Fine-Grained Chemical Entity Typing with Multimodal Knowledge Representation

Anonymous EMNLP submission

Abstract

How to extract knowledge about chemical reactions from the core chemistry literature is a new emerging challenge that has not been well studied. In this paper, we introduce a new benchmark data set (CHEMET) to facilitate the study of knowledge extraction in this new domain. Fine-grained chemical entity typing poses interesting new challenges especially because of the complex name mentions frequently occurring in chemistry literature and graphic representation of entities. At the same time, there are also interesting new opportunities to leverage external chemistry knowledge resources. We propose a novel multi-modal representation learning framework to solve the problem of fine-grained chemical entity typing by leveraging external resources with chemical structures and using cross-modal attention to learn effective representation of text in the chemistry domain. Experiment results show that the proposed framework outperforms multiple state of the art.

Chenkai **[colored sentences were added/modified/highlighted according to comments]**

1 Introduction

As the amount of research literature is growing exponentially, accurate and efficient information extraction (IE) methods are crucial for many downstream applications including question answering and knowledge reasoning. One domain largely overlooked by previous IE research is Chemistry (an example sentence is shown in Figure 2), which consists of discussion on chemicals and reactions they are involved. What benefit can it bring if we develop well-performing IE methods for chemistry domain? If a comprehensive chemistry knowledge base can be efficiently constructed, chemicals can be discovered at a faster pace since models can learn from existing reactions to infer never-imagined ones, thus benefiting downstream applications such as those in biomedical research and chemical industry.

One fundamental building-block of information extraction is fine-grained entity typing (FET), which is the task of classifying entity mentions into subset of pre-defined hierarchical classes (e.g., Person/Artist, Location/City in news domain), and doing well in such task typically requires the system to understand the mention and its context well. The task is particularly challenging for scientific articles, where domain-specific knowledge is heavily required to understand the text; for instance, in chemistry, one needs to understand the reaction mechanism in the literature described by both equation image and text (about experiment conditions), and since a reaction is based upon chemical compounds, it additionally assumes one to have knowledge about the chemical as well. Intuitively, to understand scientific articles, linking entities appearing in the text to retrieve and comprehend external information in different modalities would be very helpful. Analogically, when a person learns a cooking recipe, he or she would look up cooking instruction video (consisting of image, text, and audio) to help understanding the procedure In scientific domains, information extraction models have been widely developed for biomedical context (Liu et al., 2016; Poon and Vanderwende, 2010; Li et al., 2019a, 2017; Cho and Lee, 2019; Beltagy et al., 2019; Lee et al., 2020; Liu et al., 2018; Tian et al., 2020) However, while chemistry research shapes the foundation of many biomedical studies, there has been little work done in extracting knowledge from core chemistry research literature; previous work in Chemistry IE mainly focuses on Named Entity Recognition (NER) (e.g., recognizing chemical name spans), and there is only one work we were able to find (Nguyen et al., 2020) on task other than NER (e.g., reaction event extraction). One major difference between chemistry and

biomedical literature text lies in different chemical entity expressions, where chemical compounds in biomedical text are often expressed in natural language (e.g., water, aspirin), while in chemistry it's often complex formula-like names (e.g., 5,6-dihydroxycyclohexa-1,3-diene-1-carboxylic acid, H2O), which is hard to be understood by existing language models as such complex names do not follow morphological structure like other commonly used words like "basketball". To make the situation worse, many chemicals simply have never been coined with any nomenclature in natural language. The chemical mentions are essentially rare terms that is not best to be and thus would be not learned well by language model

Although there has been a line of method in FET applied to news domain (Choi et al., 2018; Xiong et al., 2019; Dai et al., 2019; Lin and Ji, 2019; Jin et al., 2019; López et al., 2019), none have been developed for core chemistry literature and they do not consider any types of domain-specific knowledge. While language model may have a hard time understand the chemical mention purely based on its surface form and contextual representation, we can understand the identity through it's external information (in different modalities) such as natural language description about its properties and it's structure (or graph). In the chemical typing task particularly, compound types can be well correlated with properties and physical structure

Our work is novel in that we are the very pioneers to explore FET strategies in chemistry. Utilizing external database for multimodal information retrieval, we introduce a deep learning based method that use cross-modal attention to align and embed the structure and description text of chemicals into a common space as core features for classification. As illustrated in Figure 1, some patterns of molecular substructures well align with the phrases in description text. For example the circled substructure in the is commonly appeared together with "polar aprotic".

