Unleashing the Power of Knowledge Extraction from Scientific Literature in Catalysis

Yue Zhang,[*,†,‡] Cong Wang,[¶,‡] Mya Soukaseum,[§,‡] Dionisios G. Vlachos,[*,¶,‡] and Hui Fang[*,†,‡]

†*Department of Electrical and Computer Engineering, University of Delaware, Newark, Delaware, 19711, United States*
‡*Center for Plastics Innovation, University of Delaware, Newark, Delaware, 19711, United States*
¶*Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware, 19711, United States*
§*Department of Chemical and Biological Engineering, Drexel University, Philadelphia, 19104, United States*

E-mail: zhangyue@udel.edu; vlachos@udel.edu; hfang@udel.edu

Abstract

Valuable knowledge of catalysis is often hidden in a large amount of scientific literature. There is an urgent need to extract useful knowledge to facilitate scientific discovery. This work takes the first step toward the goal in the field of catalysis. Specifically, we construct the first information extraction benchmark data set that covers the field of catalysis, and also develop a general extraction framework that can accurately extract catalysis-related entities from scientific literature with 90% extraction accuracy. We further demonstrate the feasibility of leveraging the extracted knowledge to help users better access relevant information in catalysis through an entity-aware search engine and a correlation analysis system.

1 Introduction

Plastic waste is a major environmental threat that has accumulated rapidly due to a lack of natural decomposition mechanisms and challenges in recycling.[1,2] Catalytic chemical recycling and upcycling are energy-efficient and free of carbon burden.[3] They are prime solutions to combat this global crisis and have recently drawn enormous attention to the research frontiers.[4–14] For example, several recent studies have demonstrated facile conversions of polyolefin-based plastics into value-added fuels and lubricants at mild conditions.[4–9] However, since the literature in catalytic plastic conversion has been proliferating only recently, the directly relevant knowledge is limited and changing dynamically. Researchers often extract useful information from past literature on small alkane hydroconversions and other partially relevant topics to compensate for the lack of reference. However, such manual searches are time expensive and often biased due to the large volumes and sampling errors.

The advances in Artificial Intelligence, in particular in Natural Language Processing (NLP) and Information Retrieval (IR), make it possible to systematically and efficiently extract knowledge related to catalytic plastic conversion from a large amount of scientific literature and leverage the extracted knowledge to facilitate the relevant information access and discover new knowledge.

Information extraction, like all machine learning techniques, is largely data-driven, requiring benchmark data sets for training and testing. Although chemical information extraction has received a lot of attention in recent years, few studies have focused on extracting multiple types of catalysis-related entities. CHEMDNER[15] focused on general chemical entity extraction and did not distinguish the specific role an entity plays in the catalytic process, e.g., a reactant vs. a product. ChEMU[16] is a corpus constructed for identifying paragraphs related to a single reaction from patents. Its data source is a domain with a very different writing style from scientific literature, and the task is not to extract catalyst-related entities. More recent studies centered around the articles related to material science[17,18] and focused on extracting information such as summary-level information,[19] synthesis route and procedure parameters.[20–26] These studies were mostly constrained to certain types of paragraphs, like abstract[19] or synthesis-related,[20–22,24–26] which might greatly hinder the extraction performance. Thus, there is no suitable data set for catalysis-related information extraction from scientific literature.

In this work, we construct the first benchmark data set for knowledge extraction from scientific literature in catalysis. In particular, we focus on extracting information from full-text of journal articles, and the extracted information covers six types of catalysis-related entities: Catalyst, Reaction, Reactant, Product, Characterization, Treatment. To collect relevant full-text articles, an *article selection* module is developed with the key component as a binary classifier that selects highly relevant articles related to catalysis. With the collected relevant articles, the next challenge is to annotate them based on the six entity types. This is not a trivial task as the annotators would need to go through every sentence in the article and label the entities with one of the specific types. As the numbers of annotated examples could vary significantly across different types, it would be quite labor-intensive to get enough annotations for each entity type. To overcome this limitation, we propose to reduce the unnecessary annotation effort with the help of active learning in a *data annotation* module. The goal of active learning is to collect more annotations for the entity type with fewer examples and quickly improve model performance through prioritizing the annotation of the sentences with the biggest uncertainty.

Furthermore, we develop a robust *entity extraction* model to extract the entities with the six specified types. Unlike the commonly used Conditional Random Field (CRF) models,[19] we formulate the problem as a span classification problem that automatically extracts a text span and labels it with one of the specified entity type. Furthermore, we also propose to pre-train the extraction model so that it can adapt to the domain task and achieve better extraction performance. Compared with existing work, our extraction model is robust with respect to the amount of training data and it can achieve fairly high extraction accuracy (around 90%) with a small set of annotated training data.

Finally, with the extracted information, we also develop two online systems that facilitate researchers to navigate the catalysis information more easily. The first system is an entity-aware search engine. The interface is similar to existing literature search engines, but the results are better visualized and can be further refined with the aid of the extracted entities. In particular, all the extracted catalysis-related entities are highlighted in the search results with different colors. On the side bar, top extracted entities for each type are displayed and can be selected to narrow down the search results in a more efficient way. The second system is an entity correlation analysis system with the goal of identifying related catalysis information, and the correlation of two entities is roughly estimated based on their co-occurrences in the data set. Both systems have been analyzed and shown to be useful for information access and knowledge discovery.

2 Methodology Overview

The main goal of this work is to develop a machine learning model that can automatically extract catalysis-related information from scientific articles. There is a lot of useful information we can extract from the articles related to catalysis, and this work focuses on extracting the information with the following six types: Catalyst, Reaction, Reactant, Product, Characterization, Treatment.

Machine learning algorithms rely on annotated data set to train the model. To construct the annotated data set, two annotators, both with training in chemical engineering and catalyst, went through each article and labeled the information with the above six entity types. We summarize the annotation guideline briefly in Table 1.

Figure 1 shows the overview of the developed extraction framework. Starting with a collection of literature in catalysis downloaded from online publishers' API, the first module consists of a binary classifier, which selects highly relevant articles. The highly relevant articles are then annotated by the domain experts. To ensure the quality of the annotations while minimizing the amount of data that needs to be annotated, an active learning method is developed for the data annotation component. Finally, an entity extraction component is developed to learn from the provided training data and automatically extract different types of catalysis-related entities from the articles. Within the extraction component, a domain-adaptive pre-training strategy is applied to improve the extraction accuracy. We provide more details about each component in the following sections.

3 Constructing Benchmark Data Set

3.1 Data Crawling and Preprocessing

As in any work related to scientific literature, a set of relevant data need to be first collected and downloaded from online publishers. We crawled the data from Elsevier for this work, but the general methodology is applicable to other online publishers.

To gather relevant articles related to catalytic plastic up-cycling, we first prepare a list of keywords covering the catalysts, reactants, reactions, mechanisms that are frequently discussed in catalytic plastic up-cycling. These keywords are then used as queries to collect articles through Science Direct's Search and Article Retrieval API.[27] Additional queries are also generated by concatenating keywords from two different categories together(e.g reaction + reactant, mechanism + catalyst, etc.) In total, 453 queries were used to collect relevant articles from Elsevier, and up to 6K articles are returned for each query. This results in a collection of 344,093 articles related to catalysis. The crawled articles are stored in MongoDB and uniquely identified using their DOI Identifiers.

The collected articles need to go through a series of steps to be transformed into the format that our model can take as input. ScienceDirect's Article Retrieval API returns articles that are represented in XML format and thus need to be parsed by LimeSoup[28] before text and metadata(i.e. journal, title, keywords) extraction. After removing tables and figures, each article is broken into a list of sections, each containing the corresponding title and text. Additionally, we identify the abstract by sequentially searching all sections and picking the first paragraph that has "Abstract" inside its title. After these steps, we have collected a large catalysis literature corpus with 7.8 Gigabytes of raw text. This raw text corpus is not only used to identify relevant articles as described in Section 3.2 but also used for domain adaption in Section 4.

Furthermore, text content goes through the sentence splitting and tokenization process using Stanza toolkit.[29,30] We choose to use Stanza's CRAFT biomedical model because its tokenization accuracy is on par with ChemDataExtractor[31] and can split words on hyphens, which is consistent with how BERT tokenizes its text corpus during pre-training.

3.2 Article Selection

Although ScienceDirect's Search API was used to download articles related to the queries, there are still many non-relevant articles in the corpus. For example, many articles are envi-

Table 1: Annotation Guideline

Catalysts	Identified as "metal/support," or with keywords like "metal-catalyst." If details about the catalyst composition are provided, then they are included in the catalyst text span (e.g., Ru/CeO_2, CeO_2-supported metal catalysts, Pt/H-USY (Pt:1wt.%) catalysts).
Reactants	Species that interact with the catalyst to create a product (e.g., polyethylene, plastics, PE). Reagents for catalyst synthesis are not included in this category.
Products	Species that are produced from a chemical reaction between the reactant and the catalyst (e.g., C_1-C_4, coke, liquid fuel). Intermediate species that go on to react further are not considered in this category unless there is substantial characterization/quantification of those species.
Reactions	Processes that involve the transformation of a chemical species via interactions with a catalyst (e.g., hydrogenation, isomerization, hydrocracking).
Characterizations	Any technique that is used to yield useful information about the catalyst (e.g., gas chromatography mass spectroscopy (GC-MS), powder x-ray diffraction (XRD), infrared spectrometry (IR)).
Treatments	Any intermediate steps taken to prepare the catalyst (e.g., heating, calcination, refluxing).

ronmental reports written by researchers outside the field of chemistry or polymerization research. As the data collection needs to contain knowledge related to catalytic process useful for plastic up-cycling, these articles are not useful and only add noises to the corpus.

To remove non-relevant articles in the downloaded collection, we develop a binary text classifier to predict the relevance of each article using its abstract as input. More specifically, the abstract is represented using BERT, a state-of-art pre-trained transformer-based model.[32] Following the standard procedure, we take the output of the "[CLS]" token and send it through a two-layer perceptron to make the prediction. The classifier will output a confidence score indicating to what extent the model believes the article is relevant. The classifier is trained using standard cross-entropy loss on a training data set. The training data set consists of 103 labeled highly relevant articles and 10 times more unlabeled articles randomly sampled from the whole corpus.

The developed classifier has been evaluated and shown satisfying performance. In particular, 10K articles were randomly sampled from the corpus, and the trained model is then used to rank the articles based on the confidence scores. Top-ranked articles are manually judged based on their relevance, and we found the classifier can accurately assign relevant articles with higher confidence scores.

As the classifier has shown to be effective, it is then applied to the crawled data collection and all the articles with confidence scores higher than 0.90 are considered highly relevant. After this process, we are able to identify 3,578 highly relevant articles which contain 440K sentences. All our further data annotation and extraction were carried out using this collection.

3.3 Data Annotation

The corpus needs to be annotated based on the guidance described in Table 1. Two co-authors of this work, specialized in catalysis, reviewed the corpus in every sentence, labeled and assigned relevant information into each of the six categories based on their expertise. In addition to the annotation guidelines, the annotators also follow the following two specific criteria.

