Assignment 3

Joshua Cook

Georgia Tech

CS 6460 Education Technology

Research Log

Background

We have arrived in Paris at the incubator with a narrow focus given the intense research I have been doing for the past two weeks. We have drafted a new statement for our company:

Chelle streamlines internal training, leading to field engineers who are continually up-to-date with company technicals, managers that are freed from creating training curricula and curating documentation, and teams that are empowered with a unified, easily accessible source of truth.

My current research is critical to this mission.

My research so far has led to three primary processes or AI agents that are critical to knowledge extraction: entity extraction, entity relation, and ontology mapping. These components will work together to create ontologies for each of our customer organizations. We will also maintain a set of public ontologies that can be broadly leveraged.

The theme of this week might be theory to practice. In this, I will focus on the following:

1. Named entity recognition. What models are best in class? What are the best strategies? What are assessment strategies for comparing different models/strategies?
2. LOOM (Lexical OWL Ontology Matcher). I'd like to learn more about this tool. It appears that it may have been sunset. What was it replaced with?
3. BoomerGPT. What is this tool for? Is it Python native?
4. On the topic of Python, do I think I will leverage existing packages or grasp the concepts and develop my own software.
5. Knowledge extraction as ETL.

6. Ontology curation
7. A deeper dive with "Agent-OM: Leveraging LLM Agents for Ontology Matching," Qiang, et al. (2025). c

Papers

Citation: Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., & Han, W. (2024). ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. Retrieved from https://arxiv.org/abs/2302.10205

How Found: In references of another paper

My Summary: Information extraction, using large language models; two stage framework improves the performance of LLMs in extracting entity-relation triples, named entities, and events

My Takeaway: Two stage multi return QA framework should be considered for application, explore the use of chain templates, look at interactive mode for R&D phase

Citation: Matthews, J., Love, P. E. D., Porter, S., & Fang, W. (2023). Curating a domain ontology for rework in construction: challenges and learnings from practice. Production Planning & Control. https://doi.org/10.1080/09537287.2023.2223566

How Found: Google Search

My Summary: Non-medical application application of ontology development, use of topic modeling, top down and bottom up methodologies,

My Takeaway: Developing a standardized class or can be useful across verticals, investigate the top down and bottom up approaches, these might be useful and understanding the field itself

Citation: Silva, J., Revoredo, K., Baião, F. A., & Lima, C. (2023). Enhancing Ontology Matching: Lexically and Syntactically Standardizing Ontologies Through Customized Lexical Analyzers. Semantic Web, IOS Press.

How Found: Google Search

My Summary: Customized lexical analyzers, Objective is to improve the quality of alignments between ontologies, creation of the ALIN metric to calculate similarity values

My Takeaway: The ALIN metric might be useful, if we do pursue more canonical approaches for entity, recognition, and others could be a useful resource of some of the models we could try, that said as soon as we get into the business of training models now we're in the business of training models

Citation: Darji, H., Mitrović, J., & Granitzer, M. (2023). German BERT Model for Legal Named Entity Recognition. University of Passau.

How Found: Google Search

My Summary: legal NER appplication with BERT, evaluate the performance of a fine tune model, compares to an industry standard model BiLSTM-CRF+

My Takeaway: The more I read about Bert the more I think we should stick to the approach of not using transformers, we are a small team we should focus on delivering value which can be done most efficiently using prompting

Citation: Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. arXiv preprint arXiv:2301.13294. https://arxiv.org/abs/2301.13294

How Found: Google Search

My Summary: Adaptation of LLM's to the task of named entity recognition, proposes a method that transforms the NER into a text generation task, uses a self verification strategy,

My Takeaway: Approach for converting a labeling task into a generation task using prompting, self verification is definitely something that should be explored, nearest neighbors approach is used at the end to increase accuracy worth exploring in terms of use of ensemble and complex DAGs

Citation: Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., & Kurohashi, S. (2023). GPT-RE: In-context Learning for Relation Extraction using Large Language Models. Kyoto University, Japan; Zhejiang University, China.