Since the proposed task has not been studied in the previous work, there is no dataset available for evaluating the task. To facilitate the study of this new task, we construct CHEMET[1], the first dataset for fine-grained typing in the chemistry literature domain, for which we referred to wikipedia category for ontology construction, and used distant supervision to generate training data and facilitate

[1]Both the dataset and the code will be released to public

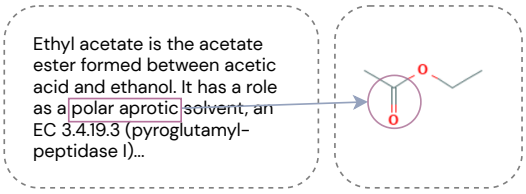

Figure 1: An example of chemical entity structure aligning with textual concepts. The circled substructure is often induce the "polar aprotic" property infer that ethyl acetate is polar aprotic

annotation procedure. The dataset was based upon a corpus of 50 open access papers from a database on a specific theme. We will discuss the data construction details in Section 2. Experimental results on the the dataset show that our method outperforms the state-of-the-art methods in entity typing. To the best of our knowledge, our method is the first to take step toward tackling fine-grained chemical entity typing.

Overall, our contributions can be summarized as the following:

- We study the task of fine-grained chemical entity typing in chemistry literature, a largely under-explored yet promising field for NLP that has a great need for information extraction methods
- We construct the first human-annotated dataset in fine-grained chemical entity typing and will release to the public.
- We introduce an novel method that utilizing multimodal knowledge representation to enrich entity mention representation
- The multimodal component of the model is based on structure and text alignment, which has never been explored before and can be applied to variety of ChemIE tasks such as relation extraction and event extraction.
- Experiments on the dataset show that our model outperforms the state-of-the-arts entity typing models.

2 Dataset

agreement metric

describe diversity

Cheng **["limited dataset" is a bit vague. Is there any such data set available? If so, we should try to use**