- *Label only non-overlapping continuous text spans.* When a text span includes

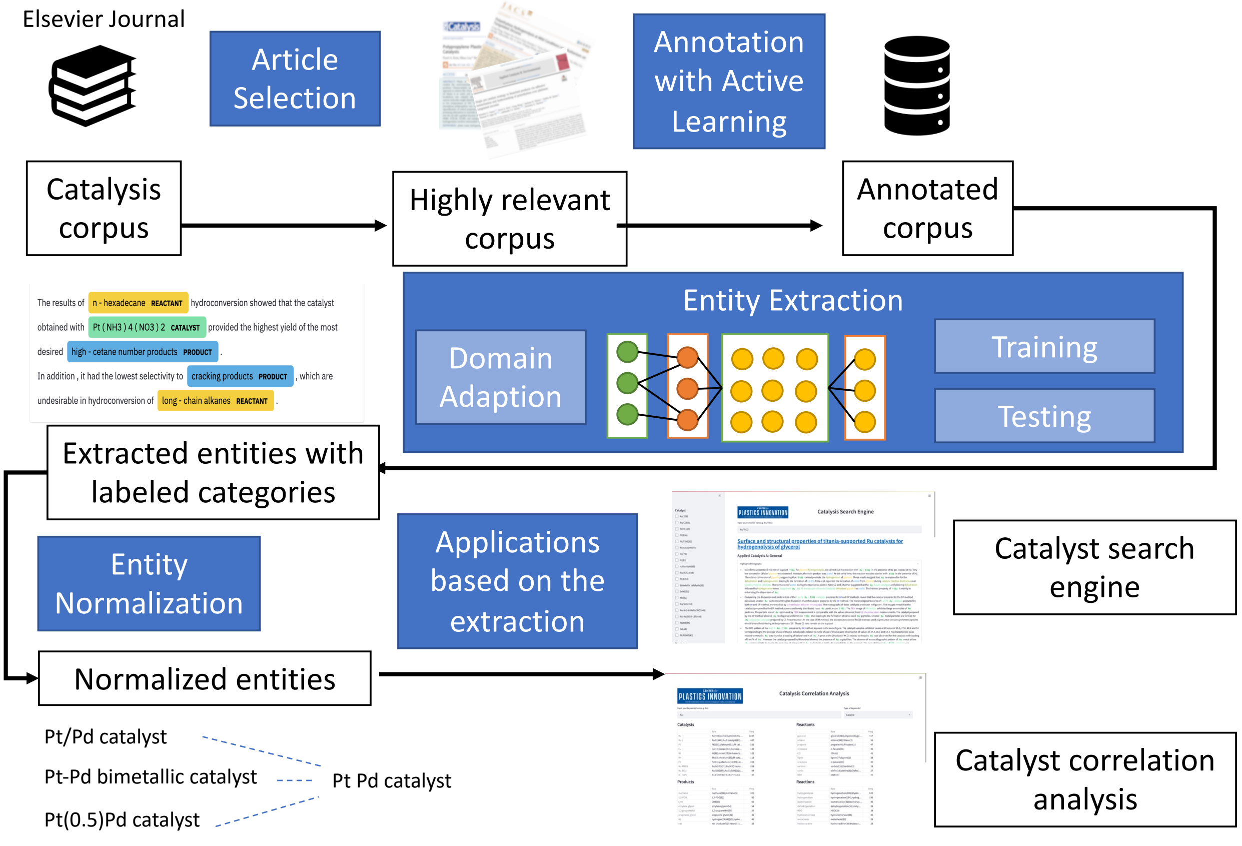

Elsevier Journal
Article Selection
Annotation with Active Learning
Catalysis corpus
Highly relevant corpus
Annotated corpus
The results of n - hexadecane REACTANT hydroconversion showed that the catalyst obtained with Pt (NH3) 4 (NO3) 2 CATALYST provided the highest yield of the most desired high - cetane number products PRODUCT .
In addition , it had the lowest selectivity to cracking products PRODUCT , which are undesirable in hydroconversion of long - chain alkanes REACTANT .
Entity Extraction
Domain Adaption
Training
Testing
Extracted entities with labeled categories
Entity Normalization
Applications based on the extraction
Catalyst search engine
Normalized entities
Catalyst correlation analysis
Pt/Pd catalyst
Pt-Pd bimetallic catalyst
Pt(0.5)Pd catalyst
Pt Pd catalyst

Figure 1: Overview of the proposed framework

mentions of multiple entities, we treat it as a single continuous span and do not consider any nested or discontinuous spans. For instance, "mono- and bi-metallic Pt and Ni catalysts" are labeled as an entity with a single continuous span instead of being split into three overlapping discontinuous spans: "mono-metallic Pt catalysts", "mono-metallic Ni catalysts", and "bi-metallic Pt and Ni catalysts". This criterion is used because (1) a single long span is often good enough to cover all the keywords needed for searching and (2) predicting nested or discontinuous spans will greatly increase the number of total possible outcomes, requiring more annotated data and leading to inferior model performance when available data is scarce.

- *Label an entity and its corresponding references.* An entity might be referred in different ways. Using abbreviations is the most common case, but there are also complicated cases. For example, $RuSiO_2$ is referenced as "ruthenium catalyst" when an author summarizes its property together with $RuTiO_2$ and $RuAl_2O_3$ in the abstract. However, the author may use the complete notation "$RuSiO_2$ sample" in the synthesis paragraph while simplifying it to "ruthenium" in the discussion of the isomerization reaction in the experiment results. Sometimes the complete composition of a catalyst is even hidden in a noun clause e.g. "TiO_2 added to Ru/SiO_2 or Ru/Al_2O_3". Such non-standard, vague referring to the same catalyst occurs many times in the annotated data set. To ensure better coverage of the extracted information, we decide to include the partial mentioning in the annotations and plan to use techniques like entity linking[33] to group all these different expressions together in our future study.

Although the examples above are about cat-

Table 2: Number of labels for each category in the annotated data sets

Categories	*Abstract*	*Full-text*
Catalyst	1125	5029
Characterization	207	708
Product	732	5108
Reactant	1322	4336
Reaction	1010	3468
Treatment	119	643

alysts, both rules are applicable to all entity types. For example, an article might have overlapped text spans when enumerating product mixtures or reactant mixtures. Partial mentions are also widely used to describe the result of characterization since the full name is often too long.

Another important decision related to data annotation is about the annotation unit. Shall the annotations be made for only abstracts or the full-text of the articles? Previous studies[15,19] often annotated only abstracts due to the increasing amount of effort required to annotate full-text. However, full-text articles certainly provide more information than abstracts.

In our work, we annotated both abstracts and full-text. Table 2 summarizes the annotation statistics of these two sets. The annotations for abstracts are based on 185 abstracts that were randomly chosen from the highly relevant corpus as described in Section 3.2. We did not include the annotations for more abstracts because the number of annotations is already sufficient to train an effective extraction model. The annotations for full-text articles are based on 50 articles, which are randomly selected out of the 185 articles whose abstract has been labeled. The full-text annotation excluded the introduction paragraphs because they often contain a large number of related work discussions and can be largely covered by the rest of the articles.

3.4 Active Learning

One observation made based on Table 2 is that the number of labels for each category is not evenly distributed. Intuitively, the categories with fewer labels, such as Treatment, may have inferior performance due to the insufficient training examples. A simple solution would be to keep labeling more articles. However, as shown in Figure 2, the average number of "Treatment" spans per sentence is rather small. It is not cost-effective to get more treatments labels through labeling more articles. To address this challenge, we propose an active learning approach to get more treatments labels efficiently.

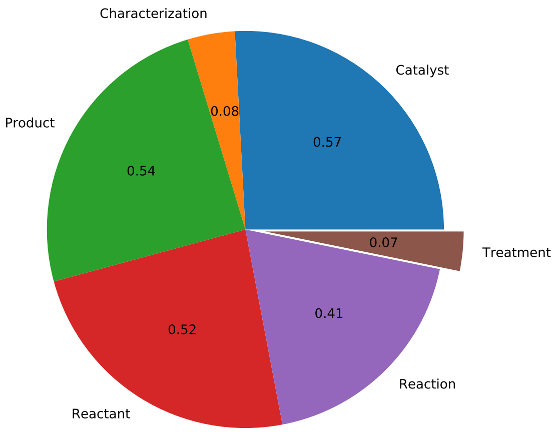


Figure 2: Average number of entities per sentence

Traditionally, data annotation is carried out first, and then the collected examples are used to train the machine learning model, which is often referred to as "supervised learning".[19] There is no feedback between the data annotation phrase and the model training phase. Active learning breaks this boundary by actively selecting the samples that are sent to annotators for labeling.[34] Compared with supervised learning, active learning can greatly reduce the number of training examples needed for the training process. The intuition behind active learning is simple: we should prioritize the labeling of examples that could potentially contribute most to the model training process. We can take the model's confidence score of its prediction as a good indicator of the examples' potential contribution. Similar to how students learn from mistakes, the model cannot learn much if we only confirm its confident prediction. We want to pick the part of examples with the biggest prediction uncertainty so that human input can make more difference.

Specifically, we first create a candidate sentence pool using the unlabeled part of our corpus. The pool is filtered by only keeping sentences similar to those existing sentences with at least one Treatment text span. We heuristically set the confidence range based on observation on the training examples. The confidence score is estimated using a preliminary model trained only using abstract and full-text annotation. For each span, we use the treatment probability output as our confidence score. The statistic shows that over 70% of Treatment sentences are correctly predicted (with around 70% accuracy) by our model with a very high confidence score. In contrast, sentences with moderate confidence scores(12.5% to 25% percentile) have much higher prediction uncertainty and, at the same time, maintain an acceptable level of prediction accuracy to prevent annotators from wasting time labeling non-treatment-related sentences. With this strategy, 946 sentences are selected and distributed to annotators. Among all 1815 new labels obtained from these sentences, 80% is of Treatment, a much higher percentage compared with those reported in Table 3. This demonstrates the effectiveness of the proposed active learning strategy in augmenting data sets with more Treatment spans.

4 Entity Extraction

We formulate the chemical information extraction task as a named entity recognition (NER) problem. The input of the model is a sequence of words and the output is different types of entities identified by the model. As explained in Section 3.3, an entity is defined as a continuous non-overlapped text span in this work.

The NER problem can be reformulated as a span classification problem, where the goal is to assign a label to each possible span. Possible labels include the six entity types used in the annotation as well as a special type called "Others", covering all the spans that do not belong to any of the six entity types. The span detection model iterates through every possible span in a sentence and classifies it to one of the possible labels. As most spans are not entities, they will be labeled as "Others". Previous studies[35,36] have shown that span classification models can reach similar performance with traditional CRF-based NER methods and is much more robust against unlabeled entities. CRF reformulates the NER problem as a sequence labeling problem, where every token will be assigned a label and then decode the final span prediction from the label sequence. For example, given a labeling sequence "[O, B-Catalyst, I-Catalyst, E-Catalyst, O]", "B-Catalyst", "I-Catalyst", "E-Catalyst", each represents the corresponding token is the beginning, internal, ending of a Catalyst span, respectively. And from the labeling sequence, a catalyst entity located between the second and fourth words of the sentence can be identified. Unlike the span classification model which can simply skip unlabeled entities during training, CRF works on the sequence level and can only skip the whole sentence. Moreover, the span classification model has the advantage that it can assign a confidence score for every possible span. This is very helpful for active learning (as described in Section 3.4) since it allows us to directly locate uncertain span predictions. On the other hand, the commonly used CRF models[19,26] can only assign a confidence score for all span predictions in the same sentence. When only specific types of entities are of interest, such global confidence scores are not useful. Thus, in this work, we adapt the span detection model[35,36] to the chemical domain and develop our span classification model accordingly.

The proposed span detection model consists of two parts: span representation and span scoring. The span representation module takes a text input, converts it into a machine-readable format, and extracts vector representations for each span. All the span representations are passed to the span scoring module and generate a probabilistic distribution for the label predictions over all types. The span scoring module is based on a biaffine scoring function. During the model training, the model's parameters will be optimized to maximize the data likelihood using the gradient descent algorithm. When testing, the list of probabilistic distributions is used to identify the most likely type for each candidate entity. Furthermore, domain adaption techniques are applied in pre-training to avoid the problem of domain shifting. Finally, all the extracted text spans are normalized to reduce the duplicates in the extraction results.

