Extracting Material Measurement Knowledge Graphs from Academic Research Papers

Stanford CS224N Custom Project

Arthur C. Campello
Department of Applied Physics
Stanford University
name@stanford.edu

Abstract

Knowledge graphs are abstract representations of connected concepts that form the basis of a machine's understanding of a subject matter. Building one with more information than is known by any one individual requires natural language processing of relevant text. This project focuses on building knowledge graphs about physical measurements that have been performed on materials of interest in condensed matter physics. Such graphs are always changing with new experiments, almost impossible to accurately construct with iterative search engine queries or generalized large language models, and critical in understanding the full picture of scientific progress concerning specific materials. Here I report the development and performance assessment of a co-occurrence proximity matrix (COPM) tool that takes in a collection of titles and abstracts of academic papers and outputs a knowledge graphs on which measurements were performed on which materials. I measure the performance of this tool over its primary parameter and compare it to that of ChatGPT employed with a structured prompt to address the knowledge task problem. While ChatGPT performs better in the data tested, the much faster nature of COPM makes it possibly more useful in certain research settings.

1 Key Information to include

- Mentor: Tony Wang

2 Introduction

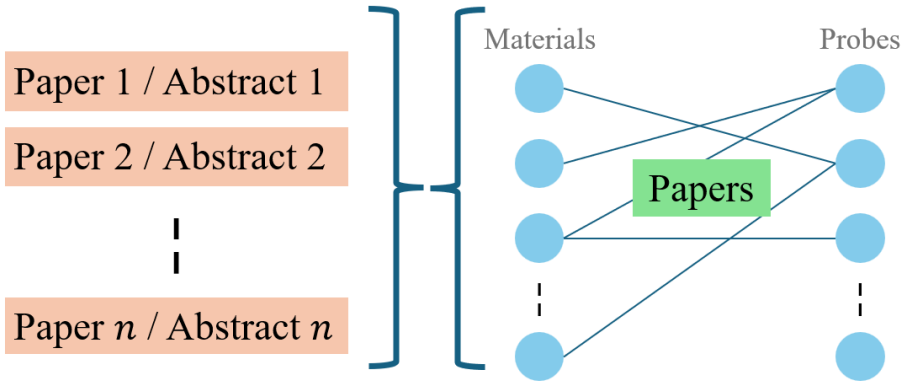
It is often said that to participate in an academic research field one must learn to "speak the language" of that field. Those familiar with such field-specific "languages" know that the nuances involved extend beyond sets of definitions. Terms and expressions for how already complex concepts interact and relate to each other comprise much of these field-languages, and can be hard to learn from online searches alone. Naturally, the more one engages in a field and communicates with mentors and peers, the more familiar these complex languages become to them. In most fields, the trade-off involved in the use of these languages is relatively harmless; communication becomes more natural at the price of small barriers to entry and lower interpretability to outsiders. A glaring exception to this harmlessness is the massive sub-field at the intersection of physics and material science known as condensed matter physics (CMP).

CMP deals with the macroscopic effects of microscopic material behaviors and the portfolio of inventions under its umbrella includes semiconductors, the transistor, superconductors, and lasers (Martin, 2017). In the United States, research in the field is largely funded by the Department of Energy, which also administers several national laboratories strongly dedicated to materials physics. Active research in CMP holds relevance to many aspects of daily life and, importantly, to the national security of nations. Despite the field's strong real-world connection to the, and to the dismay of

many, the language of condensed matter can be among the most disorienting in physics. Beyond extensive acronyms like, NMR, CDW, BCS, ARPES, DFT – normal of most field-languages – CMP appropriates ordinary words in counter-intuitive ways. The word "state" as in "states-of-matter" is used liberally and expressions for interactions between states, like "rivaling states" or one state "pinning" the other, are ill-defined but widely understood. Prefixes to states like "quasi-" and "proximate" are used to soften claims as are suffixes "-like" and "candidate." Also, the meaning of "disorder" in front of a state or material is highly context dependent and obvious only to those familiar with it. These quirks and many others make the language of condensed matter interesting from a linguistic perspective, but intensely frustrating when it comes to extracting real information.

