1. 논문제목 (국문 혹은 영문)

Designing a RAG-Based Matching Agent with Role-Differentiated LLMs

2. 분야

경영정보

3. 저자명 및 소속(국문/영문)

주저자

윤이지(Yiji Yoon)

경희대학교 빅데이터응용학과

chokogangdo1015@khu.ac.kr

교신저자

이경전(Kyoung Jun Lee)

경희대학교 경영학과 및 빅데이터응용학과

klee@khu.ac.kr

초록

본 연구는 비정형 자연어 메시지들의 의미를 해석하여 메시지의 주체들을 매칭하는 문제를 'Loose Matching'으로 정의하고, 이를 해결하기 위해, 역할별로 분리된 LLM을 활용하는 Retrieval-Augmented Generation(RAG) 기반 매칭 에이전트를 설계하였다. 제안된 시스템은 LLM으로 구성된 각 노드가 고유 역할을 수행하며, 다양한 판단 기준에 따라 메시지 간 매칭 적합성을 분석한다. 약 17,000건의 실제 및 합성 메시지를 활용한 시연 결과, 기존 임베딩 기반 검색이 처리하지 못한 난이도 높은 사례 중 약 49%의 사례들에서 상대적으로 더 설득력 높은 매칭을 성사시켰다. 본 연구는 키워드 중심 검색의 한계를 넘어, 실용적 적용이 가능한 의미 중심 매칭 에이전트를 제안한다.

1. 서론

1.1 연구배경

매칭은 서로 다른 두 집단의 참여자들이 각자 상대에 대한 선호를 가지고 있을 때, 이들을 짝지어주는 과정이다. 이론적으로는, 어떤 두 참여자가 현재의 짝보다 서로를 더 선호하는 경우가 없는 쌍방 수용 가능 상태를 '안정적(stable)'인 매칭이라고 보며, 이러한 구조는 공정성과 진실된 선호 표현을 전제로 한다[1].

그러나 온라인 커뮤니티나 플랫폼과 같은 현실 환경에서 이루어지는 매칭은 이러한 이론적 구조와는 다소 차이를 보인다. 사용자는 자신의 선호를 명시적으로 표현하지 않거나, 비정형적이고 암시적인 자연어 표현을 사용하는 경우가 많다. 예를 들어, X(구 트위터)에서는 한 사용자가 "저 뵙인데 이번에 같이 가실 분?"이라는 트윗을 게시할 수 있다. 이 트윗은 GD 콘서트 동행자를 모집하려는 의도를 내포하고 있으나, 'GD 콘서트 동행'이라는 명시적 표현은 포함되어 있지 않다. 이 트윗이 관련 동행자를 찾고 있는 사용자에게 도달하기 위해서는, 후자가 다양한 키워드 조합을 시도하거나 수동으로 피드를 탐색해야 한다.

이처럼 키워드 표현 방식의 차이, 맥락적 단서(예: '뵙'이라는 팬덤 용어), 게시 시점 등은 정확한 매칭을 어렵게 만든다. 기존 시스템은 이러한 암시적 표현이나 문화적 맥락 정보를 구조화하여 해석하지 못하며, 이로 인해 사용자는 반복적인 키워드 탐색과 표현 수정에 높은 인지적 노력을 기울여야 한다는 한계가 존재한다.

본 논문은 이러한 제약 하에서 이루어지는 의미를 중시하는 매칭 구조를 다룬다. 본 연구에서의 접근 방법은, 참여자의 메시지에 대해 의미 적합성, 역할 상보성 등을 기준으로 매칭 상대를 탐색하는 것이다. 이와 같은 구조는 전통적인 매칭 이론에서 제시하는 쌍방 선호 기반의 안정적 매칭과는 차이가 있다. 다만 도메인과 맥락이 제각각인 온라인상의 수많은 비정형 자연어 메시지들을 고려할 때, 이러한 방식이 좀 더 실용적인 매칭 메커니즘으로 작동할 가능성이 있다고 보았다. 본 논문에서는 해당 방식을 "Loose Matching"이라 명명하고, 기존의 형식적 매칭 구조 대신 비정형 표현의 의미 해석 가능성에 초점을 맞춘 매칭 문제라는 관점에서 접근한다.

1.2 연구 필요성

기존 키워드 기반 검색 시스템에 대한 보완책으로서, 최근 대규모 언어 모델(LLM)과 임베딩을 활용한 의미적 검색, 그리고 이를 결합한 Retrieval-Augmented Generation(RAG) 구조를 활용하는 시도가 등장하고 있다. 가령. Amazon의 Rufus는 사용자가 "이 재킷은 세탁 가능한가요?" 또는 "입문자에게 적합한 전동드릴은?"과 같은 자연어 쿼리를 입력했을 때, 해당 질문에 대해 제품 설명, 리뷰, Q&A 등 다양한 문서를

검색(Retrieval)하고 이를 바탕으로 적절한 응답을 생성(Generation) 하는 RAG 기반 쇼핑 도우미로 구현되어 있다[2].

현재의 LLM은 직관적 의미 판단과 맥락 해석 측면에서 한계가 있어서, 표현은 다르지만 의미적으로 연결 가능한 메시지를 정확히 구분하지 못하는 경우가 있다[3]. 게다가 단일 LLM 모델만으로는 복잡한 문제의 분해 및 해결이나 확장성 측면에서 한계를 보이며, 부정확한 응답에 대한 자가 수정을 효과적으로 수행하지 못한다[4]. RAG의 Retrieval 단계 또한 문맥 기반 검색을 가능하게 하지만, 여전히 문장 형태의 유사성에 의존하기 때문에, Loose Matching과 같은 과제에서는 표현 다양성과 정서적 맥락을 충분히 반영하기 어렵다.