4.1 Span Representation

Text representation has been a long-standing challenge in the NLP area. The traditional approach is to represent the text as a vector of indicator variables, and each indicator variable represents a handcrafted feature,[31,37–41] such as whether a word is metal or whether there is a "bimetallic" prefix in the word. With the recent introducing of neural NLP techniques(e.g. Word embedding, Long short-term memory(LSTM)), text representation can be automatically extracted,[19,20,24,42] and has been shown to be more effective than the handcraft patterns when there are enough training data.[43] In this work, we use a state-of-the-art transformer-based model, BERT,[32] to extract span representations. Since its proposal, BERT has claimed the best performance over a variety of natural language processing benchmarks[32,44] As shown in recent ChEMU benchmark,[45,46] BERT based model[26,47–49] have been widely used for chemical information extraction and become the dominating method. Lin et al.[50] provide two reason why BERT can work well on our catalysis information extraction task.

First, it is important to model "name regularity", i.e the vocabulary and naming convention used for certain entity types, in the chemical domain. For example, "Pt/TiO_2" covers two widely used noble metals and support material. And this information is a good indicator for identifying catalyst spans. BERT can address this problem by breaking words into subwords using the wordpieces algorithm. In this way, "$PtAl_2O_3$" is converted to ["Pt", "##Al", "##2", "##O", "##3"]. "##" prefix is added to all subwords except the first one as a way to distinguish them from individual words such as Pt and Al. With this strategy, the representations of "Pt/TiO_2" and "Pd/TiO_2" would be similar as they share a common subword. With the subword representation, the great diversity of chemical compound naming is reduced and therefore can be more effectively captured by the model, which can improve the model's generalization capability for new names. After subword tokenization, a subword is represented as a high-dimensional vector. Just as how we look up unfamiliar words in a dictionary, BERT looks up each subword in a huge vector table. This lookup table is stored as part of the model's parameters, working as the memory of the model. The resulting vector list becomes the initial context-free representation for each subword.

When representing a span, it is also important to consider its context, i.e., the text surrounding the text span. The context of a text span can offer lots of information about its entity type. For example, "[A] shows high selectivity for [B]" can serve as an indicator that [A] is highly likely to be a catalyst span and [B] should be some kind of product. BERT can model the context information by using a stack of transformer encoders to capture preceding or following text patterns. Each transformer encoder consists of two sub-layers: self-attention layer and feed-forward layers. Both sub-layers are connected together by LayerNorm and residual connections. The subword representations, which were context-free at the beginning, are recursively altered by their context as they pass through transformer encoders. At each residual connection, the contextual representation collected by the self-attention layer is added to the subword representation from the previous encoder. The self-attention layer works like a filter. For each subword, only information from its closely related neighbors is considered. It mainly captures phrase-level and syntactic-level relationships at lower-level encoders and composites more complicated semantic relationships at higher levels. In this way, the final representation for each subword is contextualized to reflect the surrounding text patterns.

With the contextualized subword representations, a text span is represented by a pair of subword representations, where the pair consists of the first subword in the first word of the span and the first subword in the last word of the span. As shown in Figure 3, "pentane isomers" contains two words, which can be split into three subwords. The representation of text span "pentane isomers" is the combination of the subword representation of "pent" (i.e., the first subword in the first word of the span) and that of "isomers" (i.e., the first subword in the last word of the span).

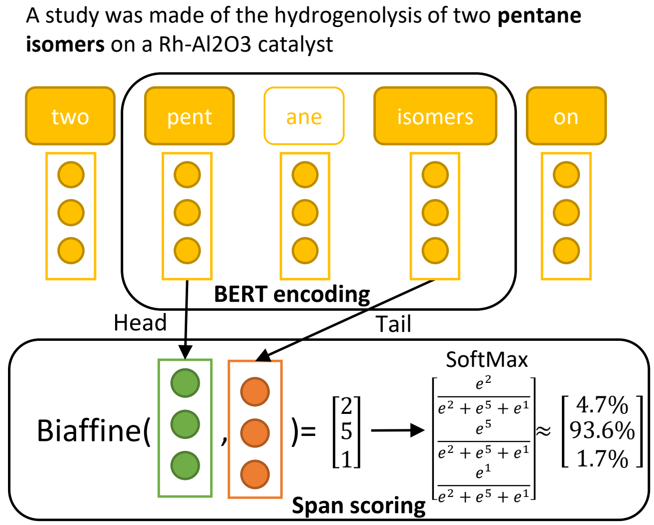

A study was made of the hydrogenolysis of two **pentane isomers** on a Rh-Al2O3 catalyst
two
pent
ane
isomers
on
BERT encoding
Head
Tail
Biaffine(
,
)=
$\begin{bmatrix}2\\5\\1\end{bmatrix}$
SoftMax
$\begin{bmatrix}\frac{e^2}{e^2+e^5+e^1}\\\frac{e^5}{e^2+e^5+e^1}\\\frac{e^1}{e^2+e^5+e^1}\end{bmatrix} \approx \begin{bmatrix}4.7\%\\93.6\%\\1.7\%\end{bmatrix}$
Span scoring

Figure 3: Span representation and scoring

4.2 Span scoring

The gathered span representations are sent to a biaffine scoring function.[51] At the beginning of the scoring process, all valid span representations are enumerated and extracted as the input of biaffine scoring function. Similar to the Cosine distance, the biaffine function measures how well the two subword representations are aligned together. For each entity type, the biaffine function captures a specific kind of bilinear alignment pattern and uses this pattern to score a span. Given a span with a pair of subword representations, i.e., $S_{ij} = [i, j]$, the raw alignment score for category k is denoted as $Score_k(S_{ij})$. All the raw scores are then converted into a probabilistic label distribution using the following softmax function.

$$O_{S_{ij},k} = \frac{e^{Score_k(S_{ij})}}{\sum_{k \in l} e^{Score_k(S_{ij})}}$$

where l denotes the set of all labels, including the six entity types as well as the "Others" category. Figure 3 shows an example when there are only 3 categories. The span representation of "pentane isomers" is first scored using biaffine scoring function for each category, and the scores are 2, 5, 1 respectively. These scores are then converted into a label probability distribution using the above softmax function. The results suggest the span should be labeled as the second category with 93.6% confidence.

4.3 Model Training

The span scoring function is learned from the training data. In particular, this work uses a variant of gradient descending algorithm called AdamW to train the model. The training objective is to maximize the logarithm likelihood of the observed data:

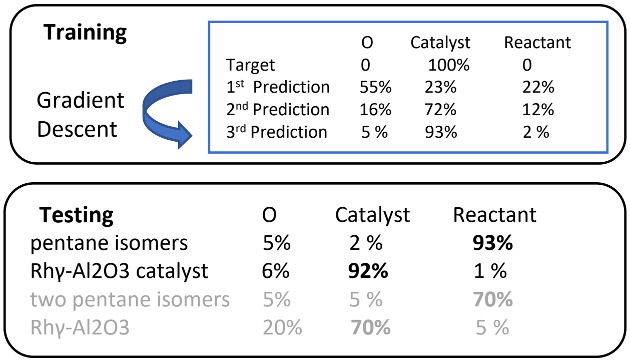
$$\sum_{(s_p,k_p)\in Pos} log(O_{s_p,k_p}) + \sum_{(s_n,k_n)\in Neg} log(O_{s_n,k_n})$$

where Pos denotes a set of all positive examples (i.e., the annotated entities), and (s_p, k_p) denotes the span representation (s_p) and the label (k_p) of an annotated entity. k_p is one of the six entity types. Similarly, Neg denoted a set of negative examples. The set is formed by sampling the spans that are not labeled. And k_n only has one possible value which corresponds to the "Others" type.

The top half of Figure 4 shows an example of the training process. For simplicity, we assume there are only two entity types: Catalyst and Reactant. And the third label "O" indicates the information that does not belong to any of the two entity types. Recall that the model training process is a process to optimize the model parameters so that the model's prediction gets closer to our target. In the example, after several rounds of optimizations, the probability for the Catalyst category becomes larger and larger, which is consistent with the target (i.e., 100% Catalyst) and demonstrates the progress of training a machine learning model.

4.4 Model Testing

The trained model can then be applied to the testing data and score every possible span. Given a sentence, the model outputs a list of every possible span together with the corresponding label probability distribution. Now the challenge is how to select a set of spans with the most confident prediction and no duplicates. We propose the following strategy to keep the most likely span prediction and remove all contradictory ones. First, all spans whose the most likely type prediction is "Others " are removed. The remaining spans are then sorted based on the confidence scores of their most likely type and pruned one by one from top to bottom. Furthermore, a span is kept if and only if it satisfies one of the two requirements: (1) it is the highest ranking span; (2) it has no overlapping with all the previous kept spans. An example of this process is illustrated in the bottom part of Figure 4, where the probabilities in bold font indicate the values used to rank those four spans. "Rh$\gamma-$Al$_2$O$_3$" is removed from the list because it overlaps with "Rh$\gamma-$Al$_2$O$_3$ catalyst" which has a higher confidence score.



	O	Catalyst	Reactant
Target	0	100%	0
1st Prediction	55%	23%	22%
2nd Prediction	16%	72%	12%
3rd Prediction	5 %	93%	2 %

Testing	O	Catalyst	Reactant
pentane isomers	5%	2 %	**93%**
Rhy-Al2O3 catalyst	6%	**92%**	1 %
two pentane isomers	5%	5 %	**70%**
Rhy-Al2O3	20%	**70%**	5 %

Figure 4: Model Training and Testing

4.5 Pre-training: Domain Adaption

Apart from training data and optimization method, the initialization of model parameters is also an important factor for the final performance, which is also demonstrated in chemical reaction extraction experiment.[26]

Before training, the parameters of span scorer are randomly initialized while the parameters of BERT were directly taken from those trained on a data set with different optimization goals. In BERT related literature, the process of training BERT on a specific task is often called "pre-training", while the training process described in the previous section is more widely referred to as "fine-tuning". Pre-training only optimizes the parameters of BERT, while fine-tuning optimize all the parameters. Pre-training usually relies on a large-scale text collection as the training data and does not require manual annotations. On the contrary, fine-tuning requires human-annotated task-specific training data. Compared with training everything from scratch, much fewer task-specific training data are needed through this "pre-training then fine-tuning" training paradigm. Therefore, we follow the same paradigm in this work.

Although pre-trained BERT may work well, it rarely produces the best performance due to the problem of domain shifting. Domain shifting means the pre-training and fine-tuning processes use data from different domains, which would decrease the effectiveness of knowledge transfer between the two processes. This is a common problem for scientific information extraction tasks, where the text collection is different from the one used for the pre-trained BERT. Google's BERT model is pre-trained on text corpus constructed by merging BookCorpus (an unpublished novel book corpus) and English Wikipedia. RoBERTa, an improved version of Google's BERT, includes more news articles, forum discussions, and stories into the training data set. Obviously, catalysis literature uses very different vocabularies and follows a more restricted writing style. SciBERT might be the closest pre-trained BERT model for our purpose. It was trained based on the scientific literature from multiple disciplines including medicine/biology, physics/mathematics, computer science, chemical, and so on. Unfortunately, the chemical related literature only takes up less than 10% of the SciBERT's training corpus.

The problem of domain shift can be solved by either training a BERT variant from scratch using a large-scale chemical literature corpus or adding an additional domain adaption process between the pre-training process and the fine-tuning process. Training a BERT variant from scratch requires a huge amount of chemical literature and hardware resources to finish, so we choose to adapt SciBERT to the chemical domain through pre-training it on a large-scale catalysis corpus.

We use the same highly relevant subset of our catalysis corpus as the annotation process since this is better aligned with the end task. After removing the text chunks that are too short, our pre-training data set contains 10.4M words and has a size of 63.7M bytes. During pre-training, the parameters of SciBERT are optimized to solve the masked language modeling (MLM) problem. Masked language modeling problem is designed to mimic the Cloze test in English class. Each time, the model takes a sentence where some part of the sentence is randomly masked. The subword embeddings and transformer encoders are trained and optimized to correctly predict the masked part based on surround context. Moreover, following the strategy used in RoBERTa, we make the MLM problem harder by masking an entire word (which might include several subwords) at once instead of just masking one subword. As the pre-training continues, SciBERT automatically learns chemical related terms and becomes ready to use for catalysis related information extraction.