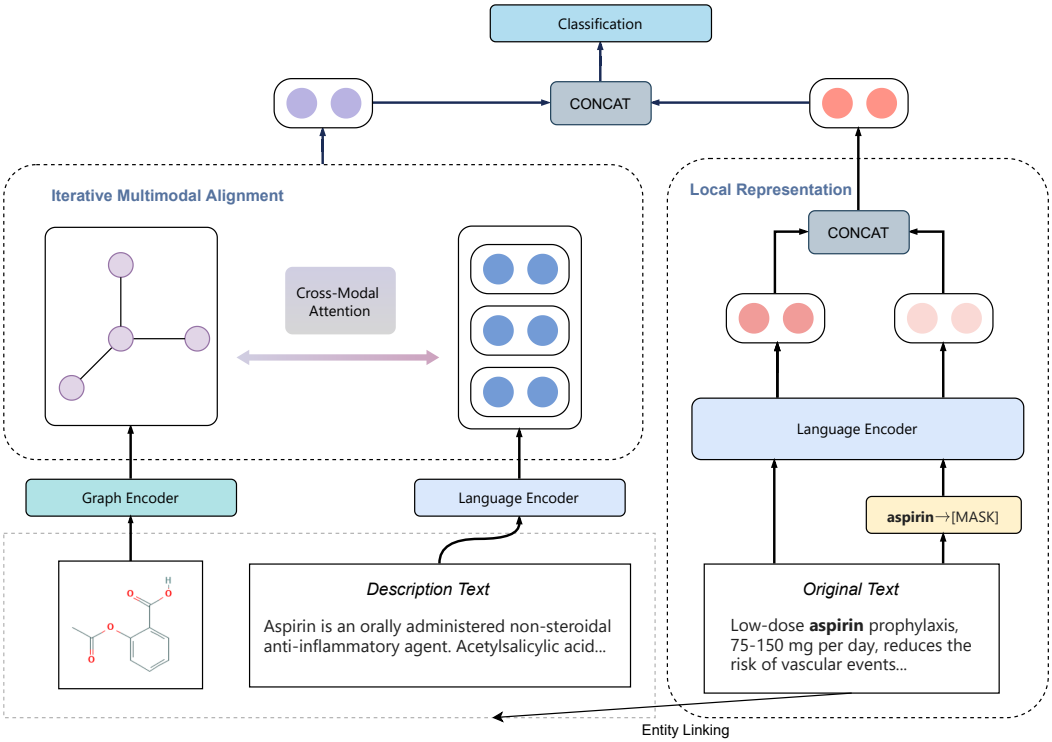

Figure 2: Our fine-grained chemical entity typing model architecture. Please refer to Section 3 for details.

Table 1: Dataset Statistics for CHEMET

Setting	Anno.	#Inst.	#Entity	#Types
Train	Distant	1000	NA	39
Dev	Human	500	NA	39
Test	Human	500	NA	39

it. My sense is that there isn't(?), so the novelty and significance of the data set could be more clearly articulated.] Due to limited dataset being available for fined-grained chemical entity typing, we have collected and annotated a dataset, CHEMET, based on a corpus of 50 papers from PubChem[2]) with Suzuki-Coupling (a popular reaction mechanism) theme; the theme was chosen to align with chemistry specialists' domain knowledge. We will discuss the steps taken to construct the dataset below.

Taxonomy Construction. *Cheng* **[how was the number 39 determined? try to give a justification or explanation of the process that reached the number 39.]** We carefully select 39 subcategories from wikipedia chemistry category page [3] as fine-grained ontology; for example, Organic chemistry→Organic compounds→Esters is a fined-grained type where right of the arrow is the sub-category of the left. *Cheng* **[The following sentence can be moved earlier in this paragraph to explain the strategy being taken. It's better to first give a description of our goal/strategy/philosophy and then describe how we do it. If there are decisions to be made, explain why we've decided to choose one options not another.]** We focused on types that are compound types commonly occurring in Suzuki-Coupling literature. The entire ontology is shown in Appendix A.

Distant Supervision In order to ease human annotators' work and to collect training data, we employed distant supervision to retrieve noisy labels for the corpus. In this step we first tokenized text using (Jessop et al., 2011), a texting mining framework for chemistry that recognize complex chemical name well. We then collected a dictionary mapping from picked types (that is, the select categories from Wikipedia) to their belonging wikipedia pages. We treated the page titles as entity names. Since a compound can have many synonyms, we queried PubChem to expand the dictionary. Finally, we used the dictionary to label the tokenized text using a well-performing string matching algorithm. *Cheng* **[Provide a reference to this string matching algorithm or elaborate.]**

Human Annotation. We hired five undergraduate

[2] https://pubchem.ncbi.nlm.nih.gov/
[3] https://en.wikipedia.org/wiki/Category:Chemistry

chemistry students as annotators. The annotators were instructed to identify and type spans in the assigned samples using Brat (Stenetorp et al., 2012) interface. To ensure the diversity of the testing data, we randomly select the test samples from the corpus for annotation. In order to mitigate annotator bias, we distributed each sentence to three annotators, and take majority vote from the results. The dataset statistics is shown in Table 1.

3 Method

Cheng **[Before describing the method, it's better to briefly motivate the method. Perhaps remind the readers what are the key (new) challenges we need to address, and then briefly describe the proposed ideas for addressing those challenges informally. This would help readers see the big picture (in terms of novelty and justification) before going over the detail of the model.]**

Chemistry literature is unique in that the mentions are often expressed in complex, unnatural forms, as shown in Figure **??**. At the same time, external databases provide multimedia data about chemical entity, such as chemical structure and natural language description. It is thus a natural idea to incorporate these different forms of representation of an entity to enhance the system's understanding on a chemicals. In our methodology, we develop an effective deep learning based model to implement such idea.