한편, Steiner 는 매칭 과정을 Screening, Selecting, Matching 의 세 단계로 구분하였다[5]. Screening은 가능한 후보군을 좁히는 초기 필터링 과정이며, Selecting은 그 중에서 적합한 대상을 선택하는 과정, Matching 은 선택된 대상과 실제로 연결되는 마지막 단계를 의미한다. 온라인 플랫폼의 검색 시스템이나 단일 LLM을 활용한 RAG 구조 등의 기존 시스템들은, Screening 및 Selecting 단계에서 의미 판단, 표현 다양성 대응 등에서 병목을 겪고 있다. 이에 Loose Matching이 요구되는 자연어 기반 매칭 상황에 대해, 다음 2가지를 핵심 문제로 정의하고자 한다.

1) 의미를 고려한 후보 탐색의 한계 (Screening 문제):
비정형 자연어 메시지는 표현 방식이 다양하고 명시적 속성이 부재하기 때문에, 기존의 검색 방식만으로는 의미적으로 적절한 매칭 후보를 탐색하기 어렵다. 이로 인해 초기 단계에서부터 유의미한 연결 가능성을 놓치게 된다.

2) 직관적 의미 판단 기준의 부재 (Selecting 문제):
탐색된 후보 중 누구와 연결되어야 하는지를 판단하는 과정에서는, 단순 유사성 외에도 사용자의 감정, 의도, 상황 맥락이 고려되어야 한다. 그러나 기존 시스템은 이를 판단할 수 있는 기준이나 체계를 제공하지 못해, 매칭 결정이 사용자 직관에 전적으로 의존하게 된다.
또한 이러한 문제에 따라 본 논문은 다음의 두 가지 Research Question을 설정한다.

RQ1. 다양한 표현과 맥락을 지닌 자연어 메시지들 중 의미 적합성이 높은 메시지간 매칭의 가능성을 높이기 위해, 어떤 방식으로 후보군을 추출(Screening)할 수 있는가?

RQ2. 단순한 유사성 판단을 넘어서, 역할 상보성 등 직관적 요소를 반영한 선별(Selecting) 체계를 어떻게 설계할 수 있는가?

2. 관련 연구

2.1 매칭의 개념과 의미 기반 접근으로의 확장

매칭은 참여자 간 선호를 기반으로 상호 수용 가능한 쌍을 형성하는 것이며, 경제학에서는 참여자가 자신의 진정한 선호를 드러내도록 하는 유인(incentives to reveal their true preference)과, 쌍들 간의 불만족을 제거하는 안정성(stability)을 주요

기준으로 삼는다. Roth는 이러한 조건을 충족하는 매칭 메커니즘을 정립하였다[1]. 한편 Steiner는 매칭을 단순한 자원 배분이 아닌 자원과 개인을 효과적으로 연결하는 사회적 장치로 해석하며, 정보 기술과 플랫폼 경제의 확산에 따라 매칭 개념이 사회기술적 맥락으로 확장되고 있음을 강조한다[5].

최근에는 대규모 언어 모델(LLM)과 임베딩 기술의 발전으로, 명시적 선호 없이도 자연어 추론을 기반으로 한 의미적 매칭이 가능해지는 방향으로 연구가 확장되고 있다. 예를 들어, LEAPME는 속성 임베딩과 메타 피처를 활용해 이질적인 속성 쌍의 의미 적합성을 분류하며[6], DeepAlignment는 워드 임베딩을 활용하여 유사도를 계산하고 Stable Marriage 알고리듬으로 매칭을 수행한다[7]. 또한, LLM의 프롬프트 설계를 통해 매칭 전략을 조정하거나, 의미에 기반한 매칭을 유도하려는 시도도 진행 중이다. LLM이 Matching, Comparing, Selecting이라는 세 전략 중 어떤 판단 구조를 따르는지 분석한 연구도 있고[8], Generator-Refiner-Scorer 구조를 통해 후보쌍 간 의미 연결의 근거를 언어적으로 명시하며 매칭 과정을 추론에 기반한 판단으로 확장한 연구도 있다[9].

이러한 시도들은 매칭의 본질을 언어적 해석과 추론의 문제로 전환시키며, 의미 기반 매칭을 기술적으로 실현하려는 접근이라 할 수 있다.

2.2 쿼리 재구성

기존 임베딩 기반 매칭 방식은 주로 코사인 유사도에 따라 문장을 비교하지만, 이는 표현 유사성에는 강점이 있어도 맥락 적합성이나 정서적 의미 연결성을 반영하는 데에는 한계가 있다. 예컨대 의미적으로 더 적절한 문장이 있음에도 불구하고 표현 유사성이 높은 문장이 선택되는 오류를 'False Vector Matching'이라 명명하며, 코사인 유사도 중심 접근의 한계를 지적한 연구도 있다[10].

이러한 문제를 극복하기 위해, 최근에는 쿼리 재구성(Query Reformulation) 전략이 활용된다. 대표적으로 (1) 쿼리 확장: 질문을 다양한 표현이나 하위 질문으로 분해, (2) 쿼리 변환: 질문 형식 재설계 또는 가상 응답 기반 재검색, (3) 쿼리 라우팅: 질문 유형 분석 기반 자동 분기 방식으로 구분된다. 이들 전략은 LLM과 결합되어 표현 다양성과 검색 포괄성을 높이는 데 기여하고 있다[11].

예를 들어, Analyze-Generate-Refine 구조는 쿼리를 분석하고 다양한 표현을 생성한 뒤 적절한 후보를 선택하는 과정을 통해 복잡한 질문을 정제한다[12]. IRCoT 구조는 초기 검색 결과를 기반으로 LLM 추론을 반복 적용하며, 쿼리를 점진적으로 보완해 검색 정확도를 높인다[13]. 본 논문에서도 LLM 기반 쿼리 재구성을 활용함으로써, Loose Matching 환경에서 Recall(검색 대상 중 실제 정답을 찾아내는 비율)을 높인다.

2.3 복합 추론 작업을 위한 역할 분리형 LLM 구조

LLM은 일반적인 도메인에서의 자연어 생성에 강점을 지니며, 추천·설명과 같은 작업에 적합하다. 그러나 복잡한 문제를 제로샷 환경에서 안정적으로 처리하는 데에는 한계가 존재하며, 이를 보완하기 위한 전략으로 역할 분리 및 태스크 분해 기반 설계가 제안되고 있다.

