Search

Updated Oct 22

[NL] Open IE Tasks

KDD Lab NL Research: Open Information Extraction (OIE)

KDD Lab Natural Language (NL) Division

Overview

Open Information Extraction from Social Media for STEM Discourses

Summary by GPT-5

> *This project builds an AI-driven pipeline that gathers and learns from public social media conversations, videos, and posts to support scientific research and reasoning. The first step—**data collection from social platforms**—automatically finds and previews relevant posts, videos, and podcasts about a topic (e.g., mental health, climate change, or AI ethics). The system then ranks useful sources (**Learning to Rank**), discovers and matches important terms (**Terminology Extraction**), identifies key facts or tables (**Table Extraction**), and links entities and relationships (**Relationship Extraction**). Multimodal **video-to-text (v2t)** tools make visual and spoken content searchable. A **Smart Crawler** continuously refines what to collect next, building a living dataset for transparent, trustworthy AI reasoning and scientific discovery.*

For ground truth rubrics and measures, see Open IE Tasks Ground Truth

Prospective Publication Venues

To guide literature reviews, research planning, preparation of papers, development of funding proposals, and selection of funding opportunities based on alignment of this work with program solicitations, current and recently funded work, etc.

Index

Task Definitions

Development

Literature Review

Prompt to Copilot:

Act as my research coordination and planning assistant. I am a project manager for a machine learning lab. We are developing a proof of concept test bed for open information extraction (Open IE or OIE), to support the development of multiple initial papers on aspects of the problem of "Open Information Extraction from Social Media for STEM Discourses", along with a project prospectus for facets of this research.

First, read our project mission statement:

"*This project builds an AI-driven pipeline that gathers and learns from public social media conversations, videos, and posts to support scientific research and reasoning. The first step—data collection from social platforms—automatically finds and previews relevant posts, videos, and podcasts about a topic (e.g., mental health, climate change, or AI ethics). The system then ranks useful sources (Learning to Rank), discovers and matches important terms (Terminology Extraction), identifies key facts or tables (Table Extraction), and links entities and relationships (Relationship Extraction). Multimodal video-to-text (v2t) tools make visual and spoken content searchable. A Smart Crawler continuously refines what to collect next, building a living dataset for transparent, trustworthy AI reasoning and scientific discovery.*"

Second, consider the following task definition:

- **Task 1. Data collection from social platforms** - scraping API

- **Task 2. Learning to rank** - search box: includes more-like-this similarity, roles
- **Task 3. Terminology extraction and matching** - entity list, synset expansion, topic modeling
- **Task 4. Table extraction** - populate a specified flat data model; one LLM baseline for filling fields
- **Task 5. Relationship extraction (REL) for instance population** - specific fields (slot-filler) for each post
  - a) to be checked against formal models
    - i. knowledge graph or other verifiable reference - for logical consistency
    - ii. semistructured data models - for (weaker) type consistency
  - b) outputs post records with automatically annotated fields
- **Task 6. Multimodal video-to-text (v2t) tasks** - for post transcripts with out-of-vocabulary terminology
  - a) automatic speech recognition (ASR
  - b) optical character recognition (OCR)
  - c) descriptive video captioning: music titles, genre, style; ambient images, backgrounds; actions
- **Task 7. Smart crawler** - continuously operating with scheduler

Third, consider the following conference and journal venues I suggested for publications:

- **Overall pipeline:** CIKM
- Tasks **1, 2, 7**: SIGIR = learning to rank (LTR) & smart crawler
- Tasks **3, 4, 5**: domain-specific journal, KDD workshop or conference
- **Data collection framework** including hosted service design, LTR: TBD (systems-oriented conference on web data mining - possibly IUI or The Web Conference)
- **Social platform monitoring** (use case of relation extraction): ICWSM, ASONAM (due to discontinuation of SocInfo)
- Task **6** (multimodal LLMs): ACII (affective computing)
- Personalization of platform crawler (**Task 1 DC/precursor**): UMAP

Fourth, look through the DBLP tables of contents for venues we are considering to set expectations for the literature reviews we are conducting, giving critical feedback on our experiment design, and making suggestions for a follow-up scouring of WikiCFP and replanning of the paper for alternate venues.

