비용 효율적인 지능형 AICC를 위한 자가 진화형 4계층 정규화 게이트

김지훈, 길기훈, 이상금*
국립한밭대학교
20227128@edu.hanbat.ac.kr, 20201859@edu.hanbat.ac.kr, sangkeum@hanbat.ac.kr

4-Layer Self-Evolving Normalization Gate for Cost-Effective Intelligent AICC

Jihun Kim, Gilhun Gil, Sangkeum Lee*
Hanbat National University

요 약

최근 비대면 상담 수요의 증가로 AICC(AI Contact Center)의 중요성이 대두되고 있으나, MZ세대를 중심으로 한 신조어 및 비정형 텍스트(Slang)의 증가는 기존 자연어 처리시스템의 성능 저하와 상담 실패를 유발하는 주요 원인이 되고 있다. 본 논문에서는 이를 해결하기 위해 Static(L1), Semantic(L2), Reasoning(L3), Evolution(L4)의 4단계 계층으로 구성된 자가 업데이트 기반 4계층 정규화 프레임워크를 제안한다. 제안된 시스템은 입력 텍스트의 복잡도에 따라 처리 계층을 동적으로 선택함으로써, 빈출 은어는 저지연 구조로 처리하고 미등록 변형어는 벡터 공간 기반 의미 유사도를 활용하여 유연하게 대응한다. 또한 상위 계층의 추론 결과를 하위 계층으로 전이하는 피드백 구조를 통해 운영 과정에서 시스템 성능을 점진적으로 개선한다. 국내 커뮤니티에서 수집한 소규모 실제 데이터를 활용한 실험 결과, 학습되지 않은 변형 은어에 대해 81.5%의 의미적 적중률을 기록하였으며, LLM 호출 비율을 기존 대비 약 0.14배 수준으로 감소시키는 비용 효율성을 확인하였다.

I. 서 론

최근 비대면 상담 수요의 증가로 AICC(AI Contact Center)가 기업과 고객 간의 핵심 소통 채널로 활용되고 있다. 그러나 MZ세대를 중심으로 한 신조어, 은어, 오타가 혼재된 비정형 텍스트는 기존 자연어 처리 시스템의 성능 저하와 상담 의도 파악 실패를 유발하는 주요 요인이다[1]. 사전 기반 형태소 분석기는 고정된 어휘 집합에 의존하므로 동적인 신조어를 미등록어로 처리하거나 오분석하는 한계를 가진다[2].

LLM을 활용한 문맥 기반 정규화 기법이 제안되고 있으나, 모든 입력을 LLM 기반 추론으로 처리하는 방식은 실시간 대규모 트래픽 환경에서 운영 비용과 응답 지연 측면의 구조적 한계를 가진다[3]. 한편, 지능형 시스템 환경에서는 실시간 데이터 처리 과정에서 정확도뿐만 아니라 운영 비용과 처리 효율을 동시에 고려한 구조적 설계의 중요성이 지속적으로 보고되어 왔다[4]. 이에 본 연구는 해시 기반 정규화, 벡터 유사도 검색, LLM 추론을 계층적으로 결합한 자가 업데이트 기반 4계층 정규화 프레임워크를 제안하며, AICC 환경에서 요구되는 처리 효율성과 정규화 정확성을 동시에 고려한 구조를 목표로 한다.

II. 본 론

본 연구에서 제안하는 4계층 레이어 프레임워크는 입력 데이터의 복잡도에 따라 처리 경로를 동적으로 분기하는 이원화된 최적화 전략을 채택한다. 이는 고빈도 · 저복잡 입력에 대해서는 처리 지연을 최소화하고, 저빈도 · 고복잡 입력에 대해서는 의미적 정확도를 우선하는 두 가지 처리 경로를 병렬적으로 운용하는 구조이다.

시스템은 그림 1와 같이 4단계 파이프라인으로 구성된다. 입력 텍스트는 복잡도에 따라 L1(Static Layer), L2(Semantic Layer), L3(Reasoning Layer)로 분기 처리되며, L4(Evolution Layer)는 상위 계층에서 축적된 지식을 하위 계층으로 전이하여 시스템을 점진적으로 개선한다.

