계와 재정렬 단계를 구분하여 해석할 수 있도록 구성되었다는 점에서 특징을 가진다. 즉 동일한 검색 대상 데이터베이스와 동일한 질의 집합을 고정한 상태에서 전처리 전략만을 달리함으로써, 전처리 방식 자체가 검색 품질과 처리 시간에 미치는 영향을 보다 명확하게 비교하고자 하였다.

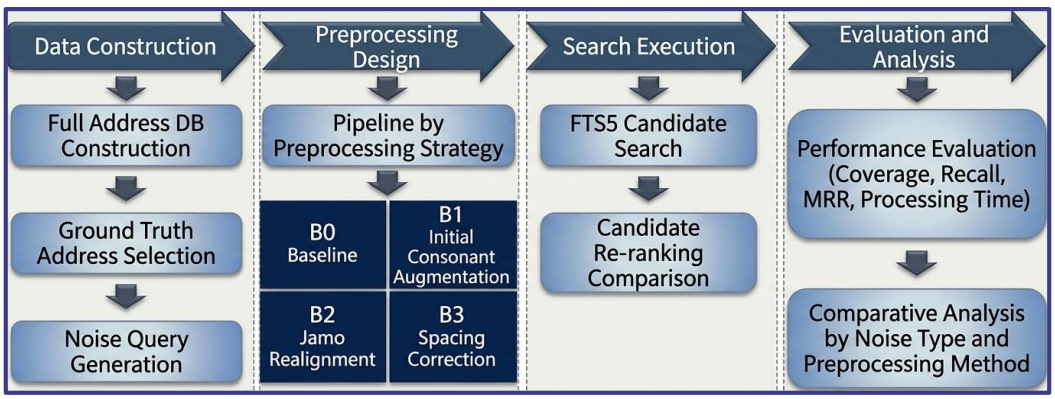

그림 1. 연구 설계 개요
Fig. 1. Overview of the research design

또한 전체 평균 성능뿐 아니라 노이즈 유형별 성능과 후보 검색 단계의 구조적 제약까지 함께 고려함으로써, 실제 서비스 환경에서 적용 가능한 전처리 전략 선택 기준을 도출하고자 하였다.

4.2 데이터 구축

본 연구에서는 주소정보누리집 및 주소기반산업지원서비스에서 제공하는 전국 도로명주소 다운로드 자료를 기반으로 전수 주소 데이터베이스를 구축하였다. 원본 데이터는 CP949 인코딩 형식의 TXT 파일이며, 이를 파싱·정제한 뒤 SQLite 데이터베이스에 적재하고 FTS5 인덱스를 구축하였다. 최종적으로 구축된 전국 전수 주소 데이터베이스의 규모는 6,414,486건이다.

평가용 정답 주소는 대전 지역 도로명주소 112,572건으로 구성하였으며, 각 정답 주소에 대해 원본 질의 1개와 노이즈 질의 4개를 생성하여 총 562,860건의 질의셋을 구축하였다. 질의셋은 CSV 형식으로 저장하였고, 각 레코드에는 질의 식별자, 정답 주소 식별자, 노이즈 유형, 질의 문자열을 포함하였다. 대전 지역은 도심·주거·상업 기능이 혼재하여 다양한 주소 형태를 포함하고, 실험 데이터 규모와 반복 가능성을 함께 확보할 수 있어 평가 대상으로 선정하였다.

다만 평가용 정답셋과 질의셋이 대전 지역 주소에 기반하므로, 본 연구 결과를 전국 단위의 주소 패턴 전반으로 일반화하는 데에는 한계가 있다. 전국 도로명주소 DB를 후보 검색 대상으로 사용하였으나, 실제 평가는 대전 지역 질의셋을 중심으로 수행되었기 때문에 지역별 도로명 구성, 건물번호 분포, 도시·농어촌 주소 특성 차이를 모두 반영한다고 보기는 어렵다. 따라서 향후 다양한 권역의 평가셋을 활용한 추가 검증이 필요하다.

이와 같은 데이터 구성은 정답 주소와 질의 간 대응 관계를 일관되게 유지하면서 전처리 전략별 검색 결과를 동일 기준에서 비교하기 위한 것이다. 또한 전체 평균 성능뿐 아니라 노이즈 유형별 성능 차이를 함께 분석할 수 있도록 설계하였다. 표 2는 본 연구에서 사용한 데이터셋의 구성과 규모를 나타낸 것이다.

표 2. 데이터셋 구성
Table 2. Dataset configuration

Category	Description
Data source	Juso.go.kr / Address-based Industry Support Service (road name address download data)
Full address database	6,414,486 records
Ground truth set	Road name addresses in Daejeon (112,572 records)
Original data format	TXT (CP949)
Search index	SQLite FTS5
Query file format	CSV (UTF-8-SIG)
Query fields	query_id, addr_id, noise_type, query
Total queries	562,860

4.3 전처리 설계

모든 실험 조건에는 공통적으로 공백 정리, 표기 통일, 특수문자 최소화 등의 기본 정규화를 적용하였다. 이를 바탕으로 네 가지 전처리 조건을 정의하였다. B0는 추가 전처리 없이 기본 정규화만 적용한 기준선이다. B1은 한글 음절을 자모 단위 시퀀스로 분해하고 초성 표현을 함께 반영하는 자모·초성 보강 조건이다. B2는 B1 기반 표현을 사용하되, 1차 검색 결과의 상위 후보에 대해 자모 단위 Levenshtein 편집거리를 계산하여 순위를 재정렬하는 방식이다. B3는 도로명 접미사와 숫자 패턴을 이용하여 공백을 삽입하거나 재배치하는 규칙 기반 띄어쓰기 보정 조건이다.

B1의 자모·초성 보강은 기본 정규화가 적용된 주소 문자열을 대상으로 한글 음절을 초성·중성·종성 단위로 분해한 뒤, 음절 순서를 유지한 선형 자모 시퀀스로 변환하는 방식으로 수행하였다. 종성이 없는 음절은 초성과 중성만 포함하고, 종성이 있는 음절은 초성·중성·종성을 모두 포함하였다. 또한 초성 축약 입력에 대응하기 위해 각 한글 음절의 초성만 추출한 초성 시퀀스를 별도로 생성하였다. 최종 B1 표현은 자모 시퀀스와 초성 시퀀스를 함께 결합한 형태로 구성하였으며, 본 연구에서는 자모 분해 결과를 별도의 n-gram 토큰으로 재구성하지 않았다.