Fifth, plan an order of paper development based not only on chronological due dates of papers during the coming 1-2 years but based on dependencies among the above planned system components. Reason carefully so as to have a contingency plan if we decide to hold the paper for a later submission, or the paper is rejected with reviews that you can subsequently analyze.

The general goal of this collaboration is to target the top AI and ML conferences: NeurIPS, ICML, ICLR, AAAI, and IJCAI, and secondarily ACL, EMNLP, and the "continental" regional conferences NAACL-HLT, AACL, and EACL. A major part of this work is aimed at WSDM and SIGIR, or secondarily RecSys. Workshops of any of these are acceptable venues if and only if the work is of too narrow or specialized a scope AND the novelty is insufficient for the main conference. User-modeling conferences such as CHI, vision conferences such as CVPR, ICCV, BCMV, computational linguistics conferences beyond ACL and EMNLP, ontology venues such as ISWC and ISWJ, are of interest if the work is VERY well aligned. Target those venues (conferences for cutting-edge work or journals for more mature work) if we have identified some facet in which results are an improvement on the state-of-the-art or represent a fundamental, substantive, and interesting advance over SOTA-competitive methods, and a better match for that conference than the corresponding tracks at top conferences.

Finally, specialty conferences such as ICWSM for social media, ASONAM and NetSci for heterogeneous information networks, K-CAP for information capture, ICDAR for document analysis conferences and recognition, ICASSP and Interspeech for speech recognition and analysis, SIGMOD/PODS for data modeling, HRI and RO-MAN for automation or human-in-the-loop mixed-initiative systems, AIES, FAccT, EAAMO for algorithmic fairness, IAAI for innovative applications, EAAI for educational applications such as public service educational videos and ads, plus irregularly held conferences similar to the now-defunct SocInfo, the apparently concluded Social Influence workshop series at IJCAI, the eCommerce workshop at SIGIR, or the sporadic SARA, would all be good venues for a relevant paper.

Prioritize deliberate research planning for highest impact: it is much more important to us to make scientifically meaningful, significant, and worthwhile contributions than to simply maximize bibliometrics such as impact factor, citation count, etc. These are of value only as a direct or side effect of significance, not in their own right.

When you have analyzed the above, tell me and ask any questions, and I will respond with a prompt for one of the specific tasks from my second instruction to you above and/or papers from my third instruction.

Response from Copilot:

Thanks, William — I've carefully reviewed your mission statement, task breakdown, venue strategy, and publication planning goals. Here's a summary of how I'm organizing this as your research coordination and planning assistant:

Project Mission Summary

You're building a **modular, AI-driven pipeline** for Open Information Extraction (OIE) from social media, with a focus on **STEM discourse**. The pipeline includes:

- **Multimodal data collection** (text, video, audio)
- **Terminology and relationship extraction**
- **Table and entity linking**
- **Smart crawling and personalization**
- **Multimodal transcription and captioning**
- **Scientific reasoning support via verifiable models**

Task Breakdown and Dependencies

Here's a dependency-aware view of your system components:

1. **Task 1: Data Collection**
   - Foundation for all downstream tasks.
   - Dependencies: API access, crawler scheduling, personalization (Task 7).
2. **Task 2: Learning to Rank**
   - Enhances search and retrieval.

Depends on Task 1 outputs; feeds into Task 7.

3. **Task 3: Terminology Extraction**
   - Supports Tasks 4, 5, and 6.
   - Requires initial data from Task 1.
4. **Task 4: Table Extraction**
   - Needs terminology and structured data (Task 3).
   - Can be prototyped with LLM baselines.
5. **Task 5: Relationship Extraction**
   - Depends on Tasks 3 and 4.
   - Outputs structured post records for reasoning.
6. **Task 6: Multimodal Video-to-Text**
   - Parallel to Task 1, but feeds into Tasks 3 and 5.
   - ASR, OCR, and captioning are subcomponents.
7. **Task 7: Smart Crawler**
   - Depends on feedback from Tasks 2 and 5.
   - Continuously improves Task 1.