How Found: Google Search

My Summary: Using LLM's for relation extraction, proposes a method that incorporates task aware representations, comparison of several relation extraction models

My Takeaway: Explore the idea of task aware retrieval, likely fine-tuning is out of the scope of possible at this time, methods for enriching data here it looks like it's beyond simple RAG approachs

Citation: Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., Roberts, K., & Xu, H. (2024). Improving large language models for clinical named entity recognition via prompt engineering. Journal of the American Medical Informatics Association, 31, 1812–1820. https://doi.org/10.1093/jamia/ocad259

How Found: Google Search

My Summary: Use of LLM in clinical NER, proposes task, specific prompts,

My Takeaway: Several useful prompt engineering techniques, rigorous use of few shot learning for extraction look at strategies for R&D process

Citation: Sui, D., Zeng, X., Chen, Y., Liu, K., & Zhao, J. (2023). Joint Entity and Relation Extraction with Set Prediction Networks. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2023.3264735

How Found: Google Search

My Summary: Advanced techniques in NLP for extracting relational triples, proposes, prediction, models, using transformers, relaxes some of the limitations of previous approaches, strong performance at relation extraction

My Takeaway: Leverages use of transformer based models, again, likely not in the cards for right now, proposes a strategy for ontological mapping to relational triples

Citation: Ciatto, G., Agiollo, A., Magnini, M., & Omicini, A. (2023). Large language models as oracles for instantiating ontologies with domain-specific knowledge. Alma Mater Studiorum – Università di Bologna.

How Found: Google Search

My Summary: KGFiller hey framework to semi automatically populate ontologies with domain specific using LLM's. Uses what is inherent to the LLM from training. Builds an initial ontology schema through a series of query templates submitted to an LLM.

My Takeaway: Pre-populate, ontology, frameworks ahead of time, possibly not using this technique, but it's useful to look at it as a strategy, look at the use of query design templates, look at strategies for error mitigation

Citation: Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents: A Survey. ACM Computing Surveys, 56(2), Article 27. https://doi.org/10.1145/3604931

How Found: Google Search

My Summary: Challenges and techniques related to NER in historical texts, highlights how ontological mapping can be applied to a line named entities with predefined ontologies

My Takeaway: Older paper, it was investigated to provide some grounding in the field, very dated likely not to be useful

Citation: Reynolds, L., & McDonell, K. (2023). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. arXiv preprint arXiv:2307.00001.

How Found: Google Search

My Summary: Exploring the efficay of zero sharp prompts for a few shot learning with GPT, suggest that few shot examples help model locate pre-existing tasks within its trained knowledge, proposes new methods for prompt programming, emphasizing optics, narrative structures, and cultural references

My Takeaway: Strategies for effective prompting, use of met prompts, toward chain of thought, be aware of how you should examples can contaminate the semantics, in general, a good read if our emphasis is nearly completely on prompt

Citation: Ashok, D., & Lipton, Z. C. (2023). PromptNER: Prompting For Named Entity Recognition. Carnegie Mellon University.

How Found: Google Search

My Summary: Leveraging LLM's and prompt base techniques to improve any NER, requires a side of entity, definition, definitions, and examples, prompt to generate potential entities and explanations, requires lab data, but very little

My Takeaway: Develop modular definitions of entities, useful for understanding how others are using prompting, useful for examining how human in the loop review is used, useful for its use metrics

Citation: Wadhwa, S., Amir, S., & Wallace, B. C. (2023). Revisiting Relation Extraction in the era of Large Language Models. Proceedings of the Conference of the Association for Computational Linguistics Meeting, 2023: 15566–15589. https://doi.org/10.18653/v1/2023.acllong.868

How Found: Google Search

My Summary: Use of a large language models for relation extraction, proposes, a novel approach, choosing chain of thought, achieving near state of the art learning, shot examples are generated by one model, then passed to another

My Takeaway: Combination of two different models, one proposes few shot examples these are past to another model that does shot learning to do the relation extraction, has me thinking about combining GPT and Claude, not a bad idea in general, great for its techniques

Citation: Zhou, W., Zhang, S., Gu, Y., Chen, M., & Poon, H. (2023). UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. Conference Proceedings.