4.6 Post processing: Text normalization

The raw form of extracted entities can be quite messy because a material, an instrument, or a concept can be described in multiple ways. Although it might not be difficult for a domain expert to tell two expressions represent the same concept, it is a very challenging task for computers if only exact matching is used to determine the matching of a concept.

To consolidate the expressions and remove duplicates, we propose the following text normalization strategy. First, we lowercase words that are not abbreviations and then lemmatize all plural words back to their singular form. The next step is to merge alternative expressions that look quite different in text. For example, "Beta Zeolite" and "β Zeolite", "hydrogen" and "H_2" are two pairs of synonyms. Besides, there exist a lot abbreviations(e.g. "PE" and "polyethylene", "Ru" and "ruthenium") in the extracted results. We merge all these expressions using a handcrafted dictionary. For catalysts, additional normalization is needed since people use different ways to connect different compositions when they write the catalyst formula, like "Pt/ZSM-5", "Pt/ZSM5". Special care is also taken to remove the weight percent notation(e.g. 1 wt% Pt/ZSM5) and the element valence notation (e.g. Cu(III)). For reaction and treatment, the tense of extracted result(e.g. hydrocracked and hydrocracking) also needs to be normalized. The final step is to remove common words that contribute little in terms of distinguishing those entities with other simpler expressions. They include cases like common suffix (e.g. catalyst, product) and connecting words(e.g. supported, containing). It is worth noting that certain phrases with very high frequency, like "hydrocarbon", "bimetallic catalyst", are also removed. They are meaningful especially when you read them in context, but, for simple context-free analysis, they have little use.

5 Evaluation of Extraction Models

This section describes the quantitative evaluation we have conducted to evaluate the effectiveness of the proposed entity extraction model over the annotated data sets. Section 6 will describe the extraction results over the large highly relevant article collection and explain how the extracted entities can be used to better facilitate researchers to access useful information.

5.1 Experiment Setup

Evaluation measures: We use standard metrics including P (Precision), R (Recall), and $F1$ to quantitatively measure the performance of entity extraction models. Precision computes the percentage of true positive results among all the predicted positive results, while recall computes the percentage of true positives among all the annotated results. F1 is a single measure that balances the trade-off of precision and recall, and is computed using the harmonic mean of precision and recall.

When deciding whether an extracted entity matches the annotated ones, one criterion is to simply rely on the *exact* matching. Under this criterion, given an extracted entity, if it has the correct type but its text span only partially overlaps with the correct entity, it will not be considered as a match. This criterion is referred as *strict* in this work. With this criterion, the evaluation measures are denoted as P_{strict}, R_{strict} and $F1_{strict}$ respectively.

However, such a strict criterion might not fully reflect the model's usefulness for chemical information extraction since a Chemical entity's name could have many variations. To take this problem, we propose another matching criterion that allows partial matching to be counted. We define the similarity between two text spans as follows:

$$S(Span_i, Span_j) = \begin{cases} 0 & different\ type \\ \frac{|Span_i \cap Span_j|}{|Span_i|} & same\ type. \end{cases}$$

We assume every prediction has a unit time cost and every annotation has a unit utility. We define P_{soft} as a measurement of time spent on reading through predictions that are worthwhile. Similarly, R_{soft} is defined as a measurement of how many total possible utilities you can collect by reading through the predictions.

$$P_{soft} = \frac{\sum_{i \in pred} \sum_{j \in gold} S(Span_i, Span_j)}{\sum_{i \in pred} 1}$$

$$R_{soft} = \frac{\sum_{i \in gold} \sum_{j \in pred} S(Span_i, Span_j)}{\sum_{i \in gold} 1}$$

where *pred* means all the predicted entities and *gold* means all the annotated entities.

$F1_{soft}$ is calculated as a harmonic mean of P_{soft} and R_{soft}.

$$F1_{soft} = \frac{2 * P_{soft} * R_{soft}}{P_{soft} + R_{soft}}$$

Note that P_{soft}, R_{soft} differs from token-level precision and recall by ensuring every entities has equal time cost or utility. Simple token-level precision and recall will put more weight on longer entities, which is not desired. Compared with soft evaluation metrics, strict evaluation metrics are special cases when $S(Span_i, Span_j)$ is set to 1 for exact matching and 0 otherwise.

Data Sets: As described in Section 3.3, three annotated data sets are used for evaluation: (1) *Abstract*, which consists of annotations for abstracts only; (2) *Full-text*, which consists of annotations for the full texts of 50 articles whole abstracts were annotated in the *Abstract* set. To avoid overlapping between two types of annotation, we exclude the abstract part from those full texts. Since each full text is much longer than the corresponding abstract, we believe this will not make a great difference. (3) *ALL*, which includes the *Abstract*, *Full text* as well as the annotations for the sentences selected through the active learning process described in Section 3.4.

The annotated data sets are used to train the model and quantitatively evaluate the performance of the proposed entity extraction model. We use 5-fold cross validation and report the average performance. Each data set is split into five equal-sized subsets . Every time four out of five subsets are used for training and the remaining one is used for testing. This process is repeated five times to ensure every subset is used as the testing subset once. We split an annotated data set into training and testing subsets in a sentence-by-sentence manner, which means sentences in a single paper might both appear in training and validation subsets.

Table 3: Statistics of the Data Sets

	Abstract	*Full-text*	*ALL*
Num. of tokens	43,474	270,984	341,306
Num. of sentences	1,584	9,261	11,801
Num. of annotations	4,515	19,292	25,622

Table 4: Performance of Entity Extraction (Micro Avg)

Data	P_{strict}	R_{strict}	$F1_{strict}$	P_{soft}	R_{soft}	$F1_{soft}$
Abstract	0.866	0.877	0.872	0.915	0.922	0.918
Full-text	0.880	0.898	0.889	0.912	0.927	0.920
ALL	0.878	0.897	0.887	0.913	0.930	0.921

Table 3 summarizes the statistic of three annotated data sets. And it is clear that the number of annotations in the *Full-text* data set is much more than those for the *Abstract* data set. And the proposed active learning algorithm is a cost-effective way to prioritize the annotation effort.

5.2 Results

We conduct four sets of experiments over the above annotated data sets. The first set evaluates the overall performance of the proposed extraction framework as shown in Figure 1. The second and third set of experiments focus on checking the effectiveness of domain adaptation and active learning respectively. The final set of experiments helps us better understand the benefit of annotating full text of articles.

5.2.1 Extraction Accuracy

We first evaluate the extraction performance of the proposed system on the three data sets. Both matching criteria, i.e., soft and strict, are used to compute the evaluation measures. The reported performance is the micro average over all the six categories. The results are shown in Table 4. Overall, our proposed extraction model can achieve around 90% accuracy over all the three data sets.

We now take a close look at the largest annotated data set (i.e., *ALL*), and look into the extraction performance for each category. The results are summarized in Table 5. The $F1_{soft}$ metric for all six types of entities surpasses 85% and the average $F1_{soft}$ reaches 92%. It is clear that our model can accurately identify the key information in the text. Moreover, the evaluation results using the strict matching criterion are also reported. In particular, $F1_{strict}$ is around 89%, which is at a similar accuracy level as previous work on extracting material-related information.[19]

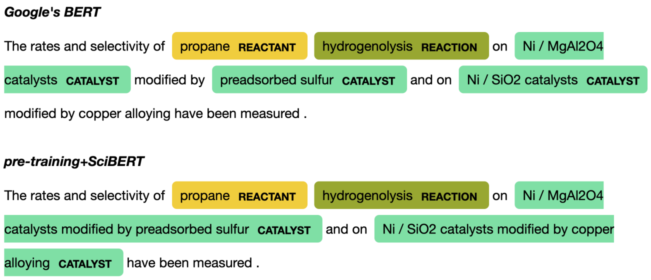
Table 5: Performance of Entity Extraction (Data set: *ALL*)

	P_{strict}	R_{strict}	$F1_{strict}$	P_{soft}	R_{soft}	$F1_{soft}$
Catalyst	0.858	0.868	0.863	0.913	0.914	0.914
Reactant	0.905	0.905	0.905	0.938	0.939	0.939
Product	0.886	0.912	0.899	0.919	0.940	0.929
Reaction	0.915	0.928	0.922	0.932	0.950	0.941
Characterization	0.808	0.851	0.828	0.849	0.894	0.870
Treatment	0.805	0.876	0.839	0.827	0.903	0.863
micro avg	0.878	0.897	0.887	0.913	0.930	0.921

Table 6: Effectiveness of Domain Adaptive Pre-training

Data	BERT	$F1_{soft}$	$F1_{strict}$
Abstract	Google's BERT	0.886	0.825
	SciBERT	0.911	0.857
	pre-training+SciBERT	**0.918**	**0.872**
Abstract+Full-text	Google's BERT	0.903	0.863
	SciBERT	0.915	0.881
	pre-training+SciBERT	**0.919**	**0.885**
ALL	Google's BERT	0.904	0.866
	SciBERT	0.917	0.882
	pre-training+SciBERT	**0.921**	**0.887**

5.2.2 Domain Adaptation



Google's BERT

The rates and selectivity of propane REACTANT hydrogenolysis REACTION on Ni / MgAl2O4 catalysts CATALYST modified by preadsorbed sulfur CATALYST and on Ni / SiO2 catalysts CATALYST modified by copper alloying have been measured .

pre-training+SciBERT

The rates and selectivity of propane REACTANT hydrogenolysis REACTION on Ni / MgAl2O4 catalysts modified by preadsorbed sulfur CATALYST and on Ni / SiO2 catalysts modified by copper alloying CATALYST have been measured .

Figure 5: Example of Domain Adaption produces better Model

We also analyze the contribution of domain adaptation techniques in our model. This time the cross validation process covers three different combinations of annotated data: *Abstract*

Table 7: Effectiveness of Active Learning

	Data	$F1_{soft}$	$F1_{strict}$
Treatment	*Abstract*	0.729	0.676
	Full-text	0.727	0.702
	ALL	**0.863**	**0.839**
micro avg	*Abstract*	0.918	0.872
	Full-text	0.920	0.889
	ALL	**0.921**	**0.887**

, *Abstract* + *Full-text*, and *ALL*. Unlike previous experiments which only use task pre-trained SciBERT, here we compare the overall performance of three different pre-trained models on the three data sets.

Table 6 shows domain adaptation helps to improve the performance. Using Google's BERT can not achieve optimal performance. In fact, as 5 shows, domain adaption helps the model to avoid breaking a long entity into several pieces. It is clear that switching the pre-training corpus from general English articles to scientific articles makes a difference. But as most of the scientific articles are in the biomedical domain, pre-training SciBERT on chemical corpus leads to a steady performance gain. Moreover, it seems that, for smaller data sets such as the *Abstract* set, the performance difference is much larger, meaning that more data can partially alleviate the domain shift problem.

5.2.3 Active Learning

Based on the results shown in Table 5, the extraction results in the *Treatment* and *Characterization* categories are worse than those in the other four categories. The performance difference could be caused by the fewer annotations for these two categories, as shown in Table 2. In fact, the worse performance of these two categories on the *Abstract* and *Full-text* data set motivated the proposed active learning idea as described in Section 3.4.