B2는 B1 기반 표현을 사용하되, 상위 후보에 대해 자모 기반 Levenshtein 편집거리를 계산하여 순위

를 재정렬하는 방식이다. 이때 편집거리 계산에는 B1 표현 중 초성 보강 부분을 제외한 자모 시퀀스만 사용하였다. 본 연구의 Levenshtein 거리는 삽입·삭제·치환 비용을 모두 1로 설정한 단순 문자열 편집거리이며, 초성·중성·종성 등 자모 유형별 가중치는 별도로 부여하지 않았다. 따라서 B2는 자모 문자열 간 최소 편집 횟수를 기준으로 질의와 후보 주소의 유사도를 보정하는 방식으로 해석할 수 있다.

또한 실제 입력 오류를 반영하기 위해 초성 입력, 받침 변형, 유사 모음 치환, 띄어쓰기 변형, 단순 오탈자, 부분 입력 등을 중심으로 노이즈 질의를 생성하였고, 일부 질의에는 두 가지 이상의 오류를 결합하여 복합 입력 상황도 반영하였다. 표 3은 본 연구에서 노이즈 질의를 생성하기 위해 적용한 주요 규칙과 유형을 정리한 것이다.

표 3. 노이즈 질의 생성 규칙 예시
Table 3. Examples of noise query generation rules

Category	Type	Description
Jamo-based errors	chosung	Keep only initial consonants
	batchim	Omit final consonant or shift it to an adjacent syllable
	vowel	Substitute with similar vowels
	spacing	Insert, delete, or shift spaces
Combined jamo errors	batchim +chosung	Combination of final consonant modification and initial consonant extraction
	chosung +spacing	Combination of initial consonant extraction and spacing variation
	batchim +spacing	Combination of final consonant modification and spacing variation
	batchim +vowel	Combination of final consonant and vowel modification
	vowel +spacing	Combination of vowel modification and spacing variation
Simple string errors	typo_delete	Random character deletion
	typo_replace	Random character substitution
	remove_spaces	Remove all spaces
	keep_tail	Keep only the tail part of the string (drop the prefix)

다만 본 연구의 노이즈 질의는 실제 서비스 로그에서 관찰되는 전체 오류 분포를 그대로 재현하기 위한 것이 아니라, 한글 주소 검색에서 자모 기반 오류가 포함된 조건을 통제적으로 비교하기 위해 설계된 실험용 질의셋이다. 이에 따라 초성 입력, 받침 변형, 유사 모음 치환 등 자모 구조와 관련된 오류가 상대적으로 높은 비중으로 포함되어 있다. 이러한 설계는 자모 기반 전처리 전략의 효과를 분석하는 데 적합하지만, 특정 전처리 전략에 유리한 평가 환경으로 해석될 가능성도 가진다. 따라서 본 연구에서는 B1의 성능 향상을 자모 기반 전처리의 보편적 우월성으로 해석하지 않고, 자모 오류가 포함된 통제 조건에서의 조건부 효과로 해석하였다.

각 전처리 전략은 동일 질의 입력에 대해 결정론적으로 적용되며, 실험 재현성을 위해 모든 변환 규칙을 명시적으로 정의하였다.

4.4 검색 실행

검색은 SQLite FTS5 기반 2단계 구조로 수행하였다. 본 연구에서 FTS5를 사용한 이유는 Elasticsearch와 같은 운영형 전문 검색엔진의 최종 성능을 검증하기보다, 동일한 데이터셋과 동일한 후보 검색 조건에서 전처리 전략 B0~B3의 상대적 효과를 통제된 환경에서 비교하기 위함이다. 전문 검색엔진은 분석기 설정, 사용자 사전 구성, 분산 인덱스 구성, 순위화 방식 조정 등 다양한 요인이 검색 결과에 영향을 줄 수 있으므로, 전처리 전략 자체의 효과를 분리해 해석하기 어렵게 만들 수 있다. 이에 본 연구는 단일 환경에서 재현 가능하고 후보 검색 조건을 일정하게 유지하기 쉬운 SQLite FTS5를 사용하였다.

다만 이러한 선택은 실험 통제와 재현성 측면의 장점이 있는 반면, 운영 검색엔진 수준의 검색 기능을 충분히 반영하지 못한다는 한계도 가진다. SQLite FTS5는 전문 검색엔진에 비해 한국어 형태소 분석, 동의어 처리, 고급 순위화 조정, 분산 확장성 측면에서 제약이 있으며, 초성 입력, 자모 오류, 행정구역명 변형 등을 정교하게 처리하는 전용 분석기를 기본적으로 제공하지 않는다. 따라서 본 연

구 결과를 운영 검색엔진 환경의 최종 성능으로 직접 일반화하기에는 한계가 있다.

1단계에서는 각 질의에 대해 전수 주소 데이터베이스에서 상위 500개 후보를 검색하였다. 2단계에서는 동일한 후보 집합을 대상으로 전처리 전략별 순위 비교를 수행하였다. 이 구조는 후보 검색 단계와 재정렬 단계의 영향을 구분하여 해석하기 위해 설계하였다. 전처리 전략에 따른 성능 차이가 후보 집합의 차이가 아니라 전처리 방식 자체에서 비롯되도록 통제하였다. 또한 대규모 질의셋 실험을 위해 검색 실행과 결과 기록을 일괄 파이프라인 형태로 구성하였다.