The overall model architecture is presented in Figure 2. Given a sentence or document S marked with mentions, we first extract external information (molecular structure and description text) by linking to PubChem, one of the mostly used chemical database; we use its search API to fetch molecule information given a molecule mention name. We also used a modified version of S that masked the entire mention name, to combat the issue of complex compound name (3.1). <ModelName> proceeds to extract features from different modalities by their corresponding encoders. The external information embeddings are then passed through multimodal alignment stage (3.2) to learn a unified representation,

3.1 Original Text Embedding

The model first encodes the original sentence with SciBERT (Beltagy et al., 2019), a Transformer (Vaswani et al., 2017) based language model pre-trained on biomedical text. Let $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_z]$, where z is the number of tokens in the sentence, after tokenization. Then we follow (Lee et al., 2017) and compute the representation for mention m by

$$\mathbf{m} = \text{FFNN}_t([\mathbf{t}_{\text{START}(m)}, \mathbf{t}_{\text{END}(m)}, \hat{t}, \phi(m)])$$

, where FFNN_t is a feed forward neural network. $\text{END}(m)$ and $\text{START}(m)$ denote start and end indices for m. $\hat{t}$ is the representation based on attention to each token in m.

3.1.1 Context-focused Embedding

Since Chemical entity are often involved with complex synonyms that are hard to be understood, we need to also produce a representation that rely less on the word structure of the mention, since the mention often not follows morphology (e.g., [3H]MK-801, NSC-406186, 8-azido-[alpha-32P]ATP). We replace the entire span of mention by [MASK]. The modified sentence is then embedded by SciBERT and the embedding for the [MASK] token is used as the corresponding context-focused representation for the mention, denoted $\mathbf{m}_{\text{MASK}}$

3.2 Multimodal Encoder for Structure and Description Text

Chenkai **[please ignore, whole thing will be changed]**

As one of our core contributions, we propose to incorporate different modalities of external features to expand chemical representation, and to combat the difficulty of understanding complex chemical mention name (e.g., (E)-3-(3,4-dihydroxyphenyl)prop-2-enoic acid) purely based on context and morphological structure.

Specifically We use API provided PubChem as the entity linker to retrieval chemical structure and description text for each chemical mention. Chemical structure refers a graph where bonds are edges and atoms are nodes, and dscription text discusses subset of chemical's experimental properties(e.g., Aspirin is an orally administered non-steroidal antiinflammatory agent).

To learn concepts from multiple modalities that better correlate with target label and to build more accurate representation of a molecule, we made use of the recently successful attention mechanism to align concepts (or molecule property) in text and substructure in molecule graph. One another benefit is that system can implicitly learn to better cluster molecules even if a chemical entity is missing some modalities (e.g., only have structure

available), since we map both modalities into the same embedding space.

Formally, let $G = (V, E)$ denote the chemical graph with a nodes, and $D = [d_{[CLS]}, d_1, d_2, ..., d_b]$ denote the sequence of$b$ tokens after tokenizing description sentences. Similar to original sentence, we embed with SciBERT so that text embedding becomes $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_b]$. We then embed the nodes in chemical graph using Graph Isomorphism Network(Xu et al., 2018), which atom features randomly initialized. We denote node representation $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2, ..., \mathbf{n}_a]$

We leverage self-attention mechanism of (Vaswani et al., 2017) to learn interaction between different modalities. To achieve this, we first stack the node and token embeddings as

$$\mathbf{X} = \begin{pmatrix} \mathbf{N} \\ \mathbf{D} \end{pmatrix}$$

Then we compute key values of the matrix by $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$,$\mathbf{K} = \mathbf{X}\mathbf{W}^K$, and $\mathbf{V} = \mathbf{X}\mathbf{W}^V$

Then the attented representation is given by

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{p_k}})\mathbf{V}$$

where $\frac{1}{\sqrt{p_k}}$ is a scaling factor in (Vaswani et al., 2017). We used an average pooling to get multimodal representation $\mathbf{E}_{CM}$

In addition, we preserve the unimodal graph representation by max pooling over the node representation, to get $\mathbf{E}_G$. We also use the CLS embeddin $d_{[CLS]}$ to represent unimodal text features.