We conduct experiments to examine the impact of new annotations obtained through active learning. In particular, the *ALL* data set include the new annotations while the *Abstract* and *Full-text* do not. Table 7 compares the average performance over all the six entity categories as well as the performance for *Treatment* category. We can see that active learning is effective in bringing more annotations for the *Treatment* category and improving the extraction accuracy accordingly. In the meantime, the average performance over all the categories also increases. If we used the standard way and collect more treatment annotation during full text annotation, we would need to five times, more even ten times more annotation effort according to the distribution statistics. However, with active learning, we only label 956 sentences which saves us months of annotation time based on our speed of finishing full text annotation. This demonstrates the value of active learning as a flexible and effective way to improve the extraction accuracy for the categories with fewer annotations.

5.2.4 Benefit of full-text annotation

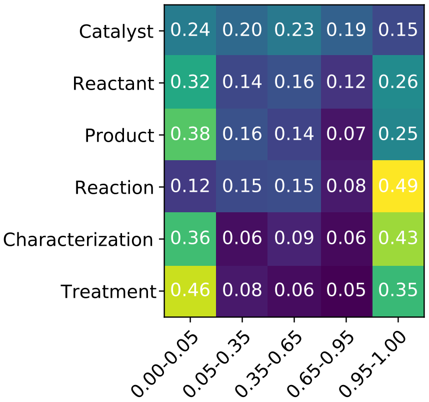


Figure 6: Full-text entity coverage ratio

We choose to annotate full text of an article since it contains more information than the corresponding abstract. However, the downside of full-text annotation lies in the increasing annotation effort. As shown in Table 3, the number of annotations in the *Full-text* set is around 5 times of those in the *Abstract* set. Please note that the full text of an article does not have overlap with its abstract as explained in the experiment setup section.

To better understand the benefit of full-text annotation, we conduct experiments to measure how much additional information we can get from full-text annotation. In particular, we collect statistics based on the similarity between the information extracted from each article's abstract and its full text. For every unique entity extracted in the full text, we calculate the maximal similarity between the entity and all entities extracted in the abstract, estimating to how much degree the entity is covered by the abstract. The similarity metric we choose is the cosine similarity between two entities' character N-gram representations. Based on the per-class maximal similarity distribution shown in Figure 6, almost half of the unique entities in full-text are not well covered by abstract. Moreover, among the six entity categories, the similarity distribution in the *Characterization* and *Treatment* categories shift heavily towards the two ends. These two entity categories are indeed often ignored or not well documented in the abstract. As for the *Reactant*, *Product* and *Reaction* categories, their similarities are more evenly distributed. This makes sense since the terminology often has a common suffix or prefix, which will lead to a non-zero similarity score between two character N-gram representations. The *Catalyst* category has the most uniform distribution since catalysts are usually combinations of a small pool of metal and inorganic elements.

6 Extraction Results and Their Applications

The previous section has demonstrated the effectiveness of the proposed extraction models. We now apply the trained model to the highly relevant documents collection (described in Section 3.2) to extract the entities in the six categories: *Catalyst*, *Reactant*, *Product*, *Reaction*, *Characterization* and *Treatment*. The collection consists of 3578 relevant documents and 440K sentences.

Table 8: Number of extracted entities

	Total	Unique	Unique after normalization
Catalyst	205K	54781	37961
Characterization	190K	4376	3558
Product	210K	25201	20199
Reactant	146K	21428	17565
Reaction	45K	6074	5264
Treatment	41K	2169	1895

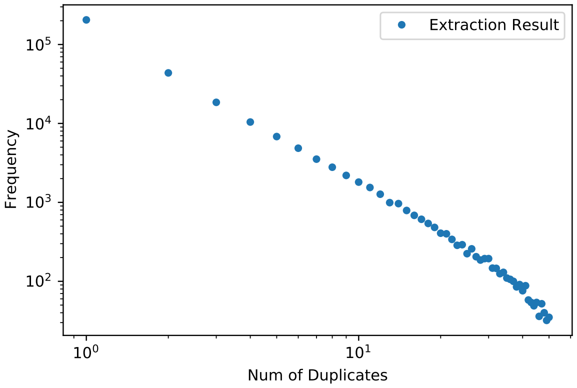


Figure 7: Duplicates Distribution

Table 8 summarizes the number of extracted entities for each entity category. As described in Section 4.6, it is important to normalize the extracted entities due to the naming variants. The table also reports the number of uniquely extracted entities and the number of unique entities after applying the proposed normalization method. It is clear that the normalization method can reduce the number of extracted entities significantly for all the six entity categories. In particular, the *Catalyst* category benefit the most from the normalization. As *Catalyst* is the most complicated type of information containing countless combinations of metal and support material, our normalization pipeline not only group alternative expressions together but also break combinations and reduce them to finer granularity representations. This may explain why the normalization is more effective for the *Catalyst* category.

We further analyze the extracted entities and to better understand the properties of the duplicated entities. In particular, we count how many times each span is duplicated in an article and the rank of the span in terms of the frequency. Figure 7 shows the relationship on a log-log graph. It is interesting to see that the trend of the extracted catalysis-related entities is consistent with what we observed for words in natural language. This pattern is well known as the Zipf's law.[52]

There are many possible ways of utilizing the extracted entities for information access and knowledge discovery. We explore two applications as case studies in this work: (1) an entity-aware search engine; and (2) an entity correlation analysis system. Both applications are accessible through Web interfaces.

6.1 Entity-aware Search Engine

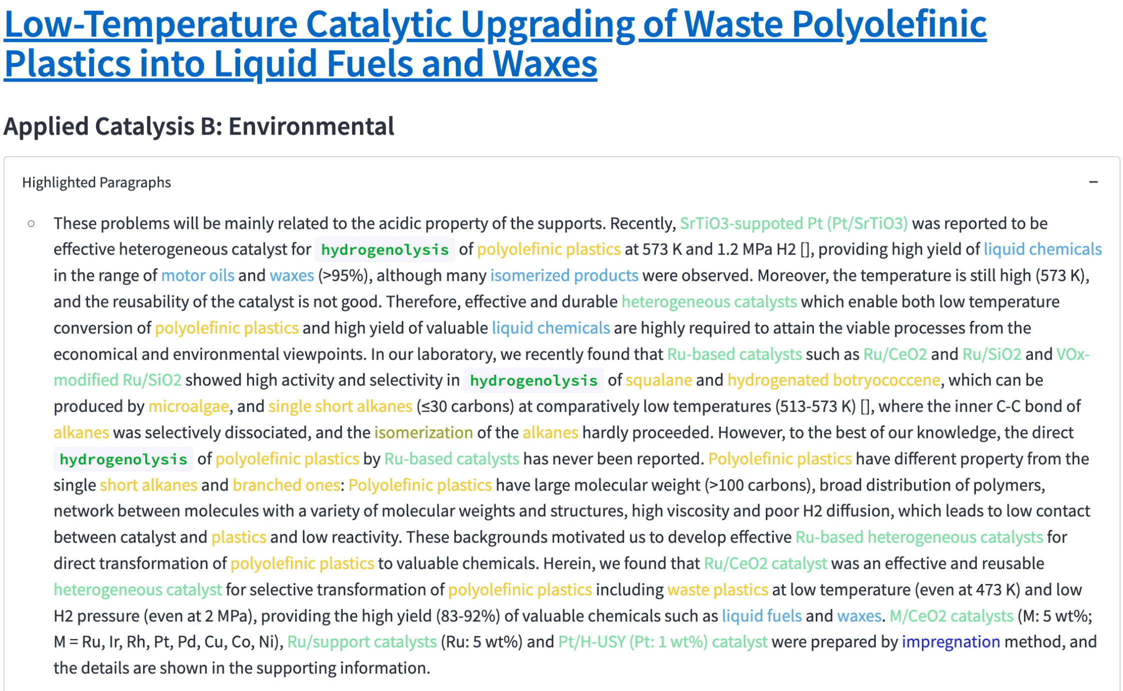
With the automatically extracted entities, we develop an entity-aware search engine customized for catalysis-related information. Similar to existing search engines, our system takes keywords as input and returns a ranked list of relevant articles. The results are ranked using a state-of-the-art ranking algorithm. The key difference between our system and existing search engines is that we leverage the extracted entities in the following unique ways to help users more efficiently navigate through the search results.

Entity-driven result representation: Existing literature search engines, such as Google Scholar and Science Direct, heavily rely on the title and abstract to generate snippet and explain the search result. This explanation, in a lot of time, is not enough and users are forced to go through the full text without any further help from the search engine. On the contrary, we break the full text of each article into a list of paragraphs, each with ten sentences, and pick three most relevant paragraphs to represent the article. Moreover, all the extracted entities in these paragraphs are also highlighted with different colors based on the corresponding categories to provide better visual guidance. An example result is shown in Figure 8.

Entity-based result filtering: Existing search engines provide limited support on result filtering. If a user is not satisfied with the search results, they may have to keep reformulating the queries. This is especially problematic in the emerging research fields where researchers do not have prior experience or references to judge whether their input criteria contain enough information for a given material of interest. On the contrary, our developed search engine displays a sidebar with options of filtering results based on the matching of different types of entities. In addition to the ability of allowing users to reformulate their queries in the search box, the sidebar provides automatically generated filtering criteria. Figure 9, shows the sidebar with filtering functions under the six categories with each containing the 20 most frequent entities extracted from the searching results. We note that both the list of literature and the checkbox of the entities are automatically generated without human intervention.

We now use an example query "plastic hydrogenolysis" to further illustrate the benefit of our developed entity-aware search engine. With the specific query, Google Scholar returns a three-line context under each item, containing the matched query words in highlights. However, no further filtering or extraction of the 6100 searching results can be performed to narrow down the searching inquiry. In contrast, our system offers secondary filtering of the initial searching results for researchers who may be interested in LDPE (Reactant) hydrogenolysis (Reaction) on Ruthenium (Catalyst) to produce fuels (Products). Upon checking the corresponding checkboxes (Figure 9) that match the interests in the above example, the most relevant literature[9] is shown in Figure 8. Our system automatically identifies the top three relevant paragraphs that contain the densest extracted information. Queries are colored differently as a visual guide based on the types. The extraction and labeling help the researchers quickly access key information delivered by the literature and identify components that are relevant in each category. In this particular case, one may promptly summarize from our automatic extraction that this literature[9] highlights the performance of Ru/CeO_2 catalyst for the hydrogenolysis of polyolefin plastics to produce liquid fuel and waxes selectively, and the authors compared their catalytic performance to previously established processes such as pyrolysis and metathesis and catalysts such as $Pt/SrTiO_3$ and Pt/H-USY.