4.5 평가 및 분석 방법

본 연구에서는 후보 검색 단계와 최종 순위 단계를 구분하여 평가를 수행하였다. 후보 검색 단계의 품질은 Coverage@k로 측정하였으며, 이는 상위 k개 후보 안에 정답 주소가 포함되는 비율을 의미한다. 최종 검색 성능은 Precision@1, Recall@1, Recall@3, Recall@5, Recall@10 및 MRR로 평가하였다. Recall@k는 상위 k개 결과 안에 정답이 포함되는 비율이며, MRR은 정답의 실제 순위를 반영하는 지표이다. 또한 실제 주소 검색 시스템에서는 잘못된 주소를 최상위 결과로 추천하는 오탐을 줄이는 것도 중요하므로, Precision@1을 추가로 산출하였다. 본 연구의 질의셋은 하나의 정답 주소에서 원본 질의와 노이즈 질의를 포함한 복수의 질의가 생성되는 구조이지만, 평가 단위는 개별 질의이며 각 질의는 하나의 정답 주소 식별자에 대응된다. 따라서 개별 질의 기준으로는 단일 정답 검색 문제로 정의되며, Precision@1은 첫 번째 추천 결과가 해당 질의의 정답 주소와 일치하는 비율로 계산된다. 이 구조에서 Precision@1은 Recall@1과 동일한 값을 갖지만, 해석 관점에서는 실제 시스템이 사용자에게 최상위로 제시하는 주소의 오탐 가능성을 평가하는 지표로 활용할 수 있다.

마지막으로 각 질의에 대한 처리 시간을 기록하여 평균 처리 시간과 p95 처리 시간을 산출함으로써 정확도와 비용을 함께 분석하였다. 질의 생성 시 부여한 노이즈 유형 정보를 기준으로 전처리 조건별 성능을 유형별로 집계하여, 어떤 오류 환경에서 어떤 전략이 상대적으로 더 효과적인지 비교하였다.

V. 연구결과 및 분석

5.1 전체 성능 비교

전처리 전략별 전체 성능 비교 결과는 표 4와 같다. 먼저 1차 후보 검색 단계의 Coverage@500은 84.84%로 나타났다. 이는 전체 질의 중 약 84.84%에서 정답 주소가 상위 500개 후보 안에 포함되었음을 의미하며, 이후 재정렬 단계가 도달할 수 있는 성능 상한을 보여준다. 다시 말해 나머지 질의는 후보 검색 단계에서 이미 정답이 누락된 경우에 해당하므로, 후속 전처리 전략만으로는 복구가 어렵다. 표 4는 전처리 전략 B0~B3의 전체 검색 성능을 Precision@1, Recall@k, MRR 기준으로 비교한 결과이다.

표 4에서 확인할 수 있듯이, B1은 Precision@1, Recall@k, MRR 모두에서 가장 우수한 성능을 보였다. B1의 Precision@1은 0.8136으로 기준선(B0)의 0.7896보다 높게 나타났으며, 이는 자모·초성 보강 기반 전처리가 정답 주소를 상위 후보군에 포함시키는 것뿐 아니라 최상위 추천 결과의 정확성 측면에서도 개선 효과를 보였음을 의미한다.

표 4. 전처리 전략별 전체 성능 비교
Table 4. Overall performance comparison by preprocessing strategy

Method	Precision@1	Recall@1	Recall@3	Recall@5	Recall@10	MRR
B0 (Baseline)	0.7896	0.7896	0.8067	0.8129	0.8197	0.8002
B1	0.8136	0.8136	0.8269	0.8315	0.8363	0.8215
B2	0.7908	0.7908	0.8142	0.8228	0.8312	0.8049
B3	0.7878	0.7878	0.8056	0.8120	0.8189	0.7988

본 연구의 단일 정답 평가 구조에서 Precision@1은 Recall@1과 동일한 값을 갖지만, 실제 서비스 해석에서는 잘못된 주소를 1위로 제시하는 오탐 가능성을 줄이는 지표로 볼 수 있다. B2는 B0 대비 일부 성능 향상을 보였으나 전반적으로 B1보다 낮은 수준에 머물렀고, B3는 대부분의 지표에서 기준선과 유사한 수준을 보였다.

특히 B1의 우수한 결과는 한글 주소 입력에서 빈번하게 발생하는 초성 입력과 자모 단위 오류에 대한 대응력이 전체 성능 향상으로 이어졌음을 시사한다. 반면 B2는 자모 기반 유사도 재정렬을 통해 일부 질의에서 정밀한 순위 조정이 가능하였으나, 전체 평균 수준에서는 B1을 상회하지 못하였다. 또한 B3는 띄어쓰기 변형에 대응하기 위한 규칙 기반 전략이지만, 전체 검색 환경에서는 평균적인 기여가 제한적이었다. 따라서 전체 성능 비교만 놓고 보면, 본 실험 환경에서 가장 실용적인 전처리 전략은 B1로 해석된다.

정확도 비교와 함께 처리 시간 결과를 보면, 전처리 전략 간 비용 차이도 분명하게 나타난다. 처리 시간 비교 결과, B2는 자모 기반 유사도 계산과 재정렬 과정이 추가되면서 평균 처리 시간과 p95 처리 시간 모두에서 가장 큰 부담을 보였다. 반면 B1은 정확도 향상 폭이 가장 크면서도 B2보다 처리 시간 증가가 작아, 품질-비용 균형 측면에서 가장 현실적인 대안으로 해석된다. B3는 처리 시간 부담이 크지 않지만 평균 성능 개선도 제한적이었다. 표 5는 이러한 전처리 전략별 평균 처리 시간과 p95 처리 시간을 비교한 결과이다.

표 5. 전처리 전략별 처리 시간 비교
Table 5. Processing time comparison by preprocessing strategy

Method	Avg. total processing time (ms)	p95 total processing time (ms)
B0 (Baseline)	158.05	376.28
B1	191.10	416.47
B2	224.43	460.35
B3	159.24	377.06

한편, 기준선(B0)과 각 전처리 전략(B1, B2, B3) 간 질의 단위 paired t-test를 수행한 결과, B1과 B2는 주요 정확도 지표에서 B0 대비 유의한 향상을 보였으며, 특히 B1이 가장 큰 평균 개선폭을 나타냈다. 반면 B3는 통계적으로 유의한 변화가 관찰되었으나 평균 차이의 방향은 전반적으로 음(-)으로 나타났다. 또한 처리 시간 검정 결과에서도 B1과 B2는 모두 기준선 대비 유의한 증가를 보였고, 특히 B2의 증가폭이 가장 컸다. 표 6은 이러한 paired t-test 결과를 요약한 것이다.

