Spring 5-27-2022

Text summarization towards scientific information extraction

Abigail Keller
DePaul University, abbykeller2@gmail.com

TEXT SUMMARIZATION TOWARDS SCIENTIFIC INFORMATION EXTRACTION

BY

ABIGAIL KELLER

A THESIS SUBMITTED TO THE SCHOOL OF COMPUTING, COLLEGE OF COMPUTING AND DIGITAL MEDIA OF DEPAUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN IN DATA SCIENCE

DEPAUL UNIVERSITY

CHICAGO, ILLINOIS

2022

DePaul University
College of Computing and Digital Media

MS Thesis Verification

This thesis has been read and approved by the thesis committee below according to the requirements of the School of Computing graduate program and DePaul University.

Name: Abigail Keller

Title of dissertation: Text Summarization towards Scientific Information Extraction

Date of Dissertation Defense: May 27th, 2022

Roselyne B. Tchoua

Advisor*

Daniela Raicu

1st Reader

Jacob Furst

2nd Reader

Peter Hastings

3rd Reader

4th Reader (if applicable)

5th Reader (if applicable)

** A copy of this form has been signed, but may only be viewed after submission and approval of FERPA request letter.*

Abstract

Despite the exponential growth in scientific textual content, research publications are still the primary means for disseminating vital discoveries to experts within their respective fields. These texts are predominantly written for human consumption resulting in two primary challenges; experts cannot efficiently remain well-informed to leverage the latest discoveries, and applications that rely on valuable insights buried in these texts cannot effectively build upon published results. As a result, scientific progress stalls. Automatic Text Summarization (ATS) and Information Extraction (IE) are two essential fields that address this problem. While the two research topics are often studied independently, this work proposes to look at ATS in the context of IE, specifically in relation to Scientific IE. However, Scientific IE faces several challenges, chiefly, the scarcity of relevant entities and insufficient training data. In this paper, we focus on extractive ATS, which identifies the most valuable sentences from textual content for the purpose of ultimately extracting scientific relations. We account for the associated challenges by means of an ensemble method through the integration of three weakly supervised learning models, one for each entity of the target relation. It is important to note that while the relation is well defined, we do not require previously annotated data for the entities composing the relation. Our central objective is to generate balanced training data, which many advanced natural language processing models require. We apply our idea in the domain of materials science, extracting the polymer-glass transition temperature relation and achieve 94.7% recall (i.e., sentences that contain relations annotated by humans), while reducing the text by 99.3% of the original document.

1. Introduction

Scientific Information Extraction (IE) has become increasingly important as the number of scientific publications and journals grows exponentially [1]. While traditional IE remains a vast and active field of research [2–4], both open source and commercial Natural Language Processing (NLP) tools[1] can be leveraged to generate labels for machine-learned models. And while crowdsourcing semantic labeling systems still need control labeling quality, the assumption is that the task is attainable for laymen [5, 6]. Scientific IE faces its own additional challenges, including the scarcity of target entities in text, the lack of annotated training data, and the fact that generating quality labels from scientific text requires expertise, which can also be costly [7]. Since most NLP tasks rely on balanced, accurate, and carefully annotated gold standard labels, we propose that an often-overlooked crucial preliminary task to any Scientific IE tool, is the generation of these labels. We further advance that the complexity of scientific facts to be extracted often requires context in addition to the structured data. For these reasons, our work tackles Scientific Automatic Text Summarization (ATS) as a prerequisite to hybrid human-machine Scientific IE. Indeed, while ATS is largely defined and evaluated in terms of generating summaries similar to those generated by humans, Scientific IE reduces the extraction of scientific facts to extracting entities, relations or attributes for example. Instead, we proposed that ATS is required not

[1] For example, NLTK, SpaCy are two Python programming libraries that include well-developed NLP tasks such as Part-of-Speech Tagging, PERSONS, LOCATIONS etc.

only to reduce the sheer amount of text to be processed by ML algorithms but also to enable experts to review data output by these algorithms and extract additional context surrounding these facts when necessary.