Furthermore, the filtering function gives re-



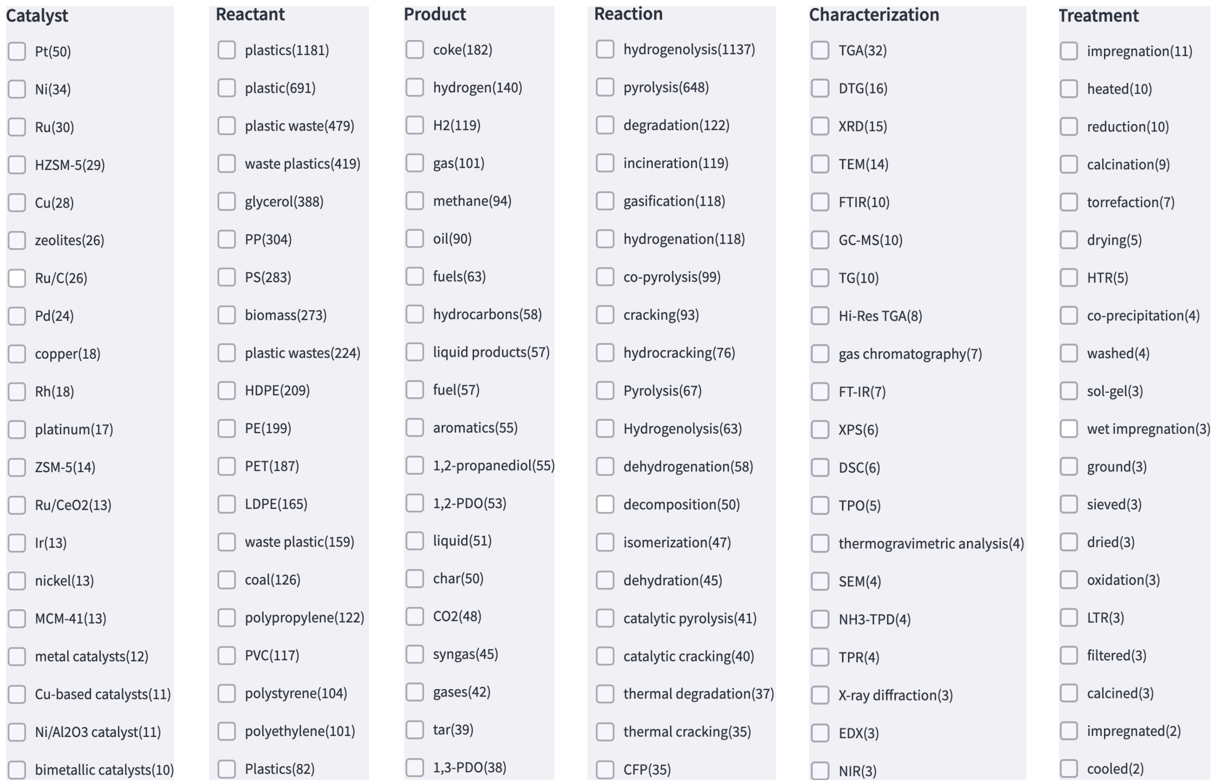
Low-Temperature Catalytic Upgrading of Waste Polyolefinic Plastics into Liquid Fuels and Waxes

Applied Catalysis B: Environmental

Highlighted Paragraphs −

○ These problems will be mainly related to the acidic property of the supports. Recently, SrTiO3-suppoted Pt (Pt/SrTiO3) was reported to be effective heterogeneous catalyst for `hydrogenolysis` of polyolefinic plastics at 573 K and 1.2 MPa H2 [], providing high yield of liquid chemicals in the range of motor oils and waxes (>95%), although many isomerized products were observed. Moreover, the temperature is still high (573 K), and the reusability of the catalyst is not good. Therefore, effective and durable heterogeneous catalysts which enable both low temperature conversion of polyolefinic plastics and high yield of valuable liquid chemicals are highly required to attain the viable processes from the economical and environmental viewpoints. In our laboratory, we recently found that Ru-based catalysts such as Ru/CeO2 and Ru/SiO2 and VOx-modified Ru/SiO2 showed high activity and selectivity in `hydrogenolysis` of squalane and hydrogenated botryococcene, which can be produced by microalgae, and single short alkanes (≤30 carbons) at comparatively low temperatures (513-573 K) [], where the inner C-C bond of alkanes was selectively dissociated, and the isomerization of the alkanes hardly proceeded. However, to the best of our knowledge, the direct `hydrogenolysis` of polyolefinic plastics by Ru-based catalysts has never been reported. Polyolefinic plastics have different property from the single short alkanes and branched ones: Polyolefinic plastics have large molecular weight (>100 carbons), broad distribution of polymers, network between molecules with a variety of molecular weights and structures, high viscosity and poor H2 diffusion, which leads to low contact between catalyst and plastics and low reactivity. These backgrounds motivated us to develop effective Ru-based heterogeneous catalysts for direct transformation of polyolefinic plastics to valuable chemicals. Herein, we found that Ru/CeO2 catalyst was an effective and reusable heterogeneous catalyst for selective transformation of polyolefinic plastics including waste plastics at low temperature (even at 473 K) and low H2 pressure (even at 2 MPa), providing the high yield (83-92%) of valuable chemicals such as liquid fuels and waxes. M/CeO2 catalysts (M: 5 wt%; M = Ru, Ir, Rh, Pt, Pd, Cu, Co, Ni), Ru/support catalysts (Ru: 5 wt%) and Pt/H-USY (Pt: 1 wt%) catalyst were prepared by impregnation method, and the details are shown in the supporting information.

Figure 8: A demonstration of automatically-generated representation of an article under the input criterion of "plastic hydrogenolysis" using the top three most relevant paragraphs recognized by the catalysis search engine.



Catalyst

□ Pt(50)
□ Ni(34)
□ Ru(30)
□ HZSM-5(29)
□ Cu(28)
□ zeolites(26)
□ Ru/C(26)
□ Pd(24)
□ copper(18)
□ Rh(18)
□ platinum(17)
□ ZSM-5(14)
□ Ru/CeO2(13)
□ Ir(13)
□ nickel(13)
□ MCM-41(13)
□ metal catalysts(12)
□ Cu-based catalysts(11)
□ Ni/Al2O3 catalyst(11)
□ bimetallic catalysts(10)

Reactant

□ plastics(1181)
□ plastic(691)
□ plastic waste(479)
□ waste plastics(419)
□ glycerol(388)
□ PP(304)
□ PS(283)
□ biomass(273)
□ plastic wastes(224)
□ HDPE(209)
□ PE(199)
□ PET(187)
□ LDPE(165)
□ waste plastic(159)
□ coal(126)
□ polypropylene(122)
□ PVC(117)
□ polystyrene(104)
□ polyethylene(101)
□ Plastics(82)

Product

□ coke(182)
□ hydrogen(140)
□ H2(119)
□ gas(101)
□ methane(94)
□ oil(90)
□ fuels(63)
□ hydrocarbons(58)
□ liquid products(57)
□ fuel(57)
□ aromatics(55)
□ 1,2-propanediol(55)
□ 1,2-PDO(53)
□ liquid(51)
□ char(50)
□ CO2(48)
□ syngas(45)
□ gases(42)
□ tar(39)
□ 1,3-PDO(38)

Reaction

□ hydrogenolysis(1137)
□ pyrolysis(648)
□ degradation(122)
□ incineration(119)
□ gasification(118)
□ hydrogenation(118)
□ co-pyrolysis(99)
□ cracking(93)
□ hydrocracking(76)
□ Pyrolysis(67)
□ Hydrogenolysis(63)
□ dehydrogenation(58)
□ decomposition(50)
□ isomerization(47)
□ dehydration(45)
□ catalytic pyrolysis(41)
□ catalytic cracking(40)
□ thermal degradation(37)
□ thermal cracking(35)
□ CFP(35)

Characterization

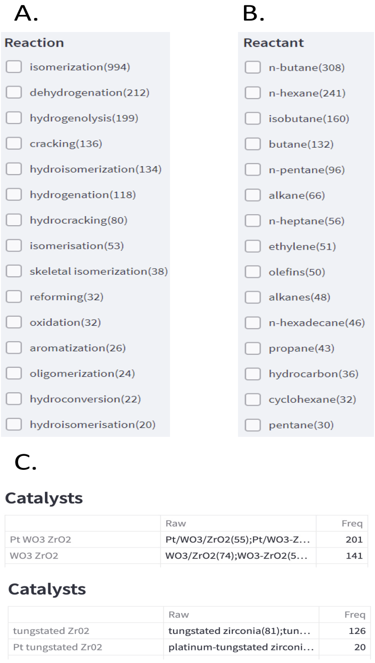
□ TGA(32)
□ DTG(16)
□ XRD(15)
□ TEM(14)
□ FTIR(10)
□ GC-MS(10)
□ TG(10)
□ Hi-Res TGA(8)
□ gas chromatography(7)
□ FT-IR(7)
□ XPS(6)
□ DSC(6)
□ TPO(5)
□ thermogravimetric analysis(4)
□ SEM(4)
□ NH3-TPD(4)
□ TPR(4)
□ X-ray diffraction(3)
□ EDX(3)
□ NIR(3)

Treatment

□ impregnation(11)
□ heated(10)
□ reduction(10)
□ calcination(9)
□ torrefaction(7)
□ drying(5)
□ HTR(5)
□ co-precipitation(4)
□ washed(4)
□ sol-gel(3)
□ wet impregnation(3)
□ ground(3)
□ sieved(3)
□ dried(3)
□ oxidation(3)
□ LTR(3)
□ filtered(3)
□ calcined(3)
□ impregnated(2)
□ cooled(2)

Figure 9: An illustration of sidebar with filter checkbox under the input criterion of "plastic hydrogenolysis" in the catalysis search engine.

searchers a comprehensive overview of essential considerations in each category, which helps them position their interests to prior knowledge in similar catalytic systems. Concerning the choices of catalysts, besides ruthenium, our searching engine identifies additional active catalysts (Figure 9), such as platinum, nickel, and solid acids (H-ZSM-5, zeolites). These results are not apparent even for experienced catalysis experts. Excitingly, the catalyst screening has provided useful insights to some of us who are specialized in catalysis. The searching results also matched well with some recently released literature,[6,53–55] some published after our database was generated, demonstrating the capabilities of predicting effective catalysts in plastic deconstruction. For example, a candidate catalyst named platinum tungstated-zirconia ($Pt/WO_3{-}ZrO_2$) appeared frequently in relevant searching results. The navigation sidebar from the query of ”platinum tungstated zirconia” suggests that this catalyst is highly relevant to the reactions of interest, such as (hydro)isomerization, hydrogenolysis, and (hydro)cracking (Figure 10.a). The correlation analysis tool identified equivalent keywords (e.g., $Pt/WO_{3/}ZrO_2$) over 500 times in the database (Figure 10.c), although this catalyst had never been reported previously for plastic deconstruction, according to the checkboxes under the entity of Reactant (Figure 10.b). Bridging the gap leads to lead to successful implement of $Pt/WO_3{-}ZrO_2$ to produce high-quality, branched fuels from plastic wastes, first reported from our research center following the catalyst screening by the searching engine. Finally, catalyst treatment is an essential factor that does not appear to be as critical as the other five entity types for people outside the catalysis field. It is, however, well known in the field that catalyst treatments can sometimes lead to a drastic change in catalytic behaviors and performances. It is noticeable that the searching engine automatically captured the so-called LTR (low-temperature reduction) and HTR (high-temperature reduction) treatments – they are thought to be able to cause noticeable differences in catalytic performances, but they are significant only with Ru catalysts. In fact, it is known in the field that HTR treatments on the Ru catalysts increase the hydrogen adsorption equilibrium, influencing the hydrogenolysis product distribution.[56] Automatic identification of such correlations is not expected because both terms are in abbreviation and are known only for researchers specializing in Ru-based catalysis.