How Found: Google Search

My Summary: Use of distillation within large language models, mission, focused instruction tuning, have distilled a model from chat GPT

My Takeaway: Distillation of GPT, I have to read more about this, it does land in the Rome of cost optimization, which should not be a priority right now, and I guess if we do this now we're in the business of maintaining a model

Citation: Yuan, C., Xie, Q., & Ananiadou, S. (2023). Zero-shot temporal relation extraction with ChatGPT. Department of Computer Science, The University of Manchester.

How Found: Google Search

My Summary: Explore the youth of large language models for extracting temporal relations between events within a document. Evaluate three different prompting techniques: zero shot, event, ranking, and chain of thought, does not outperform on vanilla tasks, but does in finding outliers

My Takeaway: Zero shot learning, event, ranking, and chain of thought, useful for looking at how the use of different prompts was explored and how the prompts were compared rigorous research on the RN around prompt engineering

Synthesis

We are nearing the end of the research phase, and my thoughts are definitely crystallizing around what I think will be our approach for automated knowledge-extraction. Nearly all of the modern research (last two years), Sees a paradigm shift from traditional NLP to Large Language Model (LLM) based approaches. Pre-LLM era research predominantly pointed towards transformer-based BERT models for knowledge extraction tasks. What I'm seeing in this research is that the future lies in leveraging existing LLMs, especially those provided by major providers such as OpenAI, Anthropic, and Google. These models perform extremely well and from a business standpoint, I would argue the following:

1. These models are more cost-effective. BERT-based models are highly effective, but they are expensive to develop and expensive to maintain. Using LLMs provided by the major providers shifts the cost from engineering to inference. There are also no maintenance costs.
2. With LLM-based approaches, we will also be able to develop much more quickly. If our work is in prompt engineering, development cycles will be significantly faster.
3. Given that we do not, and frequently will not ever, know much about the semantics of the mappings we will be preparing for customers. We will need the flexibility provided by pre-trained LLM.

All of the research points at solving knowledge extraction problems through effective prompt engineering. Here are some key strategies and considerations for these efforts:

- **Multi-Stage Prompting:** Implementing a two-stage or multi-stage prompting framework can improve information extraction performance (Wei et al., 2024).
- **Task Transformation:** Converting extraction tasks into generation tasks has shown promise in NER (Wang et al., 2023).
- **Self-Validation:** Incorporating self-verification strategies within our prompts can enhance accuracy and reliability (Wang et al., 2023).
- **Prompt Comparison Framework:** Developing a robust methodology for comparing different prompts and prompting strategies is crucial for ongoing optimization (Hu et al., 2024).

- **Semantic Contamination Awareness:** Being mindful of how few-shot examples might inadvertently influence the semantic understanding of the model (Reynolds & McDonell, 2023).
- **Model Combination:** Exploring the potential of using multiple models in tandem, where one model generates few-shot examples for another to perform the actual extraction task (Wadhwa et al., 2023).

Reflection

My research process this week was a blend of building upon the previous and exploring new directions. Many of the papers that I used this week were already in my directory of papers from previous searches. I also did new searches, especially focused on named entity recognition (NER).

I am very excited this week by how familiar I am with the key concepts in this research and how my understanding has solidified over the past few weeks. When I began, the field seemed intimidating. Of course, I haven't actually implemented anything yet, but I feel like I have a pretty solid understanding of the ideas.

One of the most valuable aspects of my research workflow has been the integration of AI, specifically Claude, as a research assistant. I developed a technique of engaging in Socratic dialogues with Claude about the papers I've read. By asking Claude to question me about the papers, I was able to gauge my own understanding and identify areas that needed further exploration. This of course, leads to ideas for the platform itself, especially the instruction portion that I haven't really started working on yet. By this, I mean that the strategy that I am using to further my own understanding of research can be a strategy that we modify, and then deploy for our users to learn the concepts presented in our guides.