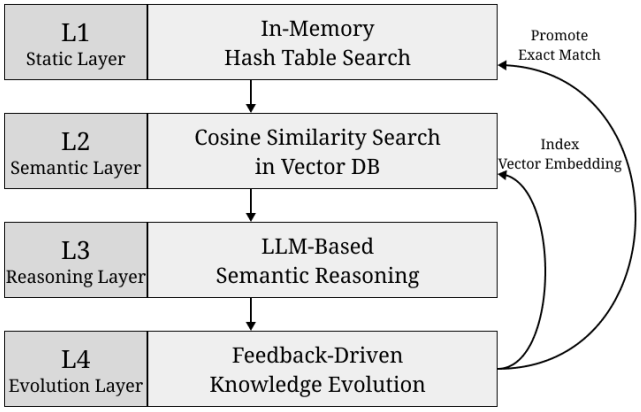

그림 1: 제안하는 자가 업데이트 기반 4계층 레이어 아키텍처

L1은 빈도수가 높은 고정형 은어를 처리하는 메모리 기반 해시 테이블이다. “ㅇㅈ(인정)”, “강추(강력 추천)”와 같이 형태가 고정되고 빈도가 높은 상위 20%의 은어를 $O(1)$ 시간 복잡도로 변환하여 대다수 트래픽을 저지연 구조로 처리한다.

L2는 형태소 파괴가 심한 변형 은어를 방어하기 위한 벡터 데이터베이 계층이다. “아아(아이스 아메리카노)”, “뜨아(뜨거운 아메리카노)”와 같은 표현은 BGE-M3 모델을 통해 고차원 벡터로 임베딩되며, 대조 학습 기반 벡터 유사도 측정[5]을 통해 가장 근접한 표준 은어 개념과 매칭된다. 이때 유사도가 임계값($\theta = 0.75$) 미만일 경우 오분석을 방지하기 위해 L3 계층으로 우회 처리한다.

L3는 미등록 신조어나 강한 문맥 의존성을 가지는 표현을 처리

하는 LLM 기반 추론 계층이다. 이 계층은 가장 높은 처리 비용과 지연을 가지지만, 앞선 계층에서 처리되지 않은 입력에 대해 최후의 의미적 추론 수단으로 활용된다.

L4는 시스템의 자가 업데이트를 담당하는 피드백 계층으로, 최근 제안된 LLM 기반 자가 수정 개념을 반영한다[6]. 은어 w의 승격 여부는 다음 식을 통해 결정된다.

$$Score(w) = \alpha \cdot Freq(w) + \beta \cdot Consistency(C_1, \ldots, C_n) \quad (1)$$

여기서 α와 β는 각각 빈도와 문맥 일관성에 대한 가중치로, 사전 실험을 통해 경험적으로 $\alpha = 0.3, \beta = 0.7$로 설정하였다. 문맥 일관성은 동일 은어가 서로 다른 세 개 이상의 문맥에서 90% 이상의 의미적 일관성을 보일 경우에만 유효한 것으로 판단한다.

검증을 통과한 은어는 유형에 따라 하위 계층으로 전이된다. 형태가 고정된 단일 어휘는 L1 해시 테이블에 즉시 삽입되며, 문맥적 변형이 잦은 표현은 BGE-M3 임베딩을 통해 L2 벡터 데이터베이스에 신규 인덱스로 등록된다. 이러한 업데이트는 시스템 재시작 없이 런타임 중 수행 가능하도록 설계되었다.

시스템 안정성을 위해 추가적인 지식 관리 정책도 적용한다. 동일한 은어에 대해 L1과 L2의 결과가 상충할 경우 처리 지연이 낮은 L1의 결과를 우선 적용하며, 저장 공간 효율성을 위해 LRU 기반 퇴출 정책을 적용하여 지식 베이스의 최신성을 유지한다.

III. 실험 및 결과

제안 시스템의 자가 업데이트 효율성과 처리 경향을 분석하기 위해, 국내 커뮤니티 '인스티즈'에서 수집한 특정 신조어 클러스터 관련 실제 데이터 125건을 활용하여 파일럿 규모의 시뮬레이션을 수행하였다.

실험은 전체 데이터를 Phase 1(Learning, 60건)과 Phase 2(Generalization, 65건)로 분할하여 진행하였다. Phase 1에서는 대다수의 은어가 미등록어로 인식되어 L3(Reasoning) 호출 비율이 86.7%에 달하였으나, 이 과정에서 L4 계층이 핵심 어휘를 학습하여 L1/L2 캐시로 점진적으로 승격시켰다. Phase 2에서는 Phase 1에서 학습되지 않은 새로운 문장이 유입되었음에도 불구하고, L2 Semantic Layer가 벡터 유사도 검색을 통해 변형 표현을 81.5%의 적중률로 식별하였다. 그 결과, 전체 트래픽 대비 L3 호출 비율이 12.3%(약 0.14배) 수준으로 감소하면서도 정규화 성능이 유지되는 경향을 확인하였다.