예를 들어, Logic-Scaffolding은 관점 기반 설명과 chain-of-thought prompting을 결합하여 중간 추론 단계를 통해 설명의 정확성과 일관성을 개선한다[14]. Meta Prompting은 하나의 LLM을 다수의 역할 인스턴스로 분할해, 각 인스턴스가 전문가 역할을 수행하며 메타 모델이 이를 조율하는 방식이다. 명시적 분리는 없지만 프롬프트 설계를 통해 다중 역할을 시뮬레이션하는 구조다[15]. AutoGen은 사용자, 검색, 실행, 조정자 등의 명시적 역할을 갖춘 에이전트들이 상호 대화를 통해 작업을 수행하는 구조를 제시하며, 단일 에이전트 대비 높은 문제 해결 성능을 보였다[16].

본 논문은 이러한 구조적 전략을 바탕으로, Loose Matching 문제를 효과적으로 처리하기 위해 역할이 분리된 여러 LLM을 사용하는 구조를 채택하였다. 각 LLM은 분석, 검색, 쿼리 재구성, 평가 등 상이한 기능을 수행하며, 이들의 협업을 통해 판단 과정을 분산하고 보다 정밀하고 유연한 매칭을 가능하게 한다.

3. 제안 시스템의 설계 구조

제안 시스템은 도메인에 구애받지 않는 일반화된 Loose Matching 구조를 지향한다. 이를 위해 당근, 숨고, 필름메이커스 등 다양한 도메인에서 수집된 실제 크롤링 데이터 또는 합성 메시지로 구성된 약 17,000개의 자연어 기반 매칭 요청 메시지를 처리할 수 있도록 설계되었으며, 이들 메시지들은 비정형 표현, 은유, 축약 등 다양한 언어적 특징을 포함하고 있다. 본 논문에서는 이 메시지들이 벡터 형태로 전처리되어 하나의 벡터 데이터베이스에 저장되어 있다고 가정하고, 입력 메시지에 대해 의미적으로 적합한 후보를 탐색하는 구조를 제안한다.

3.1 Loose Matching 판단을 위한 프레임워크

본 논문에서는 구조화되지 않은 사용자 메시지 간의 의미 기반 연결을 달성하기 위해, 메시지 간 Loose Matching을 수행하는 프레임워크를 설계하였다. 이 프레임워크는 후보 메시지의 적합성을 다각도로 판단하기 위한 네 가지 주요 기준으로 구성되며, 각 기준은 서로 다른 유형의 매칭 가능성을 포착하는 데 기여한다.

(1) Matching Target 및 상위어 유사성

메시지 내 명시된 대상의 동일 여부가 우선순위이나 그렇지 않더라도, 상위어 범주에서 일치할 경우 의미적 유사성이 있다고 간주한다. 예컨대 “바람막이”와 “가디건”은 모두 ‘아우터’라는 상위 개념으로 연결 가능하다.

(2) 표현 명시성과 해석 가능성

매칭 판단에는 상대 메시지의 의도 명확성도 중요하다. “마사지 받을 곳 찾아요”와 같이 직접적인 요청은 명시적 표현으로 해석되며, “이번 주 야근이 많았어요”처럼 간접 표현은 암시적 의도로 판단된다. 표현 명시성은 해석 난이도와 후보군 선정에 영향을 준다.

(3) 매칭 대상 유형(Type) 구분

메시지가 지칭하는 대상이 사람, 장소·이벤트·서비스, 물건 중 어디에 해당하는지를 구분함으로써, 동일 유형 간 의미 비교 및 후보 필터링이 가능하다. 이는 맥락적으로 부적절한 매칭을 사전에 제거하는 데 유용하다.

(4) 역할 상보성 여부 구분

메시지 쌍 간의 의미적 관계 유형이 동질적(homogeneous)인지, 이질적(heterogeneous)인지 구분한다. 예를 들어, “짐 옮겨주세요”와 “짐 옮겨드립니다”는 역할이 이질적이면서 역할 상보성이 있으므로 매칭 가능성이 존재하며, 반대로 “전시회 같이 가실 분”과 “전시회 동행 구해요”는 동질적인 요청으로 분류된다.

본 논문은 각 판단 기준에 대한 평가를 수행한 뒤, 이를 종합한 결과를 바탕으로 최종 연결 여부를 결정하는 구조를 채택하였다. 이는 LLM이 단일 판단 구조에서 발생시키는 의미 해석 오류를 줄이고, 다양한 매칭 가능성을 기술적으로 구조화하기 위한 설계 전략이다.