Venue Mapping

Here's how your tasks align with publication venues:

Table under this box goes here

Literature Review Planning

I'll search DBLP for recent papers from:

- **SIGIR**, **WSDM**, **CIKM**, **KDD**, **ICWSM**, **ASONAM**, **ACII**, **ACL**, **EMNLP**, **AAAI**, **IJCAI**, **NeurIPS**, **ICML**, **ICLR**, **ISWC**, **ICASSP**, **Interspeech**, **ICDAR**, **CHI**, **IUI**, **UMAP**, **K-CAP**, **NetSci**, **FAccT**, **EAAMO**, **AIES**, **EAAI**, **HRI**, **RO-MAN**

This will help us:

- Set expectations for novelty and scope
- Refine experiment design
- Identify gaps and opportunities
- Plan alternate venues via WikiCFP

Paper Development Order (Dependency + Deadline Aware)

Phase 1: Foundational Papers

Task 1 + Task 7: Smart Crawler and Data Collection Framework
→ Target: **SIGIR**, **WSDM**, **The Web Conf**, **UMAP**

- **Task 2**: Learning to Rank
  → Target: **SIGIR**, **CIKM**

Phase 2: Extraction and Modeling

- **Task 3 + Task 4**: Terminology and Table Extraction
  → Target: **KDD**, **CIKM**, **ACL workshops**
- **Task 5**: Relationship Extraction
  → Target: **ASONAM**, **ICWSM**, **ISWC**

Phase 3: Multimodal and Reasoning

- **Task 6**: Multimodal v2t
  → Target: **ACII**, **ICASSP**, **Interspeech**, **ICDAR**
- **Integrative Paper**: Full pipeline + reasoning
  → Target: **CIKM**, **AAAI**, **IJCAI**, **NeurIPS**

Contingency Planning

- If rejected: analyze reviews, revise scope, and resubmit to:
  - Workshops of the same venue
  - Alternate venues with better alignment
  - Journals for mature work (e.g., **ISWJ**, **JMLR**, **TKDE**)

Next Steps

To proceed, I'll:

1. Search DBLP for recent papers in the venues above.
2. Help you plan literature reviews and feedback loops.
3. Assist with WikiCFP scouring and alternate venue planning.

Would you like me to begin with the DBLP search now, or would you prefer to move directly into one of the specific tasks or papers you mentioned?

Task	Venue(s)	Notes
1, 2, 7	**SIGIR**, **WSDM**, **The Web Conf**, **UMAP**	Core IR and personalization
3, 4, 5	**KDD**, **CIKM**, **ACL workshops**, **ISWC**	Extraction, modeling, and reasoning
6	**ACII**, **ICASSP**, **Interspeech**, **ICDAR**	Multimodal and affective computing
Use Cases	**ICWSM**, **ASONAM**, **NetSci**, **K-CAP**	Social media and heterogeneous networks
Overall pipeline	**CIKM**, **AAAI**, **IJCAI**, **NeurIPS**	Integrative papers

Systems	**IUI**, **The Web Conf**, **SIGMOD/PODS**	Hosted services and smart crawling

TBD: Josh Garcia

1. Data collection from social platforms

Summary by GPT-5

> *This task gathers publicly available posts, videos, and podcasts from platforms like YouTube, Reddit, and X/Twitter that relate to a scientific or ethical topic. Using text and multimedia analysis, it translates speech, captions, and visual text into searchable text fields. The goal is to capture diverse public perspectives, examples, and informal reasoning. These posts form a living dataset for downstream AI learning. By organizing this content, we enable deeper study of how people discuss complex issues —helping AI systems learn from real-world dialogue rather than static documents.*