Even as measurements on materials form the basis of experimental CMP, an informed reader will often struggle to examine a paper inthe field and derive a sense of which measurements were performed on which materials. With ambiguous phrases like "transport measurements" – which could mean heat transfer or conductivity – or passive phrases one "x measurement was performed on material y" – sometimes referring to prior work and sometimes the presented work – paper abstracts sometimes plainly obfuscate presentations of measurements. Paired with writing in the CMP language, these ambiguities also make challenging to search "was x measured on material y?" in Google or sites like Google Scholar and obtain accurate or reasonable results. With the amount the Deparement of Energy is requesting in 2024 for research in condensed matter physics and material science topping $220 million USD, it is critical that the fruits of this publicly funded research are as clear and accessible to the public as possible.

This project aims to address the issue of measurement transparency by exploring approaches to extracting knowledge graphs about material measurements given paired titles and abstracts of academic papers in condensed matter. A knowledge graph is an abstract representation of knowledge as collection of connected concepts, instances, and relationships. As detailed by Peng et al. (2023), knowledge of this structure can be derived form both structured and unstructured text data. This project concerns the latter type of extraction applied to academic writing, which happens to be an active area of research – see Yu et al. (2020); Liu et al. (2020). The knowledge graphs relevant to this report connect `material` objects to `probe` objects based on whether a given material has been measured with a given probe. Small graphs connecting one or a few of each can be extracted from a single title-abstract-pairs and graphs from multiple papers can be combined to yield a complete bipartite knowledge graph where each edge corresponds to a paper. The graph extraction task can be represented schematically as follows:



As a point of focus, this report examines experimental papers relevant to quantum spin liquid (QSL) materials, whose spins exhibit long range quantum entanglement (Broholm et al., 2020). Interest in QSLs has grown rapidly in the last decade and controversies rage around which materials are true QSLs and if any are at all. The titles and abstracts of 115 prominent papers from the past decade are collected and their "true" knowledge graphs are manually completed by a person with close familiarity with the field – the author. Then a co-occurence proximity matrix (COPM) model is applied that determines the presence of key words relevant to materials and measurement probes key words the abstract and title corpuses and produces a knowledge graph for each entry. The model works by having a word base of materials and measurement probes that it may reference. Upon finding instances of these in the corupuses, it estimates whether the two are linked in the graph – i.e. whether that material has been measured with that probe – based on the proximity of words.

This proximity is determined by a threshold variable n. As expected, this model outperforms a random baseline and its performance is highly dependent on n. To compare the performance of this model against a large language model (LLM) I employed ChatGPT with a carefully crafted prompt template. While ChatGPT did perform better than the model, it does so at a much slower rate and its formatting reliability is much less. In practice, parsing a large collection of papers to create a deep knowledge graph requires the speed and reliability shown in the COPM model. Hence, neither approach is strictly optimal and the "best" solution likely requires the joint use of both to construct large knowledge graphs.

3 Related Work

The original concept of a knowledge graph was conceived by Google when it released the Google Knowledge Graph in 2012 without details on its implementation. Since this, the topic of knowledge graphs has become a popular research area in Computer Science and Information systems, as detailed by Peng et al. (2023) in its review of challenges and opportunities in the field.

Much of the recent work in this area concerns how to use abstract representations of information in practical and interpretable ways. A relevant task to this is embedding high-dimensional graphs into lower-dimensional ones that are more palatable – see Dai et al. (2020). Such embedding holds relevance to parsing concepts in academic and scientific literature as they exist in high dimensions and may be connected in myriad ways. Continued progress in embedding can pave a path to a future where a massive shared knowledge graph exists, and sub-parts of it in low dimensions may be accessed upon request.

The more directly relevant work on knowledge graphs related to this project is that concerning their extraction from data. This effort is severely complicated by the vastness of the types of media containing information pertinent to knowledge graphs and varying degrees of quality and structure to these data. Yu et al. (2020) approaches this challenge for structured, semi-structured, and unstructured data using various approaches, including analysis of a co-occurrence matrix and a residual neural network. One facet of this work involved using Wikipedia entries and construct a knowledge graph about food. The relative success of their co-occurrence matrix analysis approach inspired the approach attempted and examined in this project.