This idea is tested in the field of materials informatics, which is generating great interest as a paradigm shift within Research & Development (R&D) aimed to fundamentally accelerate the time from innovation to market in materials science; it proposes to automatically process large amounts of data for targeted design of new materials with potentially high societal impact [7–11]. The challenges faced in this field however are not unique to materials science. Indeed, while bioinformatics is more mature, the extraction of new types of entities implies appropriately generated corresponding gold standards [12–14]. Our work aims to achieve generalizability in generating summaries towards Scientific IE by leveraging weak supervision and ensemble classification. We use Snorkel [15], a data programming software, to tentatively label sentences that contain three pieces of the target entity, here a polymer-T_g pair (polymer mention, glass transition or (T_g) mention and the actual temperature); we train three corresponding models and combine them to identify important sentences. It is important to note that unlike in the case of typical relations extraction work, identifying the target relations does not assume the preliminary identification of the entities composing the relation. We achieve 94.7% recall (i.e., sentences that contain relations annotated by humans), while reducing the text by 99.3% of the original documents. In tuning our models, we prioritize recall as our end-goal is to generate training data (we want to retrieve all the important facts). As anticipated, precision is low (58.9%), but we provide an analysis of the false positive, emphasizing the reduction in text to be processed and the useful context in these additionally extracted sentences (e.g., method of measurement of T_g).

The key novelty in our approach is the reframing of Scientific ATS from a practical Scientific IE point of view. To the best of our knowledge, even extractive ATS which extracts exact sentences from text as opposed to generating equivalent sentences (abstractive ATS) does not focus on scientific information other than citations to score the importance of sentences. We use key components of the ultimate target relation to train sentence extraction models, which will ultimately be used for scientific information extraction. The contributions of the papers are centered around the design and evaluation of a new type of extractive ATS model based on combining three weekly supervised models. We demonstrate that we efficiently retrieve sentences of interest, reduce and balance the text to be later annotated by machines and/or humans. Finally, we show the importance of capturing additional sentences which contain context information related to the structured portion of the target relations.

2. Related work

There is a wealth of important information buried in textual content growing exponentially in various archives of scientific publications. Figure 1 shows examples of sentences containing important information about polymers and their glass transition temperatures that would be useful to store in a structure format for future material design. IE and ATS are two fields focused on extracting valuable information from large amounts of data as manual extraction of scientific facts is time consuming, error-prone, costly, and ultimately impractical. This is particularly true in Scientific IE as extracting scientific

facts often requires domain knowledge, hence experts' time for accurate extraction, yet a recent bibliometrics study reported that approximately 2.5 million new papers are published each year [1]. Scientific IE has typically focused on named entity recognition and relations extraction as the valuable information locked can be protein reactions or properties of polymers used to design a new drug or a new material for example. However, previous works have also shown that scientific facts can be complex and require context information to be fully understood (e.g., method of measurement of a particular material's property) [16-18]. In this paper, we propose a novel approach to scientific text summarization and present it as the first step in a scientific hybrid human-computer information extraction pipeline to accurately extract scientific facts.

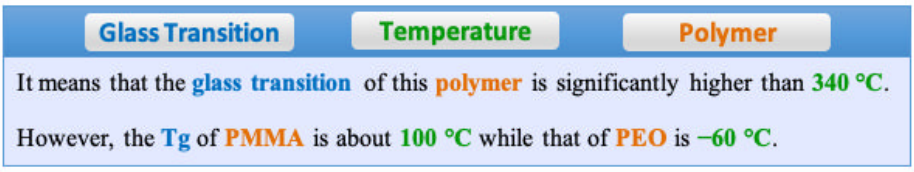

Figure 1: Sentences containing valuable polymer-T_g pair relation

ATS is an important NLP task that continues to gain more attention as the amount of textual content grows exponentially on the internet, and various archives of legal documents and scientific articles etc. [19]. Early works in ATS date back to the 1950s. In early work, the machine used statistical information derived from word frequency and distribution to compute a relative measure of significance, first for individual words and then for sentences; sentences scoring highest in significance were extracted and printed out to become the "auto-abstract." [20]. Since then, researchers have continued to improve ATS methods, which are either extractive, abstractive or hybrid. Extractive methods select the most important sentences from a document and combine them in a summary while abstractive — more recent and challenging — approaches generate a summary with new sentences [21]. Despite many advances, automatically generating accurate, complete, and human understandable summary remains a formidable challenge [22]. Difficulties include identifying *all* the most informative segments in documents, summarizing multiple documents, generating content that is similar to a human-produced summary without redundancies etc. There are several ATS surveys, a recent comprehensive survey includes abstractive which are drawing more attention with advances in deep learning [19].