We can predict the final entity type prowith the enriched multimodal features by

$$\mathbf{p} = \text{Softmax}([\mathbf{E}_C, \mathbf{E}_M, \mathbf{E}_{CM}, \mathbf{E}_D, \mathbf{E}_G])\mathbf{W}^F$$

where $\mathbf{W}^F$ is a learnable weight matrix and $\mathbf{p}$ is the final probability distribution of classes.

3.3 Training

We use cross entropy for training

4 Experiments

Since there is no other fine-grained chemical entity typing datasets to our knowledge, we evaluate fine-grained chemical entity typing on the CHEMET. The experiments can be reproduced using implementations provided in supplement material.

4.1 Baseline Methods

Cheng **[It would help to clarify what questions we can answer by comparing the proposed methods with these baselines. It seems to me that these baseline methods do NOT use the same amount of information/resources as the proposed method. If so, the improvement may have come from the fact that we have used additional information/resources, which wasn't used in the baseline methods. This wouldn't be a surprising finding as we are expected to do better with more resources/information. The more interesting question here is: what's the best way of exploiting such information? So if possible, it would be great to include stronger baseline methods that use the SAME amount of extra information. This would help showing the proposed method is better, not just because it has access to more information/resources, but also because the method can better utilize the extra information than a baseline way of using it (e.g., straightforward combination of existing methods to achieve the goal). Ideally, the baseline methods can be aligned with the most relevant previous work discussed in the related work section. This would help us empirically examine/support the novelty of the work in comparison with previous work from multiple perspectives (e.g., the perspective of tackling the complex name mentions, the perspective of multi-modal attention/embedding, and the connection between local context with molecule structure(?).]** In the experiment, we compared our method with the following state-of-the-art text classification and fine grained entity typing models,

SciBERT. SciBert (Beltagy et al., 2019) is a Transformer based language model pretrained on sample of 1.14M papers from Semantic Scholar, in which 82% are from the broad biomedical domain. A linear layer is applied on the embedding of [CLS] for classification.

BioBERT. Similar to SciBert but pretrained on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC). Similar to SciBRET, a final linear layer is applied for classification.

Latent Type Representation. Lin and Ji (2019) used a hybrid classification method beyond binary relevance to exploit type inter-dependency with latent type representation

Fine-Grained Entity Typing in Hyperbolic Space Utilized hyperbolic embeddings. If have time.

Table 2: Transductive Imputation AUC with 10% missing data

MODEL	ACCURACY	MACRO-F1	MACRO-F1
1	1	1	1

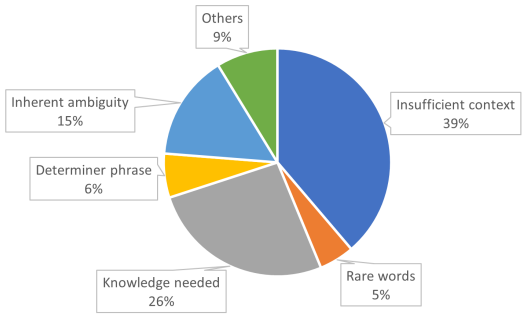

Figure 3: Distribution of remaining errors on the test set.

4.2 Implementation Detail

follow Hyspa

<ModelName> was implemented using PyTorch (Paszke et al., 2019) and Huggingface Transformers (Wolf et al., 2019) with SciBERT as text encoder. We left model and training parameters and reproducibility details in the appendix for interested readers.

4.3 Result

4.4 Ablation Study

To show the improvement made by each of the submodules in our method, we truncate model in the following way

- w/o multimodal alignment: FFNN
- w/o molecule graph:
- w/o description text:

4.5 Attention Analysis for Structure-Text Alignment

By the heatmap visualization

4.6 Error Analysis

Here we analyzed the remaining errors and categorize the into different cases (shown in Figure 3). We discuss the most common ones below

5 Related Work

5.1 Fine-Grained Entity Typing

There has been a wave of Fine-Grained Entity Typing (FET) methods in recent years (Choi et al., 2018; Xiong et al., 2019; Dai et al., 2019; Lin and Ji, 2019; Jin et al., 2019; López et al., 2019). Xiong et al. (2019) captures label correlation by employing graph convolution network on label co-occurrence matrix. Dai et al. (2019) make use of an existing entity linker to obtain noisy external data in order to enrich and disambiguate mention representation. The author also use entity linking scores as additional features. Lin and Ji (2019) exploits type inter-dependency with latent type representation. Previous FET methods, however, only focused on news domain where text comes from news or wikipedia article and speech. *Heng* **[emphasize all of the previous work only focus on general news domain or wikipedia data]**