표 6에서 확인할 수 있듯이, B1은 정확도 향상 폭과 비용 증가를 함께 고려할 때 가장 균형 잡힌 전략으로 해석된다. 반면 B2는 일정 수준의 성능 향상을 보였으나 처리 시간 증가 폭이 커 운영 효율성 측면의 부담이 크다. B3는 처리 시간 증가가 크지 않지만 평균적인 정확도 개선 효과도 제한적이어서 범용 전략으로 보기는 어렵다.

표 6. 기준선(B0) 대비 전처리 전략의 paired t-test 결과
Table 6. Summary of paired t-test results for preprocessing strategies compared with baseline (B0)

Comparison	ΔRecall @10	ΔMRR	ΔTime (ms)	p-value
B1 vs. B0	+0.0166	+0.0213	+26.62	p<.001
B2 vs. B0	+0.0115	+0.0046	+53.60	p<.001
B3 vs. B0	-0.0008	-0.0014	+0.93	p<.001

종합하면, 자모·초성 보강 기반 전처리(B1)는 본 실험 환경에서 가장 안정적으로 성능을 향상시키는 전략으로 확인되었다. 유사도 기반 재정렬(B2)은 일정 수준의 성능 향상을 보였으나 추가적인 계산 비용에 비해 이득이 상대적으로 작았고, 규칙 기반 띄어쓰기 보정(B3)은 전체적으로 뚜렷한 개선 효과를 보이지 않았다. 또한 Coverage@500이 84.84% 수준으로 측정되었다는 점에서, 전체 성능은 전처리 전략뿐 아니라 후보 검색 단계의 품질에 의해서도 구조적으로 제한된다는 점을 함께 확인할 수 있다. 결과적으로 B1은 정확도 향상과 비용 증가 간의 trade-off 측면에서 가장 우수한 Pareto 효율성을 보이는 전략으로 해석된다.

5.2 후보 검색 단계 분석

후보 검색 단계 분석 결과, Coverage@500은 0.8484로 나타났다. 이는 전체 질의 중 약 84.84%에서 정답 주소가 상위 500개 후보 내에 포함되었음을 의미한다. 반대로 약 15.16%의 질의는 후보 검색 단계에서 이미 정답 주소가 누락된 경우에 해당한다. 이러한 질의는 이후 재정렬 단계에서 정답을 복구할 수 없으므로, 사실상 전체 검색 성능의 상한을 제한하는 사례라고 볼 수 있다.

Coverage@500에서 정답 주소가 후보군에 포함되지 않은 85,357건(15.16%)의 질의를 추가로 분석한 결과, 후보 검색 실패는 단순 공백 변형이나 일반 오탈자보다 초성 축약 및 복합 자모 오류에서 집중적으로 발생하였다. 표 7은 Coverage@500에서 정답 주소가 후보군에 포함되지 않은 질의를 주요 실패 유형별로 정리한 것이다.

표 7. Coverage@500 실패 질의의 주요 유형
Table 7. Major types of candidate retrieval failures at Coverage@500

Failure type	Total queries	Failed queries	Failure rate (%)
Initial-consonant combined jamo errors	56,091	38,752	69.09
Initial-consonant abbreviation	28,326	19,414	68.54
Final-consonant + spacing errors	18,451	6,282	34.05
Vowel + spacing errors	18,808	5,531	29.41
Final-consonant + vowel errors	19,031	3,696	19.42
Partial input	56,091	9,302	16.58

표 7에서 확인할 수 있듯이, 후보 검색 실패는 초성 관련 오류에서 가장 두드러졌다.

Initial-consonant combined jamo errors는 초성 축약에 받침, 모음, 띄어쓰기 변형이 결합된 유형으로, 56,091건 중 38,752건이 실패하여 69.09%의 실패율을 보였다. Initial-consonant abbreviation 역시 28,326건 중 19,414건이 실패하여 68.54%의 높은 실패율을 보였다. 이는 초성 중심 입력이 주소 문자열의 어휘 정보를 크게 축소하여 SQLite FTS5 기반 후보 검색 단계에서 정답 주소를 Top-500 후보군에 포함시키기 어렵게 만든 결과로 해석된다.

반면 Final-consonant + spacing errors, Vowel + spacing errors, Final-consonant + vowel errors의 실패율은 각각 34.05%, 29.41%, 19.42%로 나타났으며, Partial input의 실패율은 16.58%였다. 또한 표에 제시하지 않은 받침 변형 단독 오류와 유사 모음 변형 단독 오류의 실패율은 각각 5.36%, 2.98%로 상대적으로 낮았고, 원본 질의, 공백 제거, 띄어쓰기 변형, 문자 치환 오타 유형에서는 Coverage@500 실패가 발생하지 않았다. 따라서 Coverage@500의 15.16% 실패는 무작위적으로 발생한 것이 아니라, 주로 초성 중심 축약 및 복합 자모 오류와 같이 후보 검색 단계에서 충분한 검색 단서를 제공하지 못하는 질의 유형에 집중되어 있었다.

이 결과는 최종 성능을 해석하는 데 중요한 기준을 제공한다. 가장 우수한 성능을 보인 B1의 Recall@10은 0.8363으로, Coverage@500의 값인 0.8484에 상당히 근접하였다. 이는 현재 검색 구조에서 자모·초성 보강 기반 전처리(B1)가 재정렬 단계에서 도달 가능한 성능 상한에 가까운 수준까지 정답 순위를 끌어올렸음을 의미한다. 또한 Coverage@k를 함께 측정함으로써, 최종 Recall@k와 MRR만으로는 구분하기 어려운 후보 검색 단계의 누락 문제와 재정렬 전략의 효과를 분리해 해석할 수 있다. 따라서 전처리 전략의 정교화만으로는 전체 성능 향상에 구조적 한계가 있으며, 1단계 후보 검색 품질이 전체 검색 성능을 결정하는 중요한 요소임을 확인할 수 있다.

후보 검색 단계의 한계를 완화하기 위해서는 1차 후보 검색 품질을 개선하는 접근이 필요하다. 구체적으로 초성 입력에 대응하기 위한 초성 전용 인덱스, 자모 분해 표현을 포함한 확장 인덱스, 도로명·행정구역명·건물번호 등 주소 구성 요소별 가중치를 반영한 필드 기반 검색 구조를 검토할 수 있다.