Despite these advances, scientific text summarization is often focused on identifying citations and cited work. Citation-based summarization identifies relevant aspects of the paper through publications which have cited the target paper; applying this information to ultimately score valuable sentences within the target paper [23, 24]. This exemplifies leveraging additional context in extractive approaches as most of the variations between these reside in how sentences are scored and aggregated into a summary. Other scoring methods of sentences include using term frequency or term frequency-inverse document frequency (TF-IDF), topic-words, and ontologies to identify key words and sentences for example [25]. Different extractive methods involve first identifying topics as alternative representations of concepts in the paper, before identifying important sentences (i.e., that include these topics).

While one could make an argument for abstractive methods to summarize important concepts covered in scientific articles [26, 27] to generate human-readable summaries, we argue that extractive methods are more favorable in extracting sentences that contain important scientific facts that should not be altered in the summary (at least in terms of extracting facts that are ultimately to be stored in a database and used in applications). This is also the reason why scientific IE is an active research area, especially in materials science [7, 10]. Therefore, this work proposes an intermediate approach, a semi-supervised scientific text summarization that extracts sentences containing the targeted entity to be extracted with additional metadata. Since a major challenge in applying Machine Learning (ML) techniques in scientific ATS or IE is the availability of training data and difficulty in transferring knowledge from large models to new models [7], we use Snorkel, a data programming technique, which allows to approximately label data using Python labeling function (rather than hard-coded rules).

Snorkel was previously used in novel approaches to identify sentences which contain our target relations (polymer–T_g) [28, 29]. While Snorkel customarily assumes the entities are known and learns the different accuracies of approximate rules that link the entities in the relations, we identified sentences without knowing entities a priori. Authors used three sets of labeling functions to identify parts of the relation and combined them to identify relevant sentences. They later improved this approach, which assumed the relation is contained in a single sentence and expanded to extracting blobs, to achieve 100% recall [29]. This Ensemble labeling method, or ELSIE-Blob, did not use a classification model to extract sentences. Nevertheless, while ELSIE-Blob is also concerned with extracting the components of the relation, the polymer and the T_g, it may still miss metadata about the relation. Other previous works illustrate our motivation in the need to capture additional context information, sometimes manually about the target entities and relations [16–18, 30, 31]. In our work, we anticipate that extraction of facts from relevant sentences will sometimes be automated, however, experts' interventions may on occasion be necessary. Wallace et al. also pursue this goal, using a hybrid ML and crowdsourcing approach to identify published randomized controlled trials (RCTs) [32]. They use ML classifiers to recognize citations that are deemed highly unlikely to describe RCTs, deferring to crowdsourcing otherwise. In previous work, authors extracted the complex Flory-Huggins interaction parameter, a measure of miscibility between two entities using a combination of automation and crowdsourcing [17, 18]. Similarly, in [16], authors use a hybrid pipeline of automated and manual tasks to extract polymer-T_g pairs from text. These previous works, motivate this current approach of approximate multi-classification models to identify entities along with additional related context designed for human consumption.

3. Motivation

In recent years, there has been substantial interest in how the fields of ML and NLP can assist in extracting materials science data from the scientific literature [7, 10]. While bioinformatics has long fueled advances scientific IE and text mining of biomedical publications, materials informatics, which applies the principles of informatics to materials science and engineering to improve the understanding, use, selection, development, and discovery of materials and relies on the availability of large amounts of

data, is an emerging field [8, 9]. Therefore, there is an increasing demand for extracting important materials data to feed into computational modeling tools such as CALPHAD for example [33, 34].