Chemistry text, however, is largely different from news in that it's not only heavy in domain-specific knowledge, but also often has complex mention names not following morphological structure (sometimes alphanumeric encoding). Up until now, there has been no dataset or work in chemical FET, which is important for mining compound entities from unstructured chemistry literature. We have not only built a new benchmark, but also developed an effective FET framework that incorporates external structure and text knowledge. *Heng* **[not sure what you mean by 'external structural'?]** *Cheng* **[It seems that "complex mention names" is a main new technical challenge. It would be great to amplify this message throughout the paper including a discussion of why the proposed model/architecture/framework can be expected to address this challenge and some empirical results to show the effectiveness in addressing the challenge (e.g., an example where previous methods failed to work because of the complex mention names but the proposed new method worked better. Another way to discuss the novelty here is to say the previous methods have NOT fully exploited the opportunity in our problem domain (e.g., external information/knowledge) and we propose a FET framework to enable full exploitation of external resources.]**

5.2 Multimodal Representation

Heng **[I deleted the following paragraph because it's too high-level and verbose.]**

Multi-modal knowledge representation methods have been widely applied to tasks such as visual question answering and cross-modal retrieval between image and text. One line of deep-learning based alignment methods (Diao et al., 2021; Wei et al., 2020; Ye et al., 2019; Nam et al., 2017; Li et al., 2020) involves cross-modal alignment between separately learned word and image region representation. A recent popular line of research, including VisualBERT (Li et al., 2019b) and VL-BERT(Su et al., 2019), integrates the reasoning process into pretraining, inspired from (Devlin et al., 2018). These models are fed with image-caption pairs and proceed to align regions and phrases by attention mechanism.

Different from alignment among image, text, and audio, our method involves alignment between structures and description text, which is a phenomenon specific to chemistry and has hardly been explored in previous work.

Cheng **[here it sounds like we are exploring a different kind of alignment which has not been studied before. "hardly" is vague; try to make it more specific. E.g., can we confidently say that it has NEVER been explored? or perhaps it has been explored, but we do it in a BETTER way (be specific in terms of where it's better)?]**

5.3 Knowledge-Enhanced Language Representation

Heng **[a lot more related work needs to be added in here. Check the related work in https://arxiv.org/pdf/2012.15022.pdf]** Recently, there has been a lot of work (Peters et al., 2019; Zhang et al., 2019; Qin et al., 2020; He et al., 2020; Liu et al., 2020; Yang and Mitchell, 2019; Wang et al., 2020, 2021; Xu et al., 2021) on incorporating external knowledge into language understanding. In (Liu et al., 2020), triples are injected into the sentences as domain knowledge and attach to the tokens in the sentence. (Liu et al., 2020), on the other hand, embeds words with KB concepts in an LSTM framework.

As a unique contribution, our work is the first to draw a line between local context and external (molecular) structural information

Cheng **[is there a potential here to make this idea even more general? Can we say the general idea is to leverage non-textual external resources/information?]**

6 Discussion

While we applied the multimodal entity representation technique to fine-grained chemical entity typing, the idea can be well generalized to other ChemIE tasks such as relation extraction and reaction event extraction, in which chemical entities play a major role. We will release new datasets on other ChemIE task in the near future.

7 Conclusions and Future Work

In this work, we take the first step to explore the task of fine-grained entity typing in chemistry domain and introduced a dataset, CHEMET, to facilitate the study of the task. Meanwhile, we also developed a deep-learning based model that effectively incorporates external multimodal information of chemical mentions to improve the model's understanding on chemistry text, and showed through experiments that our model achieved state-of-the-art on the dataset. We would like to point out that the multimodal entity representation can be applied to other ChemIE tasks.

One big challenge from our findings is that many chemicals cannot be linked to external database, either due to its varying mention form or the database simply does not contain that particular entity (which is relatively more obvious for newer chemistry articles). In the future, we will develop entity linking algorithm to not only match mention to database better but also do cross-document linking (i.e., retrieve context for a mention from other documents).