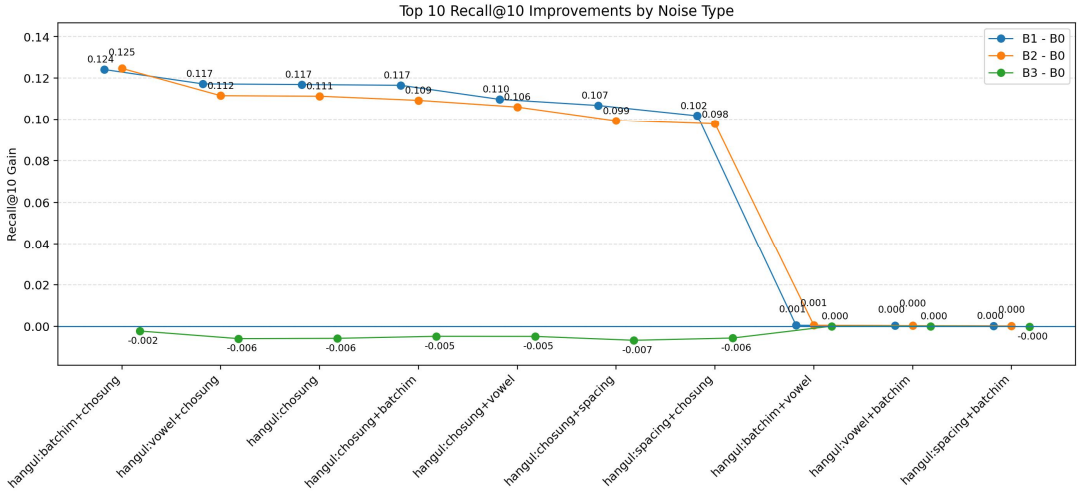

그림 2. 노이즈 유형별 Recall@10 개선폭 비교(B1-B0, B2-B0, B3-B0)
Fig. 2. Comparison of Recall@10 improvements by noise type (B1-B0, B2-B0, B3-B0)

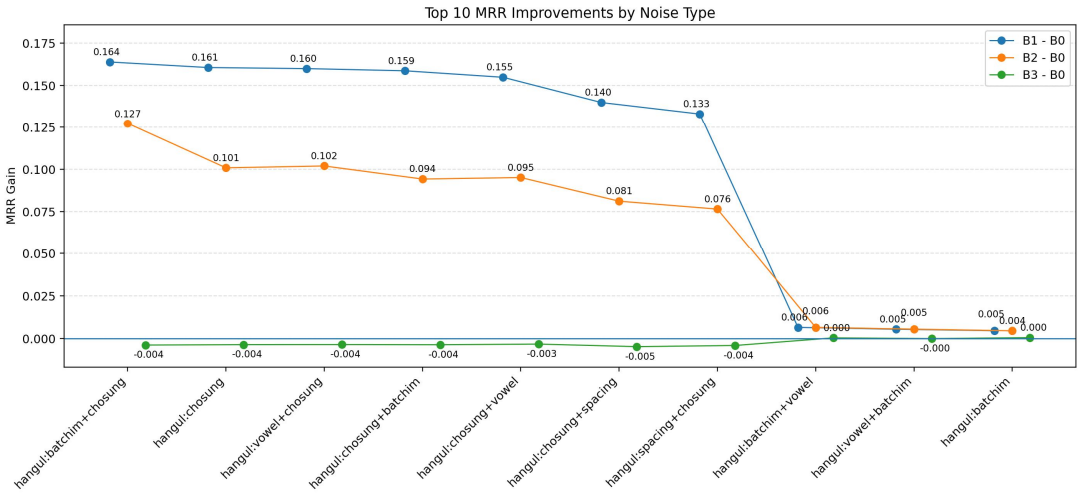

그림 3. 노이즈 유형별 MRR 개선폭 비교(B1-B0, B2-B0, B3-B0)
Fig. 3. Comparison of MRR improvements by noise type (B1-B0, B2-B0, B3-B0)

또한 운영 환경에서는 전문 검색엔진의 n-gram 분석기, 동의어 사전, fuzzy query 등을 활용한 하이브리드 후보 검색 방식도 대안이 될 수 있다. 정리하면, 본 연구의 후보 검색 단계 분석은 전처리 전략의 성능을 최종 순위 결과로만 해석하지 않고, 그 성능이 어떠한 구조적 제약 아래에서 나타난 것인지를 함께 파악했다는 점에서 의미가 있다.

5.3 노이즈 유형별 성능 분석

전처리 전략의 평균 성능 비교는 전체적인 우열을 보여주지만, 실제 주소 검색 환경에서는 입력 오류 유형에 따라 전처리 효과가 다르게 나타날 수 있다. 이에 본 연구에서는 질의 생성 과정에서 부여한 noise_type 정보를 기준으로 질의를 유형별로 구분하고, 각 전처리 조건의 Recall@10 및 MRR 개선폭을 비교하였다. 그림 2는 노이즈 유형별 Recall@10 개선폭을, 그림 3은 노이즈 유형별 MRR 개선폭을 각각 비교한 결과이다. 이를 통해 자모 기반 오류, 복합 자모 오류, 띄어쓰기 변형, 단순 문자열 오류 등 다양한 입력 환경에서 어떤 전처리 전략이 상대적으로 더 효과적인지를 분석하였다.

Recall@10 개선폭 비교 결과, B1은 여러 노이즈 유형에서 기준선(B0) 대비 가장 큰 향상을 보였다.

특히 초성 입력, 받침 변형, 유사 모음 치환, 그리고 이들이 결합된 복합 자모 오류 유형에서 개선 효과가 두드러졌다. 이는 자모·초성 보강 기반 전처리가 한글 주소 입력의 구조적 변형에 효과적으로 대응하여, 정답 주소가 상위 후보군에 포함될 가능성을 높였음을 의미한다. 반면 B2는 일부 자모 기반 오류 유형에서 제한적인 개선을 보였으나, 전반적으로 B1보다 일관된 향상 효과를 나타내지는 못하였다. B3는 띄어쓰기 변형과 관련된 일부 질의에서 보완적 역할을 수행하였으나, 전체적으로는 기준선과 유사한 수준에 머물렀다.