An important note about the existing literature on knowledge graphs is that much of it was published before the meteoric popularity of large language models with access to public data. Hence, while they report simple-to-complex models and their performances on limited data, literature on knowledge graphs is missing comprehensive comparisons with ChatGPT and other public LLMs used to construct them.

4 Approach

The first step I completed in this project was to find 115 academic papers relating to measurements on materials of interest to QSL research. On Google Scholar, I searched "Quantum Spin Liquid" and filtered papers written after 2015 – about the year when terms became a bit more standardized. I then picked the first 115 experimental papers I saw. I isolated titles and abstracts of each paper and saved them into files such that each could be referenced whenever necessary.

Following this, I manually made a list of materials and measurement types reported in each of the titles and abstracts of the paper, obtaining complete a knowledge graph for each entry. Having done this carefully, I use these "true" answers to train and/or validate models. All models tested produce an ℓ-length list of pairs of the form (`material`, `measurement probe`) that construct the knowledge graph. An important component of the loss function is to encourage the model to choose words that actually correspond to materials and measurements for these entries.

To accomplish this, I construct sets of known materials and measurement types called S_{M} and S_{P}, respectively, and respective generalized subsets M and P. Now my loss function can incorporate indicator functions $\mathbb{1}_M(\cdot)$ and $\mathbb{1}_P(\cdot)$ that return 0 if a word is not in the respective set and 1 if it is. A consideration for the loss metric is that often papers will reference multiple materials and measurements in contexts not relevant to measurements reported by the authors even in abstracts. This makes it easy for models to "over-construct" knowledge graphs. To reward the model for not

only being accurate, but not too sparse, it is helpful to incorporate cross entropy loss functions. This includes those just for lists of materials and measurement types as well as those for [`material`, `measurement probe`] pairs. To define the loss I let predicted and true knowledge sets $\hat{K}$ and K be sets of pairs (m_i, p_i) – being materials and probes. Sets $\hat{M}$, M, $\hat{P}$, P give predicted and true materials and probes, respectively. I now define

$$\mathcal{L} = -\lambda \sum_{m \in M \cup \hat{M}} \frac{\mathbb{1}_M(m)}{|M|} \log\left[\frac{\mathbb{1}_{\hat{M}}(m)}{|\hat{M}|}\right] - (1-\lambda) \sum_{p \in P \cup \hat{P}} \frac{\mathbb{1}_P(p)}{|P|} \log\left[\frac{\mathbb{1}_{\hat{P}}(p)}{|\hat{P}|}\right]$$

$$-\gamma \sum_{(m,p) \in K \cup \hat{K}} \frac{\mathbb{1}_K\left[(m,p)\right]}{|K|} \log\left[\frac{\mathbb{1}_{\hat{K}}\left[(m,p)\right]}{|\hat{K}|}\right],$$

with parameters $\lambda \in [0, 1]$, and $\gamma > 0$. The form of this loss was inspired by – but is different from – the loss metric used in Yu et al. (2020).

With this loss established, I performed evaluations on two approaches of interest to this project. The first approach was to write a script that would generate a carefully-worded prompt for OpenAI's ChatGPT (version 3.5) interface to attempt constructing the knowledge graph for a specific paper's title and abstract pair. The prompt generator used was:

> "An academic paper in condensed matter physics is titled '[`title`]' and has the following abstract: '[`abstract`]' Identify which measurement probes were performed on which materials and produce an answer in the form of a comma-separated list of [[material], [measurement probe]] for every combination. Give the overall the list in one line in square brackets and if you think no measurements were performed, just return []"

The second approach was a simple language parsing model that we call the "co-occurrence proximity model" or (COPM) This works as follows: Given the full sets M and P of possible materials and measurement probes respectively, subsets $M_a \subset M$ and $P_a \subset P$ are first obtained based on which materials and measurement probes appear in the abstract. Similar M_t and P_t are obtained for those mentioned in the title. All possible pairs (`material`, `measurement probe`) are determined for sets $M_a \cup M_t$ and $P_a \cup P_t$ and pairs are only kept if the number of words between tokens is less than integer $n \in \mathbb{N}$ or if one of the pair items appears in the title. More formally,