Automatic IE, which has the potential to unlock these data combines rules, statistical and ML models. Recent advances in Deep Learning (DL), such as the BERT language model, have revolutionized the field and outperformed previous models in various NLP tasks [35]. However, scientific IE faces specific challenges highlighted in a recent study [7], including the computer-(un)friendly format of scientific text, the lack of training data, the sparsity of the target information and the difficulties of applying models trained on general corpora to scientific text. Even leveraging other pre-trained BERT-based models [36,37], requires generating new training data for corresponding new target entities or relations. We consider our work to be a first step in scientific IE workflow, and in addressing the aforementioned challenges. Indeed, since in science, it is important to retrieve the exact information, extracting relevant sentences, that is sentences that contain the target information addresses sparsity of the data, thereby reducing the amount of (irrelevant) text to process from which training data is to be generated, and reducing the imbalance between targeted facts and other texts. Importantly, scientific text summarization addresses the complexity of the data to be extracted. Data in materials science are particularly heterogeneous, based on the significant range in materials classes that are explored and the variety of materials properties that are of interest [10]. Polymers for example are of particular interest as they are both ubiquitous and challenging to extract due to the manner in which authors report on polymeric materials [38]. Polymers are large molecules composed of many repeating units that have a wide range of properties depending on their application. Previous work has also highlighted the difficulty in extracting polymer properties for example [16–18]. The Flory-Huggins (χ) parameter measures the interaction of compounds (polymer-polymer or polymer-solvent) and can be reported in text in multiple formats, including a number, different types of equations or a graph. Moreover, there are multiple ways to measure these properties, some well-known and some customized. Even in the case of a simpler property, polymer-T_g, materials scientists may need additional details like methods of measurement and type of polymerization (bulk or mass, method of measurement or number average molecular weight. Figure 2 shows two examples of context information for polymer-T_g pairs: the first describes a decrease in T_g during an experiment rather than reporting an actual temperature while the second explains how glass transitions were measured. Because of cases like this, which are not unique to polymer science, we propose that the ultimate solution to the crucial challenge of generating training data for scientific IE may be a combination of crowdsourcing and automation, that begins with scientific ATS. ATS reduces and balances the original amount of data, addressing the aforementioned scarcity challenge, and facilitates annotations by enabling the automated labeling of exact scientific facts with potential manual labeling of additional complex, and computer unfriendly context.

Using the empirical rule that 1% of toluene lowers the Tg by 5 °C, the increase in chain mobility was dramatic, leading to a Tg below room temperature.

The soft segment glass transition temperature (Tg) and the hard segment melting temperature of the PU nanocomposites were determined via differential scanning calorimetry (DSC) using a TA Instruments Q 1000 series DSC over a temperature range of −90 to 250 °C at a ramp rate of 10 °C min-1.

Figure 2: Examples of context information for the polymer-T_g pairs

4. Methodology

The principal objective of identifying polymer sentences has been explored in prior research and influences the methodology used here [28, 29]. The methods implemented include using Snorkel systems with a distinct strategy of comparing multiple, entity-specific sets of labeling rules with the goal of extracting the relationship between them. We will instead use three models trained using these labels, one for each of the primary target entities, which include polymer names, glass transition mentions, and temperature mentions. Together, these entities constitute a scientific fact or relation, a polymer-*Tg* pair. In Snorkel, we use data programming and weak supervision to create training labels. This software system uses *labeling functions*, or a set of approximate programming rules to label data; then using a discriminative probabilistic model and limited labeled training data, it learns the relative accuracies of these labeling functions to label large amounts of data [15].

In prior work, authors used an ensemble labeling method to identify sentences that contain the polymer, the T_g and the temperature mentions; ELSIE used three sets of Snorkel labeling functions to identify each of the target entities before extracting sentences using a majority labeling technique to identify target sentences which included all three entities [28]. To achieve 100% recall, ELSIE-Blob used a novel approach inspired by Depth-First-Search and Snowball sampling to extend the search for the three entities across multiple sentences, extracting blobs instead of sentences [29] as this information was indeed sometimes spread across more than a single sentence. It is important to note that neither ELSIE, nor ELSIE-Blob used ML models. In our new approach, we now aim to use the Snorkel discriminative models to identify each of the components of the target polymer-T_g pair (polymer name, temperature and T_g mention). We expect our approach to generate more false positives as labeling functions are approximate and there are three separate models trained on limited ground truth and highly imbalanced data. We hypothesize that leveraging an ensemble of three distinct models and the convergence of their labels will achieve comprehensive extractive scientific text summarization; retrieving sentences with distinct polymer relations and valuable supplementary context.