3.2 제안 시스템의 구성 요소

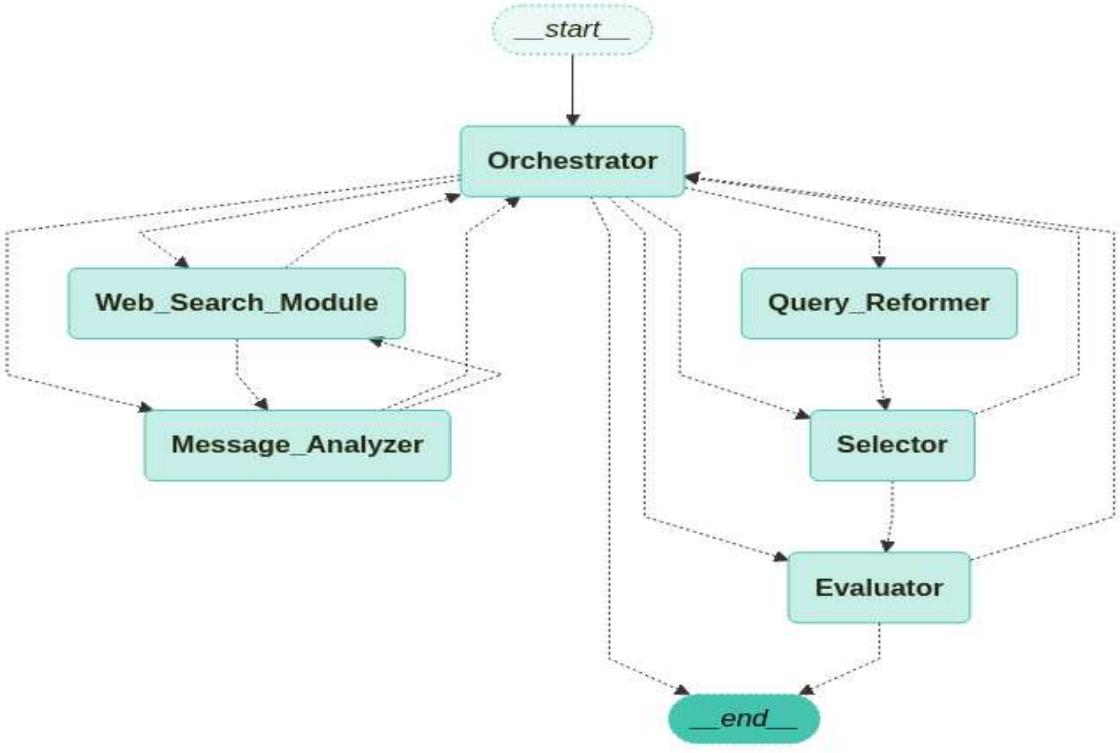

그림 1 Matching Agent 구조

본 논문에서 제안하는 시스템인 매칭 에이전트는 다양한 역할을 부여받은 LLM 들이 협업함으로써 사용자가 최종적으로 원하는 매칭 상대를 선택한다. 본 논문에서는 이러한 프로세스 내에서 판단 및 제어를 수행하는 각 구성 요소를 '노드'로 정의하며, 역할에 따른 판단 단위로서 독립적인 입력과 출력을 갖는다. 해당 시스템은 역할 구분에 따라 4 개의 메인 노드, 2 개의 보조 노드로 구성되어 있다. 4 개의 메인 노드는 Orchestrator, Query Reformer, Selector, Evaluator 이며, 2 개의 보조 노드는 Message Analyzer, Web Search Module 이다(그림 1). 각 노드는 GPT-4o 기반으로 구현되어 있으며, 텍스트 생성을 통한 대화로 프로세스가 진행된다. 또한 모든 매칭 요청 메시지는 OpenAI 의 text-embedding-3-small 모델로 벡터화되며, 검색 효율성을 위해 FAISS 를 활용한 벡터 인덱스에 저장된다. 이는 코사인 유사도 기반의 최근접 이웃 검색을 지원하며, 후술할 Query Reformer 가 매칭 대상 후보군을 추려내는 데 사용된다.

(1) Orchestrator

사용자의 자연어 입력을 받아 전체 매칭 과정을 단계별로 조율하는 중심 역할을 한다. 메시지 분석, 쿼리 생성, 후보 탐색·선택·평가 등 각 단계를 통제하며, 메시지의 명확성과 난이도에 따라 보조 모듈을 전략적으로 호출한다.

(2) Query Reformer

입력 메시지를 바탕으로 두 개의 의미 기반 쿼리를 생성하되, 각기 다른 재구성 전략(상보성 전환, 상위어 일반화, 표현 다변화 등)을 적용해 후보 메시지를 탐색한다. 이처럼 다양하게 재구성된 쿼리를 통해 임베딩 검색 범위를 넓힘으로써, 단일 표현으로는 포착되지 않는 후보까지 포함할 수 있다.

(3) Selector

후보 메시지를 평가하고 최종 후보를 선정하는 모듈로, 세 가지 하위 Selector(PersonaMatch, TypeMatch, RoleMatch)와 하나의 Main Selector 로 구성된다(그림 2).

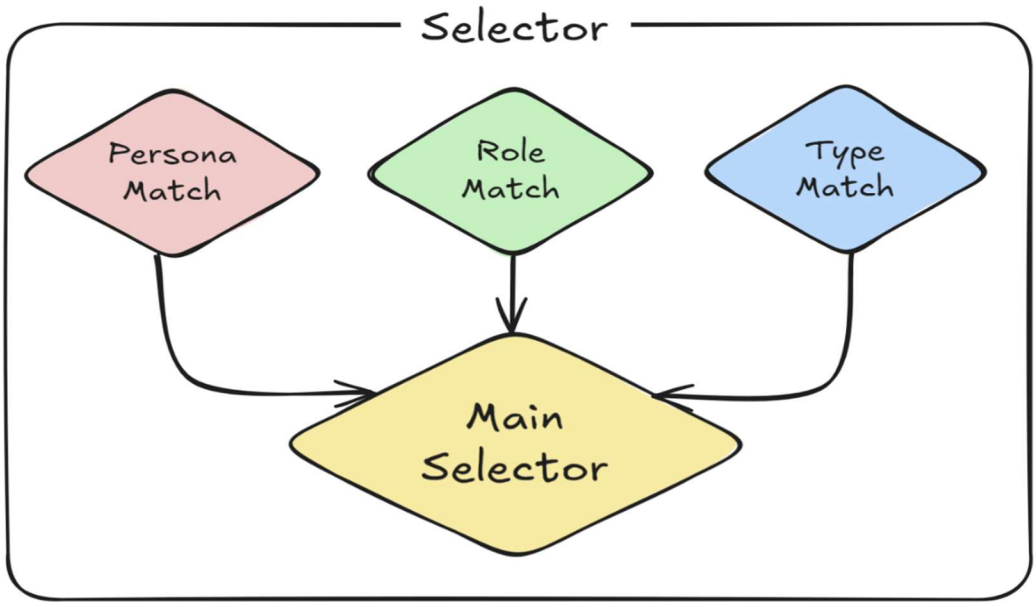

그림 2 Selector 구조

(3-1) PersonaMatch: 사용자 성향(명시적/암시적)을 추정하여 비슷한 페르소나를 가진 후보를 제안한다.

(3-2) TypeMatch: 메시지를 세 가지 유형(사람, 장소/서비스, 물건)으로 분류하여, 같은 Type 내에서만 후보를 제안함으로써 부정확한 매칭을 방지한다.