MRR 개선폭 비교에서도 전반적인 경향은 유사하게 나타났다. B1은 초성 및 자모 관련 오류 유형에서 정답 주소를 더 상위 순위에 배치하는 효과를 보였으며, 이는 단순히 정답 포함 여부뿐 아니라 상위 집중도 측면에서도 유의미한 개선이 있었음을 보여준다. 반면 B2는 일부 유형에서만 제한적인 기여를 보였고, B3는 평균적으로 큰 차이를 만들지 못하였다. 이러한 결과는 노이즈 유형별 분석에서도 전체 평균 성능 비교에서 확인된 B1의 우위가 특정 오류 환경에서 반복적으로 관찰됨을 보여준다.

이러한 결과는 전처리 전략의 효과가 모든 입력 환경에서 동일하게 나타나는 것이 아니라, 오류의 구조와 밀접하게 연결되어 있음을 시사한다. 특히 B1은 한글 자모 구조와 직접 관련된 오류 유형에서 가장 안정적인 개선 효과를 보였으므로, 실제 주소 검색 환경에서 우선적으로 고려할 수 있는 전략으로 해석된다. 반면 B2와 B3는 특정 상황에서 보조적으로 활용될 수 있으나, 전체 평균 성능과 노이즈 유형별 성능을 함께 고려할 때 범용 전략으로 보기는 어렵다. 따라서 노이즈 유형별 성능 분석은 전처리 전략의 선택이 단순 평균 지표의 비교를 넘어, 실제 입력 오류 분포에 대한 이해를 바탕으로 이루어져야 함을 보여준다.

5.4 전처리 전략별 효과 해석

이상의 결과를 종합하면, 본 연구에서 비교한 전처리 전략 가운데 B1이 가장 실용적인 대안으로 해석된다. B1은 전체 성능 비교에서 가장 높은 Recall@10과 MRR을 보였을 뿐 아니라, 노이즈 유형별 분석에서도 초성 입력, 받침 변형, 유사 모음 치환 등 한글 자모 구조와 직접 관련된 오류 유형에서 가장 안정적인 개선 효과를 나타냈다. 이는 자모·초성 보강 기반 전처리가 질의와 정답 주소 간 표기 간극을 줄이고, 정답 주소를 상위 순위에 배치하는 데 효과적으로 작용했음을 의미한다.

다만 B1의 성능 향상은 노이즈 질의셋의 구성과 함께 해석할 필요가 있다. 본 연구의 질의셋은 초성 입력, 받침 변형, 유사 모음 치환 등 자모 기반 오류를 포함하도록 설계되었으므로, B1의 결과는 모든 주소 입력 환경에서의 보편적 우월성이라기보다 자모 오류가 포함된 통제 실험 조건에서의 효과로 해석하는 것이 타당하다. 그러나 동일한 질의셋과 후보 검색 조건에서 B1, B2, B3의 성능과 처리 시간은 서로 다르게 나타났다. 특히 B2는 자모 편집거리 재정렬을 적용했음에도 비용 대비 성능 향상이 제한적이었고, B3는 평균적으로 기준선과 유사한 수준에 머물렀다. 이는 자모 또는 문자열 오류를 직접 다루는 전처리라 하더라도 전략별 효과와 비용이 동일하지 않으며, 실제 적용 시 정확도와 처리 시간의 trade-off를 함께 고려해야 함을 보여준다.

또한 B1의 Recall@10이 Coverage@500에 근접한 수준까지 도달했다는 점은, 현재 검색 구조에서 2단계 전처리 전략이 실질적으로 낼 수 있는 성능 상한에 가까운 결과라고 해석할 수 있다. 따라서 추가적인 성능 향상을 위해서는 B1과 같은 전처리 전략의 정교화뿐 아니라, 후보 검색 단계의 품질 개선도 함께 고려할 필요가 있다.

반면 B2는 자모 단위 유사도 재정렬이라는 점에서 이론적으로는 정밀한 순위 조정이 가능하지만, 실제 실험에서는 비용 대비 향상이 제한적으로 나타났다. 일부 자모 기반 오류 유형에서는 기준선 대비 개선 효과가 관찰되었으나, 전체 평균 성능에서는 B1을 상회하지 못했고 처리 시간 측면의 부담도 더 크게 나타났다. 이는 자모 기반 편집거리 재정렬이 특정 질의에서는 보완적 역할을 할 수 있으나, 전수 주소 검색 환경 전체에서 범용 전략으로 적용하기에는 효율성이 제한적일 수 있음을 시사한다. 다시 말해 B2는 “정밀하지만 무거운 전략”에 가깝

고, 실제 서비스 환경에서는 적용 범위와 비용을 함께 고려할 필요가 있다.

B3는 규칙 기반 띄어쓰기 보정 전략으로서, 띄어쓰기 변형이 포함된 일부 입력에서는 일정한 보완 효과를 기대할 수 있다. 그러나 본 실험에서는 전체 평균 성능과 노이즈 유형별 분석 모두에서 기준선과 유사한 수준에 머물렀다. 이는 공백 변형이 실제 검색 품질에 영향을 주는 경우가 존재하더라도, 전수 주소 검색 환경 전체에서는 띄어쓰기 보정만으로 큰 폭의 성능 향상을 기대하기 어렵다는 점을 보여준다. 또한 후보 검색기가 이미 일정 수준의 공백 변형 강건성을 가지는 경우, 규칙 기반 보정의 평균 기여는 더욱 제한될 수 있다.

결과적으로 본 연구의 전처리 전략 비교는 한글 주소 검색 환경에서 자모 구조와 직접적으로 연결된 오류에 대응하는 전략이 가장 큰 실효성을 가진다는 점을 보여준다. 동시에 전체 성능은 후보 검색 단계의 품질에 의해 구조적으로 제한되므로, 전처리 전략의 개선만으로는 성능 향상에 한계가 있다. 따라서 실제 서비스 환경에서는 B1과 같은 경량·효율 전략을 우선 적용하되, 추가적인 성능 향상을 위해서는 후보 검색 단계의 개선과 병행하는 접근이 필요하다. 이러한 해석은 전처리 전략을 단순한 기능 추가가 아니라, 정확도와 비용, 그리고 입력 오류 분포를 함께 고려한 설계 선택의 문제로 바라봐야 함을 시사한다.