Table 3: Multi-column table

Multi-column	
Multi-column	
X	X

Table 4: Multi-column table

Title	Category A			Category B		
	Item 1	Item 2	Item 3	Item 1	Item 2	Item 3
X	1	2	3	1	2	3
Y	1	2	3	1	2	3

Title	Category A			Category B		
	Item 1	Item 2	Item 3	Item 1	Item 2	Item 3
X	1	2	3	1	2	3
Y	1	2	3	1	2	3

Table 5: Transductive Imputation AUC with 10% missing data

Model	Dev			Test		
	Accuracy	Macro F1	Micro F1	Accuracy	Macro F1	Micro F1
BioBERT	1	2	3	1	2	3
SciBERT	1	2	3	1	2	3
Y	1	2	3	1	2	3
Our Model	1	2	3	1	2	3

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Hyejin Cho and Hyunju Lee. 2019. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics*, 20(1):1–11.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. *arXiv preprint arXiv:1807.04905*.

Hongliang Dai, Donghong Du, Xin Li, and Yangqiu Song. 2019. Improving fine-grained entity typing with entity linking. *arXiv preprint arXiv:1909.12079*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. *arXiv preprint arXiv:2101.01368*.

Qizhen He, Liang Wu, Yida Yin, and Heming Cai. 2020. Knowledge-graph augmented word representations for named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7919–7926.

David M Jessop, Sam E Adams, Egon L Willighagen, Lezan Hawizy, and Peter Murray-Rust. 2011. Oscar4: a flexible architecture for chemical text-mining. *Journal of cheminformatics*, 3(1):1–12.

Hailong Jin, Lei Hou, Juanzi Li, and Tiansi Dong. 2019. Fine-grained entity typing via hierarchical multi graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4970–4979.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Diya Li, Lifu Huang, Heng Ji, and Jiawei Han. 2019a. Biomedical event extraction based on knowledge-driven tree-lstm. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1421–1430.

Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, 18(1):1–11.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proc. The 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*.

Ying Lin and Heng Ji. 2019. An attentive fine-grained entity typing model with latent type representation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6198–6203.

Feifan Liu, Jinying Chen, Abhyuday Jagannatha, and Hong Yu. 2016. Learning for biomedical information extraction: Methodological review of recent advances. *arXiv preprint arXiv:1606.07993*.

Sijia Liu, Feichen Shen, Ravikumar Komandur Elayavilli, Yanshan Wang, Majid Rastegar-Mojarad, Vipin Chaudhary, and Hongfang Liu. 2018. Extracting chemical–protein relations using attention-based neural networks. *Database*, 2018.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Federico López, Benjamin Heinzerling, and Michael Strube. 2019. Fine-grained entity typing in hyperbolic space. *arXiv preprint arXiv:1906.02505*.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307.

Dat Quoc Nguyen, Zenan Zhai, Hiyori Yoshikawa, Biaoyan Fang, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Saber A Akhondi, Trevor Cohn, Timothy Baldwin, et al. 2020. Chemu: named entity recognition and event extraction of chemical reactions from patents. In *European Conference on Information Retrieval*, pages 572–579. Springer.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Hoifung Poon and Lucy Vanderwende. 2010. Joint inference for knowledge extraction from biomedical literature. In *Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 813–821.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2020. Erica: Improving entity and relation understanding for pre-trained language models via contrastive learning. *arXiv preprint arXiv:2012.15022*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Yuanhe Tian, Wang Shen, Yan Song, Fei Xia, Min He, and Kenli Li. 2020. Improving biomedical named entity recognition with syntactic information. *BMC bioinformatics*, 21(1):1–17.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10941–10950.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wenhan Xiong, Jiawei Wu, Deren Lei, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Imposing label-relational inductive bias for extremely fine-grained entity typing. *arXiv preprint arXiv:1903.02591*.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Song Xu, Haoran Li, Peng Yuan, Yujia Wang, Youzheng Wu, Xiaodong He, Ying Liu, and Bowen Zhou. 2021. K-plug: Knowledge-injected pre-trained language model for natural language understanding and generation in e-commerce. *arXiv preprint arXiv:2104.06960*.