(3-3) RoleMatch: 사용자와 후보 간 역할 관계(역할 상보성 여부)를 분석하며, 단순 유사성 매칭이 아니라 상호작용 가능한 조합인지 판단한다.

(3-4) Main Selector: 세 하위 Selector 의 결과를 종합해 최종 후보 1 명을 선택한다. 페르소나 유사성, 유형 일치, 역할 상보성을 모두 고려하되, 적절한 후보가 없는 경우 Orchestrator 에게 후보 재구성을 요청할 수 있다.

(4) Evaluator

매칭 최종 결정자로서, Selector 가 제안한 후보의 적합성을 종합적으로 재검토한다. 상술한 판단 기준과 더불어 과거 실패 이력과 전체 문맥도 함께 고려한다. 최대 2 회 평가 기회를 가지며, 1 차 실패 시 Orchestrator 에 사유를 전달하고, 2 차에는 가장 수용 가능한 후보를 선택해 매칭을 종료한다.

(5) Message Analyzer

입력 메시지를 의미 중심으로 구조화하여 매칭에 필요한 핵심 요소를 식별하는 역할을 한다. 단순한 키워드 수준을 넘어서, 정서, 목적, 상호작용 대상 등을 문맥적으로 해석하며, 입력 메시지가 모호하거나 의도가 복수·불명확할 경우 이를 명확화하여 Matching Target 을 추론한다.

(6) Web Search Module

메시지 내 의미 불확실한 표현(예: 고유명사, 신조어 등)을 외부 정보를 통해 해석하고 보완하는 도구이다. Tavily API 를 통해 실시간 검색을 수행하며, 가장 적절한 의미와 관련 키워드를 제공해 매칭 정확도를 높인다.

전체 매칭 프로세스는 기본적으로 Orchestrator → Query Reformer → Selector → Evaluator 의 순서로 진행된다. 이 흐름을 따라 사용자의 입력 메시지가 분석되고, 관련 후보군이 추려지고, 이 중 가장 적절한 후보가 선택되어 최종 매칭이 이루어진다. 다만 메시지의 복잡성이나 표현의 모호성에 따라, Orchestrator 는 중간 단계에서 Message Analyzer 나 Web Search Module 을 전략적으로 호출하여 의미 해석을 보완하거나 외부 정보를 활용해 이해도를 높인다.

4. 시연 및 평가

4.1 개별 노드 작동 예시

본 절에서는 전체 매칭 흐름 중에서도 개별 노드의 기능과 역할을 부각하기 위해, Message Analyzer 와 Query Reformer 의 작동 사례를 예로 들어 설명하고자 한다. 그림 3 은 목적이 불분명한 메시지에 대해, 어떤 매칭 상대를 찾아야 하는지 Message Analyzer 가 구체화해주는 사례이다.

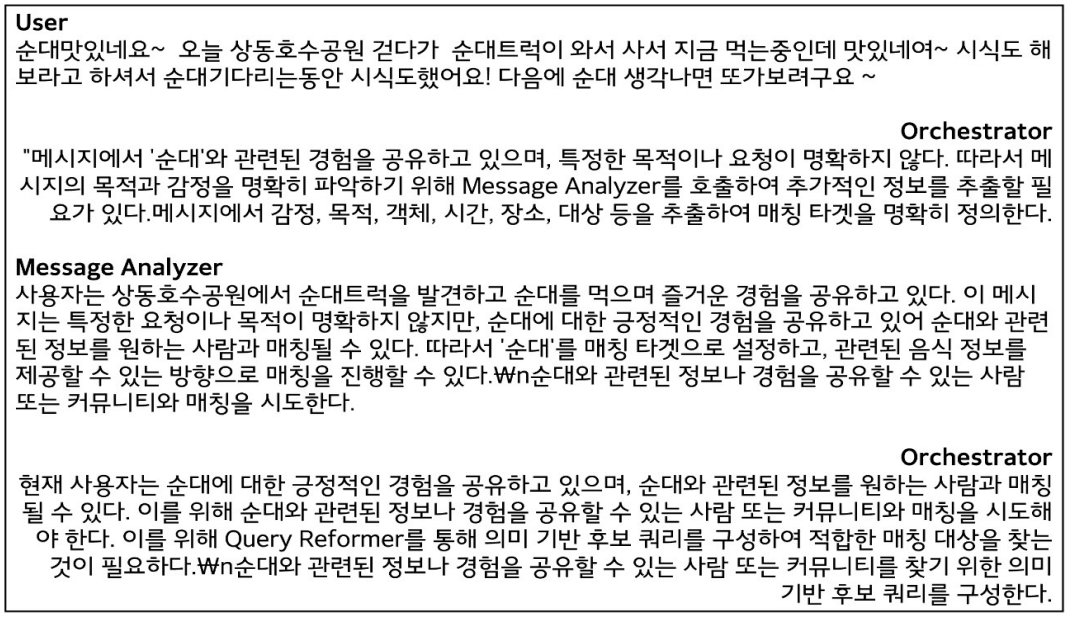

그림 3 Message Analyzer 예시

Query Reformer 는 input message 를 기반으로 '상보적인' 쿼리를 재구성하고 형태적 유사성을 맞춤으로써, 후보군을 탐색해 올 수 있다. 관련 연구에서 상술하였듯 벡터 기반 유사도 검색은 의미보다는 문장 형태에 기반하여 검색해오는 경우가 많으므로, 입력 쿼리 자체를 사용자의 메시지에 대해 상호보완적인 메시지로 변경할 경우 이에 대해 더 적절한

후보군을 탐색할 수 있다. 그림 4는 자신이 찾은 맛집을 추천하는 사용자의 메시지를, 추천이 필요한 사람처럼 재구성하는 Query Reformer 노드의 예시이다.