$$S_{\text{COPM}} = \left\{(m,p) \in (M_a \cup M_t) \times (P_a \cup P_t) : d^*(m,p) < n\right\},$$

$$d^*(m,p) = \begin{cases} d(m,p) & m \in M_a \wedge p \in P_a \\ 0 & m \in M_t \vee p \in P_t \end{cases}$$

gives the set of pairs, where $d(m,p)$ is the number of words between material m and measurement probe p. The purpose of this n threshold parameter is to keep the model from over-constructing thw knowledge graph by connecting all materials to all measurement probes it detects in the text. The idea is that lowering n will regularize the model and keep its inferences focused to the word around it. Training this model entailed finding the optimal n to minimize the model's loss.

These were tested against a baseline resulting from randomly selecting material and measurement probe pairs from the pool of "true" pairs.

5 Experiments

This section details the collection of the data used and the experiments performed to obtain the key results presented in this project.

5.1 Data

The data for this project was all sourced from Google Scholar by searching for QSL papers from the last decade. After manually examining over 250 papers, I selected 117 that were experimental and relevant to condensed matter physics. Upon further review of the selected papers I noticed that two

were actually theory papers and excluded those from the set. For each paper, transcription of the title and abstract was performed using copy-paste with special attention to manually remove special characters or symbols that would be hard to parse with a python script. For some publishers – like those in the Physical Review family – the data could be scraped from the citation export. Information about the year of publication and the web address of the paper was also recorded. A summary of the distribution of years, title lengths, and abstract lengths is shown in the histograms below.

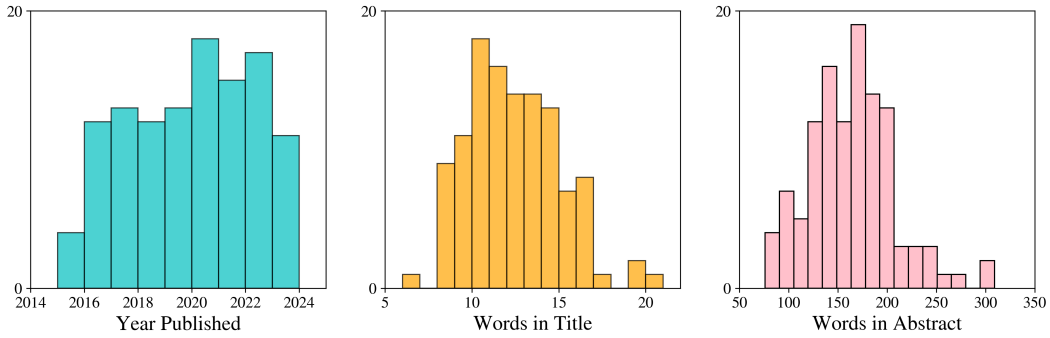


Figure 1: Histograms detailing the distributions of years published, numbers of words in titles, and numbers of words in abstracts of the QSL-focused papers used.

A link list of all the papers used is included in Appendix A.

Data on the responses to my ChatGPT prompt was also obtained my manually copying the output and pasting into python notebooks.

5.2 Evaluation method

All quantitative evaluations reported in this project are performed using the loss metric defined above, which is comprised of a sum of cross-entropy loss components. The first two cross entropy loss components reward the model for correctly finding the list of materials and measurement probes and are weighted by λ and $(1-\lambda)$, respectively. Often, papers often report multiple measurements on one type of material. It is hence usually more "important" to correctly identify the focus material than the exact measurement probes. To account for this asymmetry, $\lambda = 0.75$ is chosen. I also chose $\gamma = 3$ to reward the model for correctly identifying pairs and not just picking correct materials and probes that it "noticed in the text."

In some cases, I found instances where tested models underperformed evaluated the performance of the model as a "human evaluator" to understand what happened. Some noteworthy instances are highlighted in the "Analysis" section.