Reaction

☐ isomerization(994)
☐ dehydrogenation(212)
☐ hydrogenolysis(199)
☐ cracking(136)
☐ hydroisomerization(134)
☐ hydrogenation(118)
☐ hydrocracking(80)
☐ isomerisation(53)
☐ skeletal isomerization(38)
☐ reforming(32)
☐ oxidation(32)
☐ aromatization(26)
☐ oligomerization(24)
☐ hydroconversion(22)
☐ hydroisomerisation(20)

Reactant

☐ n-butane(308)
☐ n-hexane(241)
☐ isobutane(160)
☐ butane(132)
☐ n-pentane(96)
☐ alkane(66)
☐ n-heptane(56)
☐ ethylene(51)
☐ olefins(50)
☐ alkanes(48)
☐ n-hexadecane(46)
☐ propane(43)
☐ hydrocarbon(36)
☐ cyclohexane(32)
☐ pentane(30)

Catalysts

	Raw	Freq
Pt WO3 ZrO2	Pt/WO3/ZrO2(55);Pt/WO3-Z...	201
WO3 ZrO2	WO3/ZrO2(74);WO3-ZrO2(5...	141

Catalysts

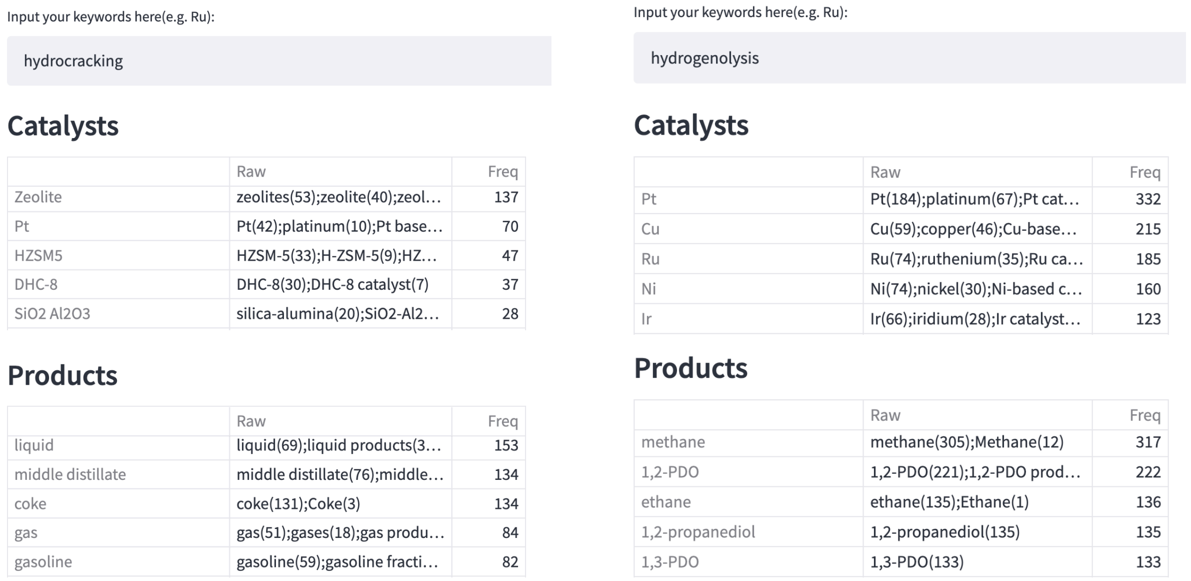
	Raw	Freq
tungstated Zr02	tungstated zirconia(81);tun...	126
Pt tungstated Zr02	platinum-tungstated zirconi...	20

Figure 10: Example of searching $Pt/WO_{3/}ZrO_2$ in search engine and searching ”tungstated” and ”ZrO2” in correlation analysis tool

6.2 Entity Correlation Analysis System

To further demonstrate our searching engine is capable of detecting advanced correlations among entities for catalysis research, we also developed a visual platform, Entity Correlation Analysis, based on automatic extraction of sentence-level co-occurrence signals. After a user enters an entity as a query, the system will automatically search for the co-occurred extracted entities and rank them based on the frequency of the co-occurrences for each of the six categories.

Figure 11 shows the comparison of correlation results for two queries: “hydrocracking” and “hydrogenolysis”. Both queries mean bond-



Input your keywords here(e.g. Ru):

hydrocracking

Catalysts

	Raw	Freq
Zeolite	zeolites(53);zeolite(40);zeol…	137
Pt	Pt(42);platinum(10);Pt base…	70
HZSM5	HZSM-5(33);H-ZSM-5(9);HZ…	47
DHC-8	DHC-8(30);DHC-8 catalyst(7)	37
SiO2 Al2O3	silica-alumina(20);SiO2-Al2…	28

Products

	Raw	Freq
liquid	liquid(69);liquid products(3…	153
middle distillate	middle distillate(76);middle…	134
coke	coke(131);Coke(3)	134
gas	gas(51);gases(18);gas produ…	84
gasoline	gasoline(59);gasoline fracti…	82

Input your keywords here(e.g. Ru):

hydrogenolysis

Catalysts

	Raw	Freq
Pt	Pt(184);platinum(67);Pt cat…	332
Cu	Cu(59);copper(46);Cu-base…	215
Ru	Ru(74);ruthenium(35);Ru ca…	185
Ni	Ni(74);nickel(30);Ni-based c…	160
Ir	Ir(66);iridium(28);Ir catalyst…	123

Products

	Raw	Freq
methane	methane(305);Methane(12)	317
1,2-PDO	1,2-PDO(221);1,2-PDO prod…	222
ethane	ethane(135);Ethane(1)	136
1,2-propanediol	1,2-propanediol(135)	135
1,3-PDO	1,3-PDO(133)	133

Figure 11: A comparison of searching results in the Catalysis Correlation Analysis platform with input criteria "hydrocracking" and "hydrogenolysis" in the entity type "reaction".

breaking via hydrogen treatments under the context of plastic decomposition, and they appear to be similar, especially for researchers outside or just entering the field. The subtle differences are underneath the bond scission mechanisms – hydrogenolysis of hydrocarbons occurs via dehydrogenation followed by C-C bond scission and hydrogenation, while hydrocracking starts with protonation following skeletal rearrangement and beta-scission. The formal process requires dissociation of hydrogen and alkyl intermediates and usually involves metals as catalysts, while the latter approach requires solid acids to protonate and stabilize the hydrocarbon and often requires a hydrogenation metal to promote beta-scission. Excitingly, our correlation analytical tool reports results that match these advanced criteria (Figure 11). That is, it correlates solid acids (zeolites, H-ZSM-5, and SiO_2-Al_2O_3), hydrogenation metal (Pt), and even commercial hydrocracking catalysts (DHC-8) with the hydrocracking process and identifies hydrogenolysis metals (Ru, Ir, Ni, Pt, and Cu), typical for hydrogenolysis process. The automatically generated products entities with high-level correlations are also consistent with what one would expect from these two reactions. Namely, the hydrocracking route produces high-quality fuels (liquid, middle distillate, gasoline) due to branching via hydroisomerization but also suffer from coke formation due to acid-catalyzed reactions; Hydrogenolysis is known for faster rate and the production of plastic diol precursors (1,2-PDO, 1,3-PDO, 1,2-pentanediol) but is also notorious for heavy methane/ethane formations. Our correlation analysis tool captures these critical characteristics from two concepts with subtle differences with no human intervention, demonstrating our software's robustness.

We note that the current version of our Catalysis Correlation Analysis system also poses a limitation, which is subject to future improvements. The current algorithm cannot fully recognize and group similar expression variants of the same input. This is a noticeable issue because researchers regularly use non-standardized expressions, variants, and abbreviations to refer to the same object. In the example of ruthenium as a catalyst, the searching input of Ru, which is the chemical symbol of ruthenium, results in 1037 direct matches in the type of catalysts. It identifies thousands of correlation information to other types of entities. However, the system found no matching result with "ruthenium" as the searching input. Similarly, the "Ru C" query, which refers to carbon-supported ruthenium catalysts, results in 487 pieces of matching information and hundreds of correlations. However, other commonly used variants that express the same, such as "Ru/C", "Ru on C", and "Ru on carbon", lead to little or no result found in the system.

7 Conclusions

Catalytic polymer deconstruction is one of the primary ways for plastic waste upcycling. However, prior literature on this topic is minimal, and scientists must rely on many partially relevant articles to shape their hypotheses and experimental designs. This is often a time-consuming and labor-intensive process. To overcome this limitation, our work aims to develop effective algorithms that allow scientists to efficiently access useful information from a large volume of scientific literature.

In this work, we focus on tackling the problem of extracting six types of entities related to catalysis from scientific literature using machine learning methods. To construct the annotation data set used to train the extraction model, we first developed a binary classifier to identify highly relevant documents and an active learning strategy to reduce the annotation effort. Furthermore, we develop an *entity extraction* component that leverages domain-adaptive pre-training and span classification to automatically extract a text span and label it with the corresponding categories. The extraction model has then been evaluated and it can achieve around 90% extraction accuracy for all six entity types. The extracted information is then used to build an entity-aware literature search engine and a catalyst correlation analysis system, which aid our collaborators in identifying several catalysts that can efficiently deconstruct polyolefins.

As the focus of this work is to develop effective extraction models, we crawled the articles from only one source. For future work, we plan to enrich our data collections with articles from more publishers such as American Chemical Society, Springer Nature, and Royal Chemical Society and also enable regular crawling from all the publishers to make sure that the data collection contains up-to-date articles. This is a critical function yet to be incorporated into our system, considering that the field of catalytic plastic upcycling is new, and the volume of literature proliferates rapidly. Moreover, we will develop more advanced entity normalization methods to further consolidate the extracted information and construct the knowledge base accordingly. Finally, we also plan to expand our extraction effort to extract other useful information related to catalytic processes, e.g. the parameters of catalyst synthesis, selectivity, yield, etc. The extraction of synthesis and reaction parameters(e.g. temperature, time) usually requires an extraction model that is different from that for entity extraction.

Acknowledgement This work was supported as part of the Center for Plastics Innovation, an Energy Frontier Research Center, funded by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences (BES), under Award Number DE-SC0021166.

8 Data and Software Availability

The code of our model and application are publicly available on GitHub.[57] We release all of our annotated data, including the abstract, full-text, and active learning part. Accompanied with the data, we provide the model checkpoints that are trained using all the data and have the best performance. Both the data and checkpoint are packed into a zenodo dataset[58] Instructions and code examples are also provided to reproduce our evaluation result in Table 5 and how to train/test our model using user-provided data. The information will be regularly updated to reflect our most recent effort, and more documentations will be added to help other researchers understand the usage of the data and codes. For people who are interested in our two system,s we also provide their latest URLs on GitHub and two short introduction to guide people run our system on their own data.

References

(1) Giacovelli, C., et al. Single-Use Plastics: A Roadmap for Sustainability (rev. 2). **2018**,

(2) Geyer, R.; Jambeck, J. R.; Law, K. L. Production, Use, and Fate of All Plastics Ever Made. *Science advances* **2017**, *3*, e1700782.

(3) Kots, P. A.; Vance, B. C.; Vlachos, D. G. Polyolefin Plastic Waste Hydroconversion to Fuels, Lubricants, and Waxes: A Comparative Study. *Reaction Chemistry & Engineering* **2022**, *7*, 41–54.

(4) Wang, C.; Xie, T.; Kots, P. A.; Vance, B. C.; Yu, K.; Kumar, P.; Fu, J.; Liu, S.; Tsilomelekis, G.; Stach, E. A., et al. Polyethylene Hydrogenolysis at Mild Conditions over Ruthenium on Tungstated Zirconia. *JACS Au* **2021**, *1*, 1422–1434.

(5) Kots, P. A.; Liu, S.; Vance, B. C.; Wang, C.; Sheehan, J. D.; Vlachos, D. G. Polypropylene Plastic Waste Conversion to Lubricants over Ru/TiO2 Catalysts. *ACS Catalysis* **2021**, *11*, 8104–8115.

(6) Vance, B. C.; Kots, P. A.; Wang, C.; Hinton, Z. R.; Quinn, C. M.; Epps III, T. H.; Korley, L. T.; Vlachos, D. G. Single Pot Catalyst Strategy to Branched Products via Adhesive Isomerization and Hydrocracking of Polyethylene Over Platinum Tungstated Zirconia. *Applied Catalysis B: Environmental* **2021**, *299*, 120483.

(7) Liu, S.; Kots, P. A.; Vance, B. C.; Danielson, A.; Vlachos, D. G. Plastic Waste to Fuels by Hydrocracking at Mild Conditions. *Science Advances* **2021**, *7*, eabf8283.

(8) Rorrer, J. E.; Beckham, G. T.; Román-Leshkov, Y. Conversion of Polyolefin Waste to Liquid Alkanes With Ru-Based Catalysts Under Mild Conditions. *JACS Au* **2020**, *1*, 8–12.

(9) Nakaji, Y.; Tamura, M.; Miyaoka, S.; Kumagai, S.; Tanji, M.; Nakagawa, Y.; Yoshioka, T.; Tomishige, K. Low-Temperature Catalytic Upgrading of Waste Polyolefinic Plastics Into Liquid Fuels and Waxes. *Applied Catalysis B: Environmental* **2021**, *285*, 119805.