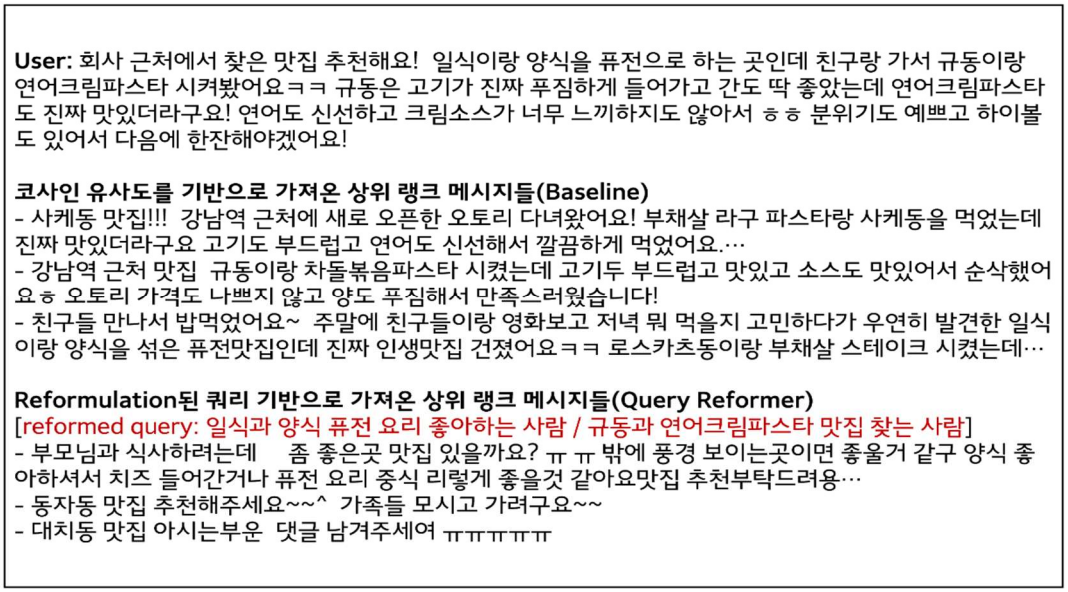

그림 4 Query Reformer 예시

4.2 Demonstration

본 절에서는 실제 사용자 메시지를 입력으로 하여, 제안한 시스템이 내부 노드 간의 협력적 상호작용을 통해 어떻게 의미 기반 매칭을 위한 판단을 수행하는지를 시나리오 형태로 시연한다. 사용자의 입력 메시지를 기점으로, 각 노드는 역할에 따라 정보를 분석·전달하며 순차적으로 의사결정에 관여한다. 이를 통해 제안 시스템이 비정형 표현에 대해 어떻게 Loose Matching 후보를 도출하는지 그 흐름을 구체적으로 살펴본다.

① 사용자 입력 (User)

“내일 2/14 노매뉴얼 플래그십스토어 가실분... 프레젠테이션 & 애프터파티까지 가시면 더 좋겠어여.”

② Orchestrator 판단

메시지 내 ‘노매뉴얼 플래그십스토어’가 고유명사로 해석됨

의미 불확실성 존재 → Web Search Module 호출

③ Web Search Module 결과

‘노매뉴얼 플래그십스토어’는 서울패션위크 기간 중 열리는 특정 브랜드 오프라인 행사

Matching Target: 패션 관련 행사 / 이벤트

도메인 및 참여 목적 명확화

④ 초기 후보 탐색 결과

후보: 사용자 ‘샤티’ - “타로 팝업스토어에 같이 가실 분”

Evaluator 판단: 부적합

(1) 도메인 불일치 (타로 vs 패션)

(2) 역할 상보성 없음

⑤ Query Reformer 재구성 쿼리 (일반화 기반)

쿼리 1: “노매뉴얼 플래그십 스토어 이벤트에 관심 있는 사람 찾습니다”

쿼리 2: “패션 행사에 관심 있는 분 찾습니다”

⑥ 후보 재탐색 및 최종 선택

최종 후보: 사용자 ‘다댬’

“본인에게 어울리는 스타일 찾고 싶은 분 계신가요? 저는 패션디자인과 전공생인데 패션 컨설팅 창업에 관심이 있어요! 혹시 옷을 사야하는데 고르기 어려우신 분, 나에게 어울리는 스타일을 잘 모르겠다 하시는 분 패션 컨설팅 도와드립니다!! 카페에서 스타일에 대해서 얘기 나누고 추후에는 같이 쇼핑도 가실 분~!”

⑦ Selector 판단 요약

도메인 정합성: 사용자의 메시지는 패션 브랜드 오프라인 행사 동행 요청이며, 후보자는 ‘패션 컨설팅’에 관심을 가진 전공생으로 패션 관련 주제에 높은 정합성을 보임

페르소나 유사성: 두 사람 모두 외향적이고 활동적인 성향을 나타내며, 스타일·패션에 대한 적극적 참여 의지를 공유

역할 상보성: 댜담은 패션에 대해 이야기하고 쇼핑을 함께할 사람을 찾고 있어, 사용자의 요청과 역할적으로 일정 수준의 상보성을 갖는다.

⑧ Evaluator 결정

최종 매칭 확정

Reasoning: 후보자는 패션에 대한 전문성과 대화 의지를 바탕으로 단순 동행이 아닌 주제 중심의 상호작용이 가능한 상대로 판단됨

이 시연은 사용자의 입력 메시지가 명확한 ‘동행 요청’이었음에도 불구하고, 단순한 키워드 일치를 넘어 정서적 표현, 역할 관계, 관심사 정합성까지 포괄하여 의미 기반 매칭이 이루어진 사례다.

특히, Web Search Module 을 통해 고유명사의 의미를 명확화하고, Query Reformer 가 표현을 일반화하면서 더욱 넓은 후보군 탐색이 가능해졌다. 이후 Selector 의 세부 평가를 통해 적합한 후보를 도출하고, Evaluator 가 정서적·상호작용적 관점에서 최종 매칭을 확정함으로써 Loose Matching 의 유연성과 실효성을 입증하였다.