표 1: 자연어 변형에 대한 일반화 성능 검증 결과

실험 단계	N	L1 (Static)	L2 (Semantic)	L3 (Reasoning)	Hit Rate
Phase 1: Learning	60	3 (5.0%)	5 (8.3%)	52 (86.7%)	13.3%
Phase 2: Generalization	65	4 (6.2%)	**53 (81.5%)**	**8 (12.3%)**	**87.7%**
변화량	-	1.2배	**9.8배**	**0.14배**	**6.6배**

효율성 및 확장성을 검증하기 위해 기대 지연 시간 $E[T]$를 다음과 같이 정의한다.

$$E[T] = (P_{L1} \cdot T_{L1}) + (P_{L2} \cdot T_{L2}) + (P_{L3} \cdot T_{L3}) \quad (2)$$

표 2와 같이 학습 후 P_{L3}가 0.123으로 급감하여 L3 오프로딩 비율과 토큰 절감률이 크게 개선됨을 확인한다.

그림 2는 누적 트래픽에 따른 L3 호출 빈도의 수렴 양상을 보여준다. Phase 2 진입 후 기울기가 급격히 완만해지며 비용 효율성이 최적화됨을 확인할 수 있다.

표 2: 제안 아키텍처의 효율성 및 확장성 분석

평가 항목	LLM 단독 (Baseline)	제안 모델 (Proposed)	개선 효과 (Impact)
L3 호출 비율	100%	12.3%	**87.7% 감소**
토큰 소모량	N (전체)	$0.12N$	**88% 감소**
처리 구조 효율	단일 추론	계층적 분산 처리	**처리량 향상**
비용 곡선	선형 증가	로그형 수렴	**운영 비용 수렴**

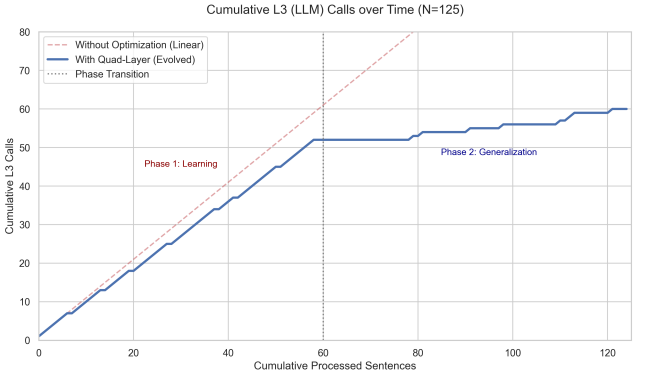

그림 2: 누적 트래픽 대비 L3 호출 수렴 곡선.

IV. 결 론

본 연구는 AICC 환경에서 발생하는 비정형 은어 문제를 완화하기 위해, LLM 기반 추론과 벡터 검색을 결합한 4단계 자가 업데이트 기반 정규화 프레임워크를 제안하였다. 제안 구조는 입력 복잡도에 따라 처리 계층을 분기하고, 상위 계층의 추론 결과를 하위 계층으로 전이함으로써 비용 효율적인 정규화 처리를 가능하게 한다. 파일럿 규모의 실험 결과, 125건의 실제 커뮤니티 데이터를 처리하는 과정에서 LLM 기반 추론 계층(L3)의 호출 비율을 12.3%(약 0.14배) 수준으로 감소시키면서도, 학습되지 않은 변형 은어에 대해 81.5%의 의미적 적중률을 유지하는 경향을 확인하였다. 이는 제한된 실험 규모 내에서 비용 절감과 정규화 품질 간의 균형 가능성을 보여준다. 본 연구는 소규모 데이터 기반 시뮬레이션에 한정된 분석이라는 한계를 가지며, 향후 연구에서는 다국어 AICC 환경으로의 확장과 대규모 상용 트래픽 조건에서의 성능 및 안정성 검증을 추가적으로 수행할 계획이다.

참 고 문 헌

[1] M. Wong, A. Alshehri, S. Kao, and H. He, "Polynorm: Few-shot llm-based text normalization for text-to-speech," in *EMNLP 2025 Industry Track*, 2025.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[3] Y. Zhang *et al.*, "A chat about boring problems: Studying gpt-based text normalization," *arXiv preprint arXiv:2309.13426*, 2023.

[4] S. Lee, S. H. Nengroo, H. Jin, Y. Doh, C. Lee, T. Heo, and D. Har, "Anomaly detection of smart metering system for power management with battery storage system/electric vehicle," *ETRI Journal*, vol. 45, no. 4, pp. 650–665, 2023.

[5] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proceedings of EMNLP*, pp. 6894–6910, 2021.

[6] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhu, Y. Yang, *et al.*, "Self-refine: Iterative refinement with self-feedback," 2023.