Planning

Here are my preliminary thoughts on the work that I will do:

1. Core Components:
    a. Applications
        i. Knowledge Extractor
            1. Application Tiers
                a. Modular system for swapping foundation models and prompts
                b. Single document extraction
                c. Diff application to existing ontology

        d. REST APIs for communication between components
    ii. Model Development and Evaluation
    iii. Prompt Strategy Library
  b. Models:
    i. Document Entity Extractor
    ii. Document Relation Extractor
    iii. Document Ontology Mapper
    iv. Ontology Entity Resolver
    v. Ontology Relation Resolver
    vi. Ontology-Ontology Resolver
2. LLM Integration:
  a. Flexibility to use OpenAI, Anthropic, and Google models
  b. Ability to select different models for each individual call
3. Prompt Engineering:
  a. Need for a process to develop prompting techniques for each application
  b. Requirement for a flexible encoding system for prompts
4. Workflow:
  a. Focus on processing text input to knowledge output
  b. Applying a "diff" to existing semantic mapping with each new document
  c. Potential need for an additional model for conflict resolution
5. Evaluation:
  a. further research required

Activity

Problem Statement

Background Information.

The field of knowledge extraction and ontological mapping is an area in natural language processing (NLP) and artificial intelligence (AI). For my purposes, this is important for organizations managing internal knowledge management and training processes. This domain focuses on automatically extracting structured information from unstructured text and organizing it into coherent, machine-readable knowledge. Once extracted, the generated ontological mapping

can be used to support additional activities, such as just-in-time information retrieval or automated training.

Key concepts in this area include:

1. Ontology
    a. A formal representation of a set of concepts within a domain and the relationships between them. Ontologies provide a shared vocabulary for modeling a domain.
2. Entity Extraction
    a. The process of identifying and classifying key information (entities) in text into predefined categories such as person names, organizations, locations, etc.
3. Relation Extraction
    a. The task of detecting and classifying semantic relationships between entities in text.
4. Ontology Mapping
    a. The process of finding correspondences between concepts in different ontologies.
5. Knowledge Base
    a. A technology used to store complex structured and unstructured information used by a computer system.

The challenge lies in developing efficient, accurate, and scalable methods to extract knowledge from diverse document sources, map this knowledge to existing ontologies, and present it in a form that is useful for non-expert users. This involves not only technical challenges in AI and NLP but also considerations of user experience, data integration, and the specific needs of different organizational contexts.

General Problem Statement.

Organizations struggle to efficiently extract, organize, and utilize knowledge from their unstructured textual data. This challenge encompasses:

- Information Overload - the volume of documents and data within organizations is growing exponentially.
- Knowledge Silos - important information is often scattered across different departments, systems, versions, documents
- Manual Curation - traditional methods of knowledge organization rely heavily on manual effort

- Lack of Standardization - challenging to integrate information from diverse sources
- Rapid Obsolescence - knowledge can quickly become outdated, requiring constant updates
- Domain Specificity - each organization or industry has its own specialized vocabulary and concepts solutions.
- Accessibility - even when knowledge is properly organized, it may not be easily accessible
- Scalability - knowledge management systems often struggle to keep pace with growth

Scholarly Support.

- Information Overload:
    - Data creation and replication will grow to an annual rate of 23% through 2025, reaching 181 zettabytes (Taylor, P, 2023).
    - Knowledge workers spend 47% of their time searching for information (Gartner, 2023).
- Knowledge Silos
    - 55% of organizations report that knowledge silos are a significant barrier to effective decision-making (Deloitte, 2020).
- Manual Curation
    - Workers spend 19% of their time searching for and gathering information (Chui, M. et al., 2012)
- Lack of Standardization:
    - Organizations using 7-10 information management systems or repositories have gone from 3.6% in 2013 to 6.2% in 2018 and now 14.42% in 2023. This represents approximately 100% growth every five years in the number of systems. Additionally, interoperability is a challenge with expanding technology stacks. Most content systems (74%) are not connected to other lines of business systems. Only 26% of document, content, and records management systems integrate with other core applications (AIIM, 2023).
- Rapid Obsolescence:
    - The average half-life of skills is now less than five years, and in some tech fields it's as low as two and a half years (Tamayo, et al., 2023).
- Domain Specificity:
    - In a systematic review of ontology matching techniques, Otero-Cerdeira et al. found that domain-specific knowledge remains a significant challenge, with over 60% of surveyed approaches requiring

manual input or domain expertise for effective ontology alignment (Otero, et al., 2015).