Ⅵ. 결 론

본 연구는 전수 도로명주소 데이터베이스 기반 검색 환경에서 한글 주소 질의에 대한 자모 기반 전처리 전략의 효과를 분석하였다. 이를 위해 전국 도로명주소 6,414,486건으로 검색 데이터베이스를 구축하고, 대전 지역 정답 주소 112,572건을 기반으로 원본 및 노이즈 질의를 포함한 562,860건의 질의셋을 구성하였다. 전처리 전략은 기준선(B0), 자모·초성 보강(B1), 자모 기반 유사도 재정렬(B2), 규칙 기반 띄어쓰기 보정(B3)으로 구분하였으며, SQLite FTS5 기반 2단계 검색 구조에서 Coverage@k, Precision@1, Recall@k, MRR, 처리 시간 지표를 이용하여 성능을 비교하였다.

실험 결과, Coverage@500은 84.84%로 나타났으며, B1이 Recall@10 83.63%, MRR 0.8215로 가장 우수한 성능을 보였다. 특히 B1은 초성 입력, 받침 변형, 유사 모음 치환 등 한글 자모 구조와 직접 관련된 오류 유형에서 가장 안정적인 개선 효과를 보였다. 반면 B2는 일부 오류 유형에서 제한적인 향상을 보였으나 비용 대비 효율이 크지 않았고, B3는 전반적으로 기준선과 유사한 수준에 머물렀다. 이러한 결과는 자모 오류가 포함된 통제 실험 조건에서 자모·초성 보강 기반 전처리가 가장 실용적인 전략임을 보여준다.

다만 본 연구의 질의셋은 규칙 기반으로 생성된 노이즈 질의를 포함하므로, 실제 서비스 로그의 전체 오류 분포를 그대로 대표한다고 보기는 어렵다. 또한 평가용 정답셋은 대전 지역 주소에 기반하므로 전국 단위 일반화에도 한계가 있다. 향후에는 광역시, 도 지역, 농어촌 지역 등 다양한 주소 구조를 포함한 평가셋과 실제 검색 로그를 활용하여 지역별·오류 유형별 전처리 효과를 추가 검증할 필요가 있다.

동시에 본 연구는 전체 성능이 후보 검색 단계의 품질에 의해 구조적으로 제한된다는 점도 확인하였다. 즉 전처리 전략의 개선만으로는 성능 향상에 한계가 있으며, 추가적인 향상을 위해서는 후보 검색 단계의 개선이 함께 이루어질 필요가 있다. 따라서 실제 서비스 환경에서는 B1과 같은 경량·효율 전략을 우선 적용하되, 향후에는 질의 확장, 검색 토큰 가중치 조정, 하이브리드 검색 방식 등 후보 검색 품질을 높일 수 있는 방향과 병행하여 검토할 필요가 있다.

종합하면, 본 연구는 대규모 전수 도로명주소 데이터베이스와 562,860건의 질의셋을 기반으로 자모 기반 전처리 전략을 동일 조건에서 비교하였다. 또한 전처리 전략을 단계적으로 구분하여 자모·초성 보강, 자모 기반 재정렬, 규칙 기반 띄어쓰기 보정의 효과를 분리하여 분석하였다. 아울러 Precision@1, Recall@k, MRR, 평균 및 p95 처리 시간을 함께 평가하여 정확도와 비용 간의 trade-off를

제시하였다. 나아가 Coverage@500 실패 질의를 추가 분석하여 후보 검색 단계의 구조적 한계와 주요 실패 유형을 구체화하였다.

이러한 분석 결과를 바탕으로 본 연구의 학술적·실무적 의의는 다음과 같이 정리할 수 있다. 첫째, 학술적 측면에서 본 연구는 한글 주소 검색 환경에서 자모 기반 전처리 전략의 효과를 단일 성능 지표에 한정하지 않고 Precision@1, Recall@k, MRR, 처리 시간, 후보 검색 실패 유형을 함께 고려하여 정량적으로 비교·분석하였다는 점에서 의의가 있다. 특히 후보 검색 단계와 재정렬 단계를 구분하여 해석함으로써, 전처리 전략의 효과가 후보 검색 품질과 구조적으로 연결되어 있음을 실증적으로 제시하였다. 또한 자모·초성 보강, 자모 기반 재정렬, 규칙 기반 띄어쓰기 보정 전략의 효과와 한계를 동일 조건에서 비교함으로써, 한글 주소 검색에서 전처리 전략의 상대적 기여를 보다 체계적으로 설명하였다.

둘째, 실무적 측면에서 본 연구는 대규모 도로명주소 검색 환경에서 어떤 전처리 전략을 우선 적용할 수 있는지에 대한 선택 기준을 제시하였다는 의의를 가진다. 실험 결과 B1은 정확도 향상과 처리 시간 증가를 함께 고려할 때 가장 균형 잡힌 전략으로 나타났으며, B2와 B3는 적용 범위와 비용을 함께 검토할 필요가 있음을 확인하였다. 또한 전체 검색 성능이 후보 검색 단계의 품질에 의해 구조적으로 제한될 수 있음을 보여줌으로써, 실제 서비스 환경에서는 전처리 전략과 후보 검색 단계 개선을 병행해야 함을 제시하였다.

References

[1] J. H. Na, H.-G. So, G. R. Yeom, J.-O. Lee, and H.-J. Oh, "Test Set Construction for Quality Evaluation of NAK Portal's Search Service and the Status Analysis", Journal of Korean Society of Archives and Records Management, Vol. 22, No. 4, pp. 25-43, Nov. 2022. https://doi.org/10.14404/JKSARM.2022.22.4.025.

[2] H.-G. So, G. R. Yeom, and H.-J. Oh, "Construction of Pilot System to Improve Search Quality in National Archives of Korea Portal and Effects Validation", Journal of Korean Society of Archives and Records Management, Vol. 23, No. 2, pp. 117-135, May 2023. https://doi.org/10.14404/JKSARM.2023.23.2.117.