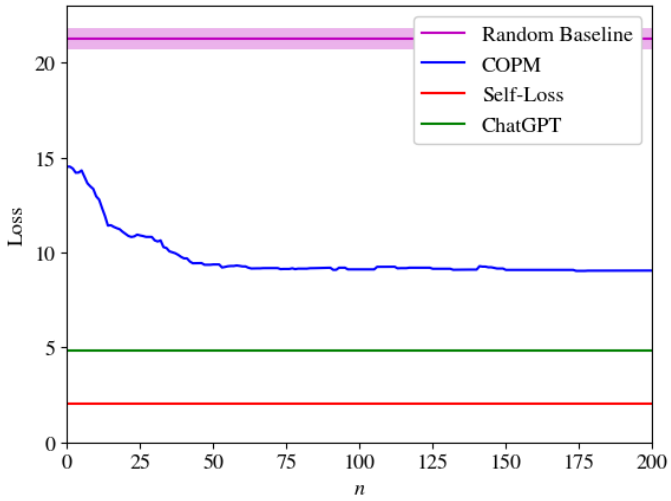
5.3 Experimental details

With the loss metric established, experiments were run to compare the performances of the approaches mentioned. An upper-bound baseline loss was set by the "random loss" which was that from testing random true entries against each other. This baseline has some uncertainty due to the randomness, but a working method should perform better. The lower bound loss was the "self-loss:" that from the loss function between data and itself.

For the ChatGPT approach, the relevant experiment was calculating the loss across the collection of papers and compare it to the baselines. For the COPM approach, I experimented with the effect of n, the threshold parameter on the loss. This elucidated the effect of proximity between material measurement and probing terms and their probability of connection in the knowledge graph.

5.4 Results

The results of the experiments performed are summarized in the following plot, which shows the defined loss for the two baselines, the ChatGPT approach, and the COPM model with n-dependence up to n=250 words.



The randomness of the baseline derived from randomly sampling true knowledge graphs paired is reported to have an average $\boxed{L_{\text{rand}} = 21.3 \pm 0.5}$. The stanard deviation based on the matching randomness is shown in the magenta bar in the figure. The lowest possible loss – that from matching true graphs with themselves – is $\boxed{L_{\text{true}} = 2.0}$. Using ChatGPT with the structured prompt above yields an loss of $\boxed{L_{\text{ChatGPT}} = 4.8}$, which is strong considering the cross-entropy nature of the loss used. As expected, the COPM loss has a substantial n-dependence, but it was surprising to see that the loss appears to decrease with n and does not increase again at a limit where n spans the corpus of the text. The lowest measured loss was $\boxed{L_{\text{COPM}} = 9.0}$ at n=174.

6 Analysis

With insights into specific instances of this knowledge extraction task, it is not surprising that ChatGPT outperformed the COPM model overall in performance. It's LLM nature allows it both to understand nuances of language relevant to resolving ambiguities involved in this task. Given the instructions in the format of the prompt above, ChatGPT was even able to extract pieces of information that an informed observer would expect could mislead a language model. For example, one abstract uses the phrasing "A2IrO3 (A=Na, Li)" to mean "Na2IrO3 and Li2IrO3" – i.e. that A can be substituted for either element. ChatGPT was able to identify that this meant two materials were studied. One sense in which ChatGPT performed poorly was a tendency to "overconstruct" the knowledge graph by interpreting claims about uncovered insight into materials as actual measurements performed. For example ChatGPT may interpret a sentence like "Here we measure spin-singlet excitations in Herbertsmithite using inelastic neutron scattering." to signify that the "spin-singlet excitations" are a measurement and not an interpretation of an inelastic neutron scattering measurement. A failure of ChatGPT not captured by the loss was its occasional deviation from the syntactic structure the prompt prescribed. Despite being prompted to produce answers in one line and with a convenient bracketing scheme, ChatGPT would occasionally deviate from this and persist with poor answers even when prompted to regenerate a solution. Though seemingly trivial, this lack of syntactic certainty may be an important consideration for the use of LLMs in constructing knowledge graphs.

Compared to the ChatGPT approach, the COPM model had the advantage of a reference base of materials and measurement probes to compare against the texts. This prevents the model from categorizing strongly worded paper author interpretations as measurements performed. The weakness of this model is its rigidity an inability to detect nuance in prose and adapt to the different ways of expressing the same measurement technique. Qualitatively, the performance of COPM was comparable to that of an informed person skimming the abstract and finding a verbal "table" of measurements performed. In cases where the author explained these clearly – and for a reasonable n – COPM performed very well. Looking at examples, one sees that COPM performs worse than ChatGPT in cases where claims about the significance of findings overshadows or "pushes aside"

clear descriptions of meeaurements, which is the case in many papers in this competitive field. A clear advantage of COPM over ChatGPT is its speed and syntactic reliability. With this strength, a natural way to modify COPM for improved performance is to have an auxiliary identifier of instances where COPM works as indended versus instances where a larger model may be needed to complete the task.