- Accessibility:
    - In a survey of 1,000 knowledge workers across industries, 72% reported difficulty finding and accessing the information they need to do their jobs effectively, despite the presence of knowledge management systems in their organizations (Cross, et al., 2004)
- Scalability:
    - A study of 179 multinational corporations found that as organizations grew in size and complexity, the effectiveness of their knowledge management systems decreased. Specifically, companies with over 10,000 employees reported a 23% lower satisfaction rate with their knowledge management systems compared to smaller firms (Heisig, et al., 2016).

Specific Problem Statement.

In addressing these general problems, I believe the system I will be developing here in addition to the product that I am working on with my startup, that we can tackle these:

- **Rapid skill obsolescence**, with a half-life of 2.5 to 5 years, creates a continuous need for efficient knowledge updating and dissemination within organizations
- **Domain-specific knowledge and vocabulary** create barriers to developing universally applicable ontology mapping solutions
- **Manual curation** of knowledge consumes 19% of workers' time, indicating a need for more automated knowledge extraction and organization methods

Closing Commentary.

The future of work is uncertain, and the problems that we identified here are central to this uncertainty. If the problems that we described here cannot be solved. We are likely to see some of the following effects:

- organizations will struggle to remain competitive outdated skills and knowledge
- inefficiencies and knowledge management will lead to increase costs and reduced productivity
- manual curation will continue to consume significant work hours, especially amongst the most valuable employees

What is so exciting about the work that we are doing is that the applications of this work are quite broad. While the exigencies of starting a company requires us to focus on the future of work, if we are successful, the applications are significantly broader. Perhaps we could even apply some of what we are doing to K-12 education. All told, it is exciting to be part of the future of knowledge management.

References

Association for Intelligent Information Management (AIIM). (2023, April 20). *2023 State of the Intelligent Information Management Industry*. AIIM. https://www.aiim.org/industrywatch2023

Chui, M., Manyika, J., Bughin, J., Dobbs, R., Roxburgh, C., Sarrazin, H., Sands, G., & Westergren, M. (2012, July 1). *The social economy: Unlocking value and productivity through social technologies*. McKinsey Global Institute. https://www.mckinsey.com

Cross, R., & Sproull, L. (2004). More than an answer: Information relationships for actionable knowledge. Organization science, 15(4), 446-462.

Deloitte. (2020). *2020 Deloitte Global Human Capital Trends: The social enterprise at work: Paradox as a path forward*. Deloitte Development LLC. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/human-capital/us-2020-deloitte-global-human-capital-trends.pdf

Gartner, Inc. (2023, May 10). *Gartner survey reveals 47% of digital workers struggle to find the information needed to effectively perform their jobs*. https://www.gartner.com/en/newsroom/press-releases/2023-05-10-gartner-survey-reveals-47-percent-of-digital-workers-struggle-to-find-the-information-needed-to-effectively-perform-their-jobs

Heisig, P., Suraj, O. A., Kianto, A., Kemboi, C., Perez Arrau, G., & Fathi Easa, N. (2016). Knowledge management and business performance: global experts' views on future research needs. Journal of Knowledge Management, 20(6), 1169-1198.

Otero-Cerdeira, L., Rodríguez-Martínez, F. J., & Gómez-Rodríguez, A. (2015). Ontology matching: A literature review. Expert Systems with Applications, 42(2), 949-971.

Tamayo, J., Doumi, L., Goel, S., Kovács-Ondrejkovic, O., & Sadun, R. (2023). Reskilling in the age of AI: Five new paradigms for leaders—and employees. *Harvard Business Review*, 101(5), 86–95. https://hbr.org/2023/09/reskilling-in-the-age-of-ai