(10) Celik, G.; Kennedy, R. M.; Hackler, R. A.; Ferrandon, M.; Tennakoon, A.; Patnaik, S.; LaPointe, A. M.; Ammal, S. C.; Heyden, A.; Perras, F. A., et al. Upcycling Single-Use Polyethylene Into High-Quality Liquid Products. *ACS central science* **2019**, *5*, 1795–1803.

(11) Ellis, L. D.; Orski, S. V.; Kenlaw, G. A.; Norman, A. G.; Beers, K. L.; Román-Leshkov, Y.; Beckham, G. T. Tandem Heterogeneous Catalysis for Polyethylene Depolymerization via an Olefin-Intermediate Process. *ACS Sustainable Chemistry & Engineering* **2021**, *9*, 623–628.

(12) Serrano, D.; Aguado, J.; Escola, J. Developing Advanced Catalysts for the Conversion of Polyolefinic Waste Plastics Into Fuels and Chemicals. *ACS Catalysis* **2012**, *2*, 1924–1941.

(13) Zhang, F.; Zeng, M.; Yappert, R. D.; Sun, J.; Lee, Y.-H.; LaPointe, A. M.; Peters, B.; Abu-Omar, M. M.; Scott, S. L. Polyethylene Upcycling to Long-Chain Alkylaromatics by Tandem Hydrogenolysis/Aromatization. *Science* **2020**, *370*, 437–441.

(14) Tennakoon, A.; Wu, X.; Paterson, A. L.; Patnaik, S.; Pei, Y.; LaPointe, A. M.; Ammal, S. C.; Hackler, R. A.; Heyden, A.; Slowing, I. I., et al. Catalytic Upcycling of High-Density Polyethylene via a Processive Mechanism. *Nature Catalysis* **2020**, *3*, 893–901.

(15) Krallinger, M. et al. The CHEMDNER Corpus of Chemicals and Drugs and Its Annotation Principles. *Journal of Cheminformatics* **2015**, *7*, S2.

(16) He, J.; Nguyen, D. Q.; Akhondi, S.; Druckenbrodt, C.; Thorne, C.; Hoessel, R.; Afzal, Z.; Zhai, Z.; Fang, B.; Yoshikawa, H.; Albahem, A.; Cavedon, L.; Cohn, T.; Baldwin, T.; Verspoor, K. M. ChEMU 2020: Natural Language Processing Methods Are Effective for Information Extraction From Chemical Patents. *Frontiers in Research Metrics and Analytics* **2021**, *6*.

(17) Olivetti, E.; Cole, J.; Kim, E.; Kononova, O.; Ceder, G.; Han, T. Y.; Hiszpanski, A. Data-Driven Materials Research Enabled by Natural Language Processing and Information Extraction. *Applied physics reviews* **2020**, *7*, 041317.

(18) Kononova, O. V.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and Challenges of Text Mining in Aterials Research. *iScience* **2021**, *24*.

(19) Weston, L.; Tshitoyan, V.; Dagdelen, J.; Kononova, O.; Trewartha, A.; Persson, K. A.; Ceder, G.; Jain, A. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction From the Materials Science Literature. *Journal of chemical information and modeling* **2019**, *59*, 3692–3702.

(20) Kim, E.; Huang, K.; Tomala, A. C.; Matthews, S.; Strubell, E.; Saunders, A.; McCallum, A.; Olivetti, E. Machine-Learned and Codified Synthesis Parameters of Oxide Materials. *Scientific Data* **2017**, *4*.

(21) Mysore, S.; Jensen, Z.; Kim, E. J.; Huang, K.; Chang, H.-S.; Strubell, E.; Flanigan, J.; McCallum, A.; Olivetti, E. The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. LAW@ACL. 2019.

(22) Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. Text-Mined Dataset of Inorganic Materials Synthesis Recipes. *Scientific Data* **2019**, *6*.

(23) Friedrich, A.; Adel, H.; Tomazic, F.; Hingerl, J.; Benteau, R.; Maruscyk, A.; Lange, L. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. ACL. 2020.

(24) Mysore, S.; Kim, E. J.; Strubell, E.; Liu, A.; Chang, H.-S.; Kompella, S.; Huang, K.; McCallum, A.; Olivetti, E. Automatically Extracting Action Graphs from Materials Science Synthesis Procedures. *ArXiv* **2017**, *abs/1711.06872*.

(25) Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T. Inferring Experimental Procedures From Text-Based Representations of Chemical Reactions. *Nature Communications* **2021**, *12*.

(26) Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. Correction to Automated Chemical Reaction Extraction from Scientific Literature. *Journal of Chemical Information and Modeling* **2021**, *61*, 4124.

(27) `https://www.elsevier.com/solutions/sciencedirect/librarian-resource-center/api`.

(28) `https://github.com/CederGroupHub/LimeSoup`.

(29) Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020.

(30) Zhang, Y.; Zhang, Y.; Qi, P.; Manning, C. D.; Langlotz, C. P. Biomedical and Clinical English Model Packages for the Stanza Python NLP Library. *Journal*

of the American Medical Informatics Association **2021**,

(31) Swain, M. C.; Cole, J. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of chemical information and modeling* **2016**, *56 10*, 1894–1904.

(32) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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). 2019; pp 4171–4186.

(33) Zheng, J.; Howsmon, D. P.; Zhang, B.; Hahn, J.; McGuinness, D. L.; Hendler, J. A.; Ji, H. Entity Linking for Biomedical Literature. *BMC Medical Informatics and Decision Making* **2015**, *15*, S4.

(34) Tchoua, R. B.; Ajith, A.; Hong, Z.; Ward, L. T.; Chard, K.; Audus, D. J.; Patel, S.; de Pablo, J. J.; Foster, I. T. Active Learning Yields Better Training Data for Scientific Named Entity Recognition. *2019 15th International Conference on eScience (eScience)* **2019**, 126–135.

(35) Li, Y.; lemao liu,; Shi, S. Empirical Analysis of Unlabeled Entity Problem in Named Entity Recognition. International Conference on Learning Representations. 2021.

(36) Fu, J.; Huang, X.; Liu, P. SpanNER: Named Entity Re-/Recognition as Span Prediction. ACL. 2021.

(37) Jessop, D. M.; Adams, S. E.; Willighagen, E.; Hawizy, L.; Murray-Rust, P. OSCAR4: A Flexible Architecture for Chemical Text-Mining. *Journal of Cheminformatics* **2011**, *3*, 41.

(38) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry. *Journal of Cheminformatics* **2011**, *3*, 17 – 17.

(39) Rocktäschel, T.; Weidlich, M.; Leser, U. ChemSpot: A Hybrid System for Chemical Named Entity Recognition. *Bioinformatics* **2012**, *28*, 1633–1640.

(40) Leaman, R.; Wei, C.-H.; Lu, Z. tmChem: A High Performance Approach for Chemical Named Entity Recognition and Normalization. *Journal of Cheminformatics* **2015**, *7*, S3.

(41) Khabsa, M.; Giles, C. L. Chemical Entity Extraction Using CRF and an Ensemble of Extractors. *Journal of Cheminformatics* **2015**, *7*, S12.

(42) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chemistry of Materials* **2017**, *29*, 9436–9444.

(43) Yadav, V.; Bethard, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. COLING. 2018.

(44) Beltagy, I.; Lo, K.; Cohan, A. SciBERT: Pretrained Language Model for Scientific Text. EMNLP. 2019.

(45) He, J.; Nguyen, D. Q.; Akhondi, S. A.; Druckenbrodt, C.; Thorne, C.; Hoessel, R.; Afzal, Z.; Zhai, Z.; Fang, B.; Yoshikawa, H.; Albahem, A.; Cavedon, L.; Cohn, T.; Baldwin, T.; Verspoor, K. M. Overview of ChEMU 2020: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. CLEF. 2020.

(46) He, J. et al. An Extended Overview of the CLEF 2020 ChEMU Lab: Information Extraction of Chemical Reactions from Patents. CLEF. 2020.

(47) Yang, H.; Hsu, W. H. Named Entity Recognition from Synthesis Procedural Text in Materials Science Domain with Attention-Based Approach. SDU@AAAI. 2021.

(48) Mahendran, D.; Gurdin, G.; Lewinski, N. A.; Tang, C.; McInnes, B. T. Identifying Chemical Reactions and Their Associated Attributes in Patents. *Frontiers in Research Metrics and Analytics* **2021**, *6*.

(49) Wang, J.; Ren, Y.; Zhang, Z.; Zhang, Y. Melaxtech: A report for CLEF 2020 - ChEMU Task of Chemical Reaction Extraction from Patent. CLEF. 2020.

(50) Lin, H.; Lu, Y.; Han, X.; Sun, L. A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land? EMNLP. 2020.

(51) Yu, J.; Bohnet, B.; Poesio, M. Named Entity Recognition as Dependency Parsing. ACL. 2020.

(52) `https://en.wikipedia.org/wiki/Zipf%27s_law`.

(53) Peczak, I. L.; Kennedy, R. M.; Hackler, R. A.; Wang, R.; Shin, Y.-S.; Delferro, M.; Poeppelmeier, K. R. Scalable Synthesis of Pt/SrTiO3 Hydrogenolysis Catalysts in Pursuit of Manufacturing-Relevant Waste Plastic Solutions. *ACS applied materials & interfaces* **2021**,

(54) Sun, M.; Zhu, L.; Liu, W.; Zhao, X.; Zhang, Y.; Luo, H.; Miao, G.; Li, S.; Yin, S.; Kong, L. Efficient Upgrading of Polyolefin Plastics Into C5–C12 Gasoline Alkanes Over a Pt/W/Beta Catalyst. *Sustainable Energy & Fuels* **2022**,

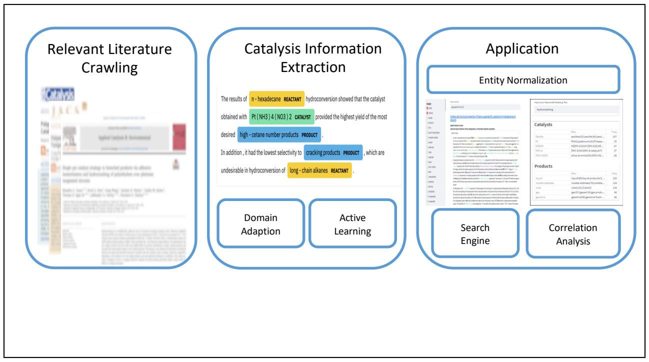
(55) Peczak, I. L.; Kennedy, R. M.; Hackler, R. A.; Wang, R.; Shin, Y.; Delferro, M.; Poeppelmeier, K. R. Scalable Synthesis of Pt/SrTiO3 Hydrogenolysis Catalysts in Pursuit of Manufacturing-Relevant Waste Plastic Solutions. *ACS Applied Materials & Interfaces* **2021**, *13*, 58691–58700.

(56) Bond, G. C.; Coq, B.; Dutartre, R.; Ruiz, J. G.; Hooper, A. D.; Proietti, M. G.; Sierra, M. C. S.; Slaa, J. C. Effect of Various Pretreatments on the Structure and Properties of Ruthenium Catalysts. *Journal of Catalysis* **1996**, *161*, 480–494.

(57) `https://github.com/nsndimt/CatalysisIE`.

(58) `https://doi.org/10.5281/zenodo.6533264`.

TOC Graphic



The results of n - hexadecane REACTANT hydroconversion showed that the catalyst obtained with Pt (NH3) 4 (NO3) 2 CATALYST provided the highest yield of the most desired high - cetane number products PRODUCT .

In addition , it had the lowest selectivity to cracking products PRODUCT , which are undesirable in hydroconversion of long - chain alkanes REACTANT .