Bishan Yang and Tom Mitchell. 2019. Leveraging knowledge bases in lstms for improving machine reading. *arXiv preprint arXiv:1902.09091*.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10502–10511.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

A Dataset Ontology

Chemistry	Organic_Chemistry	Organic_Compounds	Aromatic_Compounds	
Chemistry	Organic_Chemistry	Organic_Compounds	Aromatic_Compounds	Aryl_Groups
Chemistry	Organic_Chemistry	Organic_Compounds	Carbenes	
Chemistry	Organic_Chemistry	Organic_Compounds	Esters	
Chemistry	Organic_Chemistry	Organic_Compounds	Ethers	
Chemistry	Organic_Chemistry	Organic_Compounds	Hydrocarbons	Alkanes
Chemistry	Organic_Chemistry	Organic_Compounds	Hydrocarbons	Alkenes
Chemistry	Organic_Chemistry	Organic_Compounds	Hydrocarbons	Alkynes
Chemistry	Organic_Chemistry	Organic_Compounds	Organic_Acids	Carboxylic_Acids
Chemistry	Organic_Chemistry	Organic_Compounds	Organic_Acids	Phosphonic_Acids
Chemistry	Organic_Chemistry	Organic_Compounds	Organic_Acids	Phosphinic_Acids
Chemistry	Organic_Chemistry	Organic_Compounds	Organic_Acids	Sulphinic_Acids
Chemistry	Organic_Chemistry	Organic_Compounds	Organic_Acids	Sulphonic_Acids
Chemistry	Organic_Chemistry	Organic_Compounds	Organohalides	
Chemistry	Organic_Chemistry	Organic_Compounds	Organometallic_Compounds	
Chemistry	Organic_Chemistry	Organic_Compounds	Organonitrogen_Compounds	Amides
Chemistry	Organic_Chemistry	Organic_Compounds	Organonitrogen_Compounds	Amines
Chemistry	Organic_Chemistry	Organic_Compounds	Organonitrogen_Compounds	Nitriles
Chemistry	Organic_Chemistry	Organic_Compounds	Organonitrogen_Compounds	Nitro_Compounds
Chemistry	Organic_Chemistry	Organic_Compounds	Heterocyclic_Compounds	
Chemistry	Organic_Chemistry	Organic_Compounds	Organophosphorus_Compounds	Phosphinic_Acids_And_Derivatives
Chemistry	Organic_Chemistry	Organic_Compounds	Organophosphorus_Compounds	Phosphonic_Acids_And_Derivatives
Chemistry	Organic_Chemistry	Organic_Compounds	Organosulfur_Compounds	Sulfonic_Acids
Chemistry	Organic_Chemistry	Organic_Compounds	Polycyclic_Organic_Compounds	
Chemistry	Organic_Chemistry	Organic_Compounds	Organic_Polymers	
Chemistry	Organic_Chemistry	Organic_Compounds	Reactive_Intermediates	
Chemistry	Organic_Chemistry	Organic_Compounds	Thiols	
Chemistry	Organic_Chemistry	Organic_Compounds	Other_Compounds	
Chemistry	Organic_Chemistry	Functional_Groups	Carbonyl_Group	Ketones
Chemistry	Organic_Chemistry	Functional_Groups	Carbonyl_Group	
Chemistry	Organic_Chemistry	Functional_Groups	Hydroxyl_Group	
Chemistry	Organic_Chemistry	Functional_Groups	Acyl_Groups	
Chemistry	Organic_Chemistry	Functional_Groups	Amides	
Chemistry	Organic_Chemistry	Functional_Groups	Amines	
Chemistry	Organic_Chemistry	Functional_Groups	Esters	
Chemistry	Organic_Chemistry	Functional_Groups	Ethers	
Chemistry	Organic_Chemistry	Functional_Groups	Nitriles	
Chemistry	Inorganic_Chemistry	Inorganic_Compounds		
Chemistry	Chemical_Reactions	Catalysis	Catalysts	

Figure 4: Ontology Screenshot