4.3 평가

본 절에서는 제안 시스템의 평가를 위해, 두 가지 유형의 실험을 구성하였다. 첫 번째 실험은 구조화되지 않은 단방향 요청 메시지를 기반으로, 비정형 표현 간의 의미 기반 매칭 가능성을 검토하는 데 초점을 둔다. 반면 두 번째 실험은, 매칭 이론에서 핵심으로 간주되는 '쌍방 수용 가능성' 개념을 메시지 쌍에 적용해 본 것이다. 즉, 매칭 에이전트가 각 메시지를 의미 기반 기준에 따라 상대적으로 적합한 연결 후보로 간주했을 때, 그 판단이 쌍방에서 동시에 성립할 수 있는지를 평가하고자 한 실험이다. 이에 실험에는 두 종류의 데이터셋이 사용되었다. 첫 번째는 상술한 약 17,000 건의 자연어 메시지셋이며, 구조화되지 않은 단방향적 요청이 다수를 차지한다. 두 번째는 34 명의 남성과 34 명의 여성이 연애 상대 매칭을 위해 작성한 것으로 가정된 총 34 쌍의 정합성 기반 메시지 쌍이다. 본 실험은 모든 참여자가 짝을 갖는 안정적 매칭의 조건을 고려하여, 각 메시지 쌍이 상호 선택 가능성을 갖는 조합인지 제안 시스템의 의미 기반 판단에 따라 평가하는 방식으로 설계되었다.

첫번째 데이터셋을 통해, 기존 임베딩 기반 검색 및 단일 LLM 으로 구성된 Baseline 과 제안 시스템의 성능을 비교하는 실험을 진행하였다. Baseline 이 매칭에 실패한 80 건의 난이도 높은 메시지들을 대상으로 제안 시스템을 적용한 결과, 약 49%에서 더 설득력 있는 매칭이 도출되었다. 특히 역할 상보성이나 페르소나 유사성 등의 기준을 고려했을 때, 단순 유사성 기반 구조보다 더 정교한 판단이 가능함을 확인하였다. 가령 '중국어를 알려달라' 는 메시지에 대해, Baseline 은 관심사 일치를 근거로 들어 똑같이 중국어 잘하는 친구가 생겼으면 좋겠다는 메시지를 매칭하였다. 반면 제안 시스템은 중국어 재능 기부 목적의 후보를 선별하여, 사용자가 실질적으로 상호작용이 가능한 후보를 매칭하였다.

또한 두번째 데이터셋을 통한 실험 결과, 총 34 쌍 중 15 쌍(약 44%)이 의미 기반 기준에 따라 양측 모두에게 수용 가능하다고 판단되었으며 이를 통해 제안 시스템이 상호 적합한 연결의 일부를 형성할 수 있음을 확인하였다. 다만, 의미적으로 유사한 후보가 복수 존재하는 경우에는 우선순위 판단이 일관되지 않거나 특정 쌍이 명확히 도출되지 않는 한계가 있었으며, 이는 향후 후보 정렬 방식 및 선호 기준의 명확화 필요성을 시사한다.

5. 결론 및 논의

본 논문은 비정형 자연어 기반의 Loose Matching 문제를 다루며, 이를 해결하기 위한 구조적 접근으로 역할 분리형 RAG-LLM 기반 매칭 에이전트를 설계하고 실제 메시지를 통한 시연 및 평가를 수행하였다. 기존 키워드 중심 검색이나 단일 모델 기반 시스템이 처리하기 어려웠던 표현 다양성과 맥락 해석 문제에 대해, 역할 분리와 판단 기준 세분화를 통해 일정 수준의 대안을 제시할 수 있음을 확인하였다.

본 논문이 보여준 기여는 다음 두 가지이다. 첫째, 의미 기반 연결이 필요한 Loose Matching 환경에서 단순 유사성이 아닌 Matching Target 의 상위 개념, 표현 명시성, 역할 상보성, 대상 유형 구분 등을 기준으로 삼아 판단할 수 있는 틀을 구성하고, 이를 설계에 반영해본 것이다. 이는 직관에 기반한 매칭을 기술적으로 구현할 수 있는 하나의 방향성을 제시한다. 둘째, 단일 모델에 판단을 집중시키는 방식 대신, 다중 노드 구조에서 매칭 판단 기능을 역할별로 분산하여 해석-재구성-탐색-평가가 연속적으로 이루어지도록 한 점이다. 매칭 에이전트가 복잡한 문맥 해석이나 비정형 메시지 연결에서 유연성을 높일 수 있는 구성으로 작동 가능함을 일부 시연하였다. 이러한 제안 시스템의 구조는 단순 키워드 일치에 기반한 기존 추천·검색 시스템이 가진 맥락 해석의 한계를 보완하며, 온라인 플랫폼, 고객지원, 파트너 매칭, 내부 업무지원 시스템 등 다양한 비정형 정보 흐름 속에서 실질적인 연결 효율을 높일 수 있는 가능성을 보여준다.

그러나 본 시스템에는 한계가 존재한다. 우선, 현재 구조는 여전히 표현 유사성에 대한 의존도가 높아, 표현 방식이 다르더라도 의미적으로 연결 가능한 메시지를 제대로 포착하지 못하는 사례가 나타난다. 또한 복수의 수용 가능한 후보가 존재하는 경우, 그중 누구를 우선적으로 선택할 것인지에 대한 판단 기준은 불명확하다. 이는 본 시스템이 구조적으로 부적합한 후보를 걸러내는 데 최적화되어 있기 때문이며, 상대적으로 정밀한 선호 판단이나 후보 간 비교에는 제한이 존재한다. 이러한 구조는 도메인이 불분명하거나, 탐색적 매칭이 요구되는 환경에서는 유용할 수 있다. 그러나 특정 도메인에서 정확한 매칭과 제약조건 간 정밀한 매칭이 요구되는 경우, 우선순위 선정 및 제약 조건 판단의 정교함이 더 중요해진다. 또한 두 번째 실험처럼 쌍방의 선호를 동시에 고려해야 하는 매칭 상황에서는 해당 방식이 상대적으로 취약할 수 있다. 마지막으로, 판단 기능을 역할별로 분리한 구조는 복잡한 의미 해석을 정교하게 수행하는 데는 유리하나 그에 따른 LLM 호출 빈도 증가와 연산 비용 부담이라는 현실적인 제약을 동반한다. 특히 고성능 LLM 을 다수 호출해야 하는 구조적 특성상, 실제 응용에서는 효율성과 비용 간 균형을 고려한 설계 보완이 요구된다.