7 Conclusion

This report presents and compares two strong language processing approaches to the task of constructing knowledge graphs from unstructured data in the form of paper titles and abstracts in the field of condensed matter physics. It marks a small step toward an exciting possible future for science where research and findings are fed into a single, large knowledge graph and parsed on demand to anyone curious for structured information. As it stands now, accessing critical information relevant to material physics from online searching and paper sleuthing alone is critically slow due to the opaque language of condensed matter physics and commonly obfuscated information about performed material measurements. Though opposites in their computational complexities, both the approaches of structured ChatGPT prompts and analysis of co-occurrence matrices perform modestly well in the task of knowledge graph construction. The approaches seem to perform well in complementary cases and paired with some intermediate-fidelity models and a model selection scheme, could prove incredibly powerful toward the difficult task of knowledge graph construction.

Near-term work in this vein can take myriad directions, but could focus on the creation of a small-language model optimized to tackle this task that can do so quickly and cover vast areas of research. Another interesting avenue to explore is the possibility of a knowledge graph that helps researchers decide their next steps and focus areas. For example, a completed knowledge graph for condensed matter physics would enable a researcher with some measurement probes at their disposal to search determine the best measurements to perform given the tools at their disposal to maximize their impact on the field – and/or their odds of being published.

References

C. Broholm, R. J. Cava, S. A. Kivelson, D. G. Nocera, M. R. Norman, and T. Senthil. 2020. Quantum spin liquids. *Science*, 367(6475):eaay0668.

Yuanfei Dai, Shiping Wang, Neal N. Xiong, and Wenzhong Guo. 2020. A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics*, 9(5).

Jiaying Liu, Jing Ren, Wenqing Zheng, Lianhua Chi, Ivan Lee, and Feng Xia. 2020. Web of scholars: A scholar knowledge graph. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2153–2156, New York, NY, USA. Association for Computing Machinery.

Joseph D. Martin. 2017. Resource Letter HCMP-1: History of Condensed Matter Physics. *American Journal of Physics*, 85(2):87–97.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 56(11):13071–13102.

Haoze Yu, Haisheng Li, Dianhui Mao, and Qiang Cai. 2020. A relationship extraction method for domain knowledge graph construction. *World Wide Web*, 23(2):735–753.

A Appendix

For reference I include the full list of papers from which abstracts and titles were extracted for use in this project.

Paper 001Paper 002Paper 003Paper 004Paper 005Paper 006Paper 007Paper 008Paper 009Paper 010Paper 011Paper 012Paper 013Paper 014Paper 015Paper 016Paper 017Paper 018Paper 019Paper 020Paper 021Paper 022Paper 023Paper 024Paper 025Paper 026Paper 027Paper 028Paper 029Paper 030Paper 031Paper 032Paper 033Paper 034Paper 035Paper 036Paper 037Paper 038Paper 039Paper

040Paper 041Paper 042Paper 043Paper 044Paper 045Paper 046Paper 047Paper 048Paper 049Paper 050Paper 051Paper 052Paper 053Paper 054Paper 055Paper 056Paper 057Paper 058Paper 059Paper 060Paper 061Paper 062Paper 063Paper 064Paper 065Paper 066Paper 067Paper 068Paper 069Paper 070Paper 071Paper 072Paper 073Paper 074Paper 075Paper 076Paper 077Paper 078Paper 079Paper 080Paper 081Paper 082Paper 083Paper 084Paper 085Paper 086Paper 087Paper 088Paper 089Paper 090Paper 091Paper 092Paper 093Paper 094Paper 095Paper 096Paper 097Paper 098Paper 099Paper 100Paper 101Paper 102Paper 103Paper 104Paper 105Paper 106Paper 107Paper 108Paper 109Paper 110Paper 111Paper 112Paper 113Paper 114Paper 115