Task page: [NL] OIE Data Collection

Research Planning

Team**:** James Chapman (lead), Jahid Choton, Chipo Sekabanja

Literature Review: Data Collection

Development: Data Collection

2. Learning to Rank

Summary by GPT-5

> *After collecting candidate posts, this step teaches the AI to prioritize the most relevant, informative, and trustworthy sources. It uses both human feedback and automated similarity measures to score posts based on clarity, originality, and connection to research questions. By learning to rank posts effectively, the system becomes selective—surfacing high-quality examples for reasoning and summarization while filtering noise. This mirrors how scientists scan literature for credible evidence, enabling the AI to identify which posts or discussions most strongly advance understanding within a domain like neuroscience or climate science.*

Team Lead: Mobina Golmohammadi
Other Team Members: Bruce Jia, Yihong Theis, *KDD developer TBD* (tentatively Tatenda Sekabanja)

Learning to Rank (LTR) *aka* Machine Learned Ranking (MLR) - *material to import from task thread of Teams channel*

search box: includes more-like-this similarity, *roles*

Literature Review: Learning to Rank (LTR)

Development: Learning to Rank (LTR)

material to import from task thread of Teams channel

substance list, *synset expansion*

3. Terminology extraction and matching

Summary by GPT-5

> *This step identifies key scientific or thematic terms used in the collected posts and aligns them with known vocabularies or ontologies. For example, it can detect when "microbiome health" or "carbon offset" appears under slightly different names. Matching terms across sources supports consistent labeling and helps track how the public and experts use language differently. The result is a structured glossary that evolves with new discussions. This process strengthens both machine and human comprehension, enabling better retrieval, summarization, and interpretation of emerging ideas in complex interdisciplinary fields.*

Team Leads: Timothy Tucker and Jahid Choton
Other Team Members: Manny Adeniji, James Chapman

Literature Review: Term Extraction

Development: Term Extraction

Specs from William Hsu - Tue 07 Oct 2025

- prompt link here
- GHP IE progress tracking chat (for spreadsheet & wiki updates)
- KDD Dev channel

4. Table extraction

Summary by GPT-5

> *Many online posts, papers, and videos present data in tables, charts, or captions. This task automatically extracts such structured information—like survey results, nutritional data, or climate indicators—from text or images. By converting these visuals into standardized data tables, we make quantitative evidence searchable and comparable across contexts. Extracted tables can then feed models for reasoning, visualization, and simulation. This helps scientists and students explore "found data" from the web while maintaining links to its sources, supporting reproducibility and interpretability in AI-assisted research.*

Team Lead: Bruce Jia *(tentative)*
Other Team Members: Jahid Choton, Patrick Stingley

Specs from William Hsu - Tue 07 Oct 2025

- material to import from task thread of Teams channel populate a specified flat data model
- one LLM baseline for filling fields

Literature Review: Structured Information Extraction

Development: Structured Information Extraction

5. Relationship extraction (REL) for instance population

Summary by GPT-5

> *Relationship extraction identifies how entities (people, concepts, or objects) connect within posts—for example, linking "vaccination" to "reduced infection risk" or "carbon emissions" to "temperature rise." This task populates a knowledge graph where nodes represent entities and edges represent their relationships. Over time, this forms a living map of public reasoning and evidence. Such structured representations allow AI models to explain why certain claims are made or debated. In essence, REL transforms scattered online statements into machine-readable scientific arguments that can be queried, verified, and built upon.*