향후 연구는 다음 두 가지 방향으로의 확장이 가능하다. 첫째, 현재 시스템은 1:1 메시지 매칭에 초점을 두고 있지만, 실제 온라인 커뮤니티나 플랫폼 환경에서는 다자간 연결 역시 일반적이다. 예를 들어, 스포츠 팀 모집이나 공동 활동의 경우처럼, 단일 후보가 아니라 역할 분산, 성향 다양성, 상호 보완성 등을 고려한 그룹 단위 매칭 구조가 요구된다. 이는 단순 메시지 간 연결을 넘어 집단 내 의미적 정합성과 구조적 균형까지 고려해야 한다. 둘째, 매칭 이론에서 논의되는 다대다 매칭이나 양방향 안정적 매칭 문제를, 자연어

기반 입력만으로 얼마나 실현할 수 있는지에 대한 연구가 요구된다. 사용자의 선호 순위, 연결 제약 조건 등을 자연어로부터 정밀하게 추출하고, 이를 LLM의 프롬프트 설계와 판단 흐름 안에 녹여내는 방식으로 언어 기반의 매칭 알고리듬 구조를 구현할 수 있을 것으로 예상된다.

참고문헌

[1] A. E. Roth, “The economics of matching: Stability and incentives,” Mathematics of Operations Research, vol. 7, no. 4, pp. 617-628, Nov. 1982, doi: 10.1287/moor.7.4.617.

[2] Amazon Web Services, “Scaling Rufus: The Amazon generative AI-powered conversational shopping assistant with over 80,000 AWS Inferentia and AWS Trainium chips for Prime Day,” AWS Machine Learning Blog, May 2024. [Online]. Available: https://aws.amazon.com/ko/blogs/machine-learning/scaling-rufus-the-amazon-generative-ai-powered-conversational-shopping-assistant-with-over-80000-aws-inferentia-and-aws-trainium-chips-for-prime-day/

[3] Y. Wang, S. Lai, W. Zhang, Y. Liu, and S. Zhang, "ReMatch: Retrieval-enhanced schema matching with LLMs," Pattern Recognition and Artificial Intelligence, vol. 1, no. 1, pp. 1-10, 2024. [Online]. Available: https://doi.org/10.1007/s44336-024-00009-2

[4] Z. Liu, M. Zhang, and J. Xu, "Schema alignment with LLMs: A thought prompting approach," preprint, Preprints.org, May 2025. [Online]. Available: https://www.preprints.org/manuscript/202502.0406/v1

[5] P. Steiner, “Economy as matching,” Política & Sociedade, vol. 18, no. 43, pp. 14-45, Sep./Dec. 2019.

[6] J. Peeters, D. Weyns, and S. Heymans, “LEAPME: Learning-based Property Matching with Embeddings,” 2023 IEEE International Conference on Big Data (Big Data), Sorrento, Italy, 2023, pp. 4458-4467. doi: 10.1109/BigData58614.2023.10384108.

[7] H. Xu, J. Tang, Y. Qu, and Z. Li, "DeepAlignment: Unsupervised ontology matching with refined word vectors," in Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI), Stockholm, Sweden, Jul. 2018, pp. 1906-1912.

[8] M. J. Kim, T. Rekatsinas, and J. Fan, “Match, Compare, or Select: An Investigation of Large Language Models for Entity Matching,” Proceedings of the VLDB Endowment, vol. 16, no. 9, pp. 2183-2196, 2023.

[9] J. Xu, T. Wang, X. He, Z. Liu, and M. Zhang, "Matchmaker: Schema matching with self-improving compositional LLM programs," preprint, OpenReview, Mar. 2024. [Online]. Available: https://openreview.net/forum?id=vR2MWaZ3MG

[10] Y. Wang, Y. Liu, S. Lai, W. Zhang, and S. Zhang, "MeTMaP: Metamorphic testing for detecting false vector matching problems in LLM augmented generation," in Proc. 17th ACM Int. Conf. Web Search Data Min. (WSDM), Mérida, Mexico, 2024, pp. 1166-1176. [Online]. Available: https://doi.org/10.1145/3650105.3652297

[11] Y. Liu, P. Huang, C. Zheng, L. Ding, L. Wang, D. Xiong, D. Tao, and Z. Tu, "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2301.00375, 2023. [Online]. Available: https://arxiv.org/abs/2301.00375

[12] K. Lin, J. Wang, M. Yu, H. Zhang, S. Wang, and J. Callan, "Analyze, generate and refine: Query expansion with LLMs for zero-shot open-domain QA," in Findings Assoc. Comput. Linguist.: ACL 2024, pp. 10572-10592. [Online]. Available: https://aclanthology.org/2024.findings-acl.708/

[13] K. Hashimoto, E. M. Ardehaly, Y. Yamaguchi, K. Matsubara, N. Takemoto, Y. Tsuruoka, and Y. Miyao, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," in Proc. 61st Annu. Meeting Assoc. Comput. Linguist. (ACL), pp. 9983-10000, 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.557/.

[14] B. Rahdari, H. Ding, Z. Fan, Y. Ma, Z. Chen, A. Deoras, and B. Kveton, "Logic-Scaffolding: Personalized aspect-instructed recommendation explanation generation using LLMs," in Proc. 17th ACM Int. Conf. Web Search Data Min. (WSDM), Mérida, Mexico, 2024, pp. 1078-1081. doi: 10.1145/3616855.3635689.

[15] M. Suzgun and A. T. Kalai, "Meta-prompting: Enhancing language models with task-agnostic scaffolding," arXiv preprint arXiv:2401.12954, 2024. [Online]. Available: https://arxiv.org/abs/2401.12954

[16] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," arXiv preprint arXiv:2308.08155, 2023. [Online]. Available: https://arxiv.org/abs/2308.08155