[3] Y. Bae, H. Kim, J.-H. Lim, H.-K. Kim, and K. J. Lee, "2-Phase Passage Re-ranking Model based on Neural-Symbolic Ranking Models", Journal of KIISE, Vol. 48, No. 5, pp. 501-509, May 2021. https://doi.org/10.5626/JOK.2021.48.5.501.

[4] S. Mudasir, A. W. Agha, S. M. Hussain, and S.-T. Chung, "Field-Adaptive Dense Retrieval of Structured Documents", Journal of Korea Multimedia Society, Vol. 28, No. 8, pp. 1001-1014, Aug. 2025. https://doi.org/10.9717/kmms.2025.28.8.1001.

[5] S.-H. Kim and H.-G. Cho, "Improved First-Phoneme Searches Using an Extended Burrows-Wheeler Transform", KIISE Transactions on Computing Practices, Vol. 20, No. 12, pp. 682-687, Dec. 2014. http://dx.doi.org/10.5626/KTCP.2014.20.12.682.

[6] P. M. Park, "A Study on a Preprocessing Framework for Unrefined Road Name Address Fields Using Character-Based Similarity Algorithms", Master's thesis, Dept. of Computer Engineering, Chungnam National University, pp. 1-34, Feb. 2023.

[7] G. J. Park, W. M. Song, E. J. Kim, and M. W. Kim, "Development of Efficient Address Cleaning System for CRM", Proc. Korean Computer Congress, Jeju, Korea, Vol. 34, No. 1, pp. 128-132, Jun. 2007.

[8] J. Y. Kim, H. J. Kim, and J. W. Lee, "Street Name Address Parsing Model based on Bidirectional Gate Recurrent Unit through Automatic Construction of Training Data", Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, Vol. 41, No. 5, pp. 301-310, Oct. 2023. https://doi.org/10.7848/

ksgpc.2023.41.5.301.

[9] J. Lee, "Development of Middleware for Integration of Administrative Data using Open Source and Deep Learning", Research Report, University of Seoul, pp. 1-18, Sep. 2020.

[10] S. Seok and J. Lee, "Development of Geocoding and Reverse Geocoding Method Implemented for Street-based Addresses in Korea", Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, Vol. 34, No. 1, pp. 33-42, Mar. 2016. http://dx.doi.org/10.7848/ksgpc.2016.34.1.33.

[11] K. Lee, J.-Y. Kim, and B.-G. Kang, "A Method of Recognizing and Validating Road Name Address from Speech-oriented Text", Journal of Internet Computing and Services, Vol. 22, No. 1, pp. 31-39, Feb. 2021. http://dx.doi.org/10.7472/jksii.2021.22.1.31.

[12] S. Saravit, J.-H. Bae, K.-H. Lee, and W.-S. Cho, "Global Address Data Quality Verification and Improvement Techniques using Deep Learning", Journal of KIIT, Vol. 20, No. 12, pp. 15-24, Dec. 2022. http://dx.doi.org/10.14801/jkiit.2022.20.12.15.

[13] I. Kim and W. H. Jeon, "A Study on Automation of Road Name Address Layer Construction using Convolutional Neural Network (CNN): Focusing on Real Road Width and Building Layers", Journal of the Korean Cadastre Information Association, Vol. 25, No. 3, pp. 125-134, Dec. 2023. https://doi.org/10.46416/JKCIA.2023.08.25.3.125

[14] E. L. dos Santos, B. C. Ávila, F. M. Hasegawa, and C. A. A. Kaestner, "Query Expansion and Noise Treatment for Information Retrieval", Proc. CACIC 2003, La Plata, Argentina, pp. 717-728, Oct. 2003.

[15] J. Lee, J. Kim, and G. G. Lee, "Exploring Back Translation with Typo Noise for Enhanced Inquiry Understanding in Task-Oriented Dialogue", Proc. 11th Dialog System Technology Challenge, Prague, Czech Republic, pp. 185-192, Sep. 2023.

[16] W. Feng, B. Liu, D. Xu, Q. Zheng, and Y. Xu, "GraphMR: Graph Neural Network for Mathematical Reasoning", Proc. EMNLP 2021, Punta Cana, Dominican Republic, pp. 3395-3404, Nov. 2021. https://doi.org/10.18653/v1/2021.emnlp-main.273.

[17] S.-Y. Park, S.-H. Kim, and H.-G. Cho, "An Original Text Protecting Search Method for Huge Korean Documents", Journal of KIISE: Computing Practices and Letters, Vol. 18, No. 7, pp. 563-567, Jul. 2012.

[18] S. Lee, "Performance Analysis of a Korean Word Autocomplete System and New Evaluation Metrics", Journal of the Korean Society of Marine Engineering, Vol. 39, No. 6, pp. 656-661, Jul. 2015. http://dx.doi.org/10.5916/jkosme.2015.39.6.656.

[19] J. H. Park and M. G. Kang, "Extraction and Basic Analysis of Spatial Information in 120 Dasan Call Civil Complaint Texts through Named Entity Recognition Modeling", Journal of Korea Planning Association, Vol. 59, No. 1, pp. 145-160, Feb. 2024.

저자소개

박 상 현 (Sanghyeon Park)

2010년 2월 : 충남대학교 토목공학과(학사)
2026년 2월 : 서울사이버대학교 인공지능학과(공학사)
2024년 3월 ~ 현재 : 고려사이버대학교 융합정보대학원 석사과정
관심분야 : 인공지능, 빅데이터 분석, 자연어 처리, 정보검색, 공간정보

도 성 룡 (Sungryong Do)

2008년 2월 : 상명대학교
소프트웨어학부(이학사)
2010년 2월 : 상명대학교
컴퓨터과학과(컴퓨터과학석사)
2016년 8월 : 상명대학교
컴퓨터과학과(컴퓨터과학박사)
2012년 4월 ~ 2017년 4월 :
현대오트론 품질팀 선임연구원
2017년 7월 ~ 2020년 8월 : 상명대학교 산학협력중점교수
2020년 9월 ~ 2023년 8월 : 경민대학교 컴퓨터SW과
조교수
2023년 9월 ~ 현재 : 고려사이버대학교 컴퓨터공학부
조교수
관심분야 : 소프트웨어 공학, 소프트웨어 품질,
소프트웨어 안전, 소프트웨어 프로세스