Team Lead: Yihong Theis *(tentative)*
Other Team Members: Ahsan Zaidi

Specs from William Hsu - Tue 07 Oct 2025

Material to import from task thread of Teams channel

specific fields (slot-filler) for each post

1. to be checked against
   a. **neuroscience ontology (for E-R consistency)**
   b. scientific data ontology (for type consistency)
   c. affective content ontology (social network and sentiment analysis)
2. outputs post records with automatically annotated fields

multivariable Bayesian conditioning - see: Most Probable Explanation (MPE) task

Literature Review: Relationship Extraction and Probabilistic Reasoning

Development: Relationship Extraction and Probabilistic Reasoning

6. Multimodal video-to-text (v2t) tasks

Summary by GPT-5

> *This task converts spoken, visual, and on-screen text from videos into structured textual descriptions. Using speech recognition, image captioning, and optical character recognition (OCR), it translates multimedia posts into data that can be searched and analyzed like text. It captures gestures, emotions, sound effects, and narrative flow—turning videos into analyzable "dialogue" objects. This allows AI systems to study how reasoning unfolds visually and verbally, such as in debates, interviews, or explainer videos. For scientists, it unlocks a rich new source of public reasoning and informal communication for analysis.*

Team Lead: Patrick Stingley *(tentative)*
Other Team Members: Bruce Jia, David Kurtenbach, Chipo Sekabanja

material to import from task thread of Teams channel

for post transcripts with out-of-vocabulary terminology

1. automatic speech recognition (ASR)
2. optical character recognition (OCR)
3. descriptive video captioning: music titles, genre, style; ambient images, backgrounds; actions

Literature Review: Video-to-Text (v2t) Tasks

Speech Recognition

Optical Character Recognition

Descriptive Video Captioning

- *Descriptive fMRI - **Bruce***

Development: Video-to-Text (v2t) Tasks

Speech Recognition Services

*Using and running them - link to transparent/autonomous service **and** API docs here*

Optical Character Recognition Services

*Using and running them - link to transparent/autonomous service **and** API docs here*

Descriptive Video Captioning Services

*Using and running them - link to transparent/autonomous service **and** API docs here*

material to import from task thread of Teams channel

continuously operating with scheduler

7. Smart Crawler

Summary by GPT-5

> *The Smart Crawler orchestrates the whole pipeline. It learns what kinds of content are most relevant and adjusts its searches accordingly. Using feedback from the ranking, terminology, and relationship-extraction stages, it refines queries to discover new, high-value posts automatically. This adaptive behavior makes the crawler a kind of autonomous research assistant—continuously exploring online discussions, identifying knowledge gaps, and gathering fresh examples for analysis. By closing the loop between data discovery and reasoning, it supports scalable, transparent, and domain-adaptable AI research for scientists in any field.*

Team Lead: William H. Hsu & David Kurtenbach
Other Team Members: Chipo Sekabanja; *outreach to Josh Weese (2013-2014) & review notes by Carter Nelson (2017-2019), ScraPy interface* by *Jordan Roth (c. 2019-2020)*

Literature Review: Smart Crawler

Development: Smart Crawler

References

This research work focuses on extracting procedural information in the form of recipes from published scientific literature with application to nanomaterials synthesis. From our overall goal of producing recipes from free text, we derive the technical objectives of a system consisting of pipeline stages: document acquisition and filtering, payload extraction, recipe step extraction as a relationship extraction task, recipe assembly, and presentation through an information retrieval interface with question answering (QA) functionality. This system meets computational information and knowledge management (CIKM) requirements of metadata-driven payload extraction, named entity extraction, and relationship extraction from text. Results, key novel contributions, and significant open problems derived from this work center around the attribution of these holistic quality measures to specific machine learning and inference stages of the pipeline, each with their performance measures.

Navigation

- **[NL] Chat | Channel | Team | Wiki**

- **[OIE] Chat | Channel | Plan | Lit | Dev | Wiki**

Return

- Open IE Project Home Page - [NL] Open Information Extraction (OIE)
- KDD Lab Natural Language (NL) Division Page - Natural Language
- KDD Lab Research Divisions Page - Divisions
- KDD Lab Research Page - Research
- **KDD Lab Home Page -** KDD Lab (private wiki)

Related content

Send feedback

[NL] Open IE Publication Plan
William Hsu

[NL] Open Information Extraction (OIE)
Kyle Draeger

[P&G] Research Project Template - WIP
William Hsu

Natural Language
Kyle Draeger

Timothy Tucker
William Hsu

Add a comment

Add a reaction