4.1 Data

The data originated from a keyword search from *Macromolecules*[2], a journal which specializes in materials science, and specifically polymer research. The dataset consists of 36 scientific articles and 10,821 total sentences which have been reviewed and labeled by materials scientists (i.e., extracted polymer-T_g pairs from the documents). We previously split the data into sentences and matched the polymer-T_g pairs to sentences within these documents [28]. The data is highly imbalanced with positive sentences (i.e., containing relevant information) accounting for only 48 of the 10,821 total sentences (0.4%). We now separate the ground truth in three different sets, one for each of the target entities. That is, using the original ground truth, the *polymer* model is trained using data in which polymer names are identified, the temperature and T_g models are trained on sentences that contain a temperature and/or glass transition mentions. Temperature and glass transition mentions are often related meaning

[2]https://pubs.acs.org/journal/mamobx

sentences identified in these models will often converge. The T_g model allows for and was constructed to achieve higher precision with fewer false positives. The *polymer* model required a unique set of labels which involved meticulous examination to identify any sentence with a polymer name or abbreviation. For each model, we split the data into 80–20 splits (with 8,656 training samples and 2,165 testing samples). Since we then need to combine output from the three models, the overall performance was measured through thirty randomized trials. Note that the imbalance in the data means that there are often only a few positive instances in the test set and highlights the efficiency of our models. None of the original data was preprocessed prior to modeling.

4.2 Snorkel

As previously mentioned, Snorkel is a software system which allows for highly efficient and accurate probabilistic labeling through weak supervision without the need for extensive hand labeled training data [15]. This approach provided the central framework behind the modeling. The labeling functions allow Snorkel to create an initial set of labeled data which can then be used to model the correlations between those outputs. These correlations are then used to create new confidence-weighted training labels to reduce noise and conflicts.

4.2.1 Text Preprocessing

The sample of sentences were not preprocessed or transformed prior to using Snorkel's modeling system. However, certain preprocessors inherent to Snorkel's labeling functions were used for basic data cleaning where needed. The preprocessors used involve transforming all text to lowercase, removing numeric characters, punctuation, and verifying parentheses are used properly.

4.2.2 Labeling Functions

The three entities which are paramount to identifying valuable polymer relations include glass transition or T_g mentions, temperature mentions, and the polymer names. Labeling functions are python functions which iterate through the text data provided and return True, False or None (i.e., “TG”, “Junk” (no TG present), and “Abstain”). The main difference between “Junk” and “Abstain” is that a “Junk” labels is a False labels asserting that the entity is not found, while an “Abstain” allows for a different labeling function to label the sentence. For example, not finding an acronym does not imply not finding a polymer name (abstain), while not finding a number is immediately equivalent to not finding a temperature (junk)s. These provide the initial labels within Snorkel which are then used to model their correlations and accuracies to produce a final set of probabilistic training labels. The T_g model includes six total labeling functions which largely encompass checks for any mention of a glass transition temperature. This model overlaps substantially with temperature, so much so that they were originally combined into a single model. Eventually, they were separated as we discovered the individual models allowed for better recognition of key material while avoiding redundant information. The temperature model includes five labeling functions which consider a broad range of information related to each entity. The model intends to identify temperatures while filtering out any sentence which does not contain a number and temperature in some form. The model further considers whether a connection exists between the detected temperature and a glass transition or polymer mention.

Conversely, the primary focus of the polymer model was to identify a single entity associated with a polymer name or abbreviation. This model also includes a total of five labeling functions but largely focused on keyword searches in conjunction with pattern identification using regular expressions while both the T_g and temperature models focused more heavily on the latter. Polymer names are not easily identifiable but do, in some cases share common patterns, including the prefix "poly" and abbreviations also beginning with the character "P". However, there are several instances where this is not the case. Unfortunately, no complete dictionary for polymer names exists and the standardized International Union of Pure and Applied Chemistry (IUPAC)[3] naming conventions often result in lengthy and, hence, rarely used names.

4.3 Word Embedding

To classify sentences, the model needs vectorized sentences — not strings. A count vectorizer, or matrix of token counts, was the original word embedding method used for each model. However, the polymer model performed well using term frequency-inverse document frequency (TF-IDF) vectorization; The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general [39]. Vectorization and the analyzer used within the TF-IDF vectorizer was found to be crucial to identifying the correct entities within each model.

4.4 Classification models

In this section, we describe each of the models in more detail. The classification models learn the embeddings of relevant sentences and separate them from other sentences. After experimenting with multiple classifiers, we substituted the default discriminative model in Snorkel with a Support Vector Machine (SVM) classifier with different kernels. While differences were not always noticeable, SVMs generally outperformed other classifiers. SVMs are known to perform well with high-dimensional numeric data. Hence, we mostly varied the pre-processors, vectorizers and SVM kernels across classifiers.

Glass Transition (Tg) Model

The model created to identify T_g mentions consists of six labeling functions. The first identifies keywords related to any mention of "glass transition" temperatures and their possible variants. The second labeling function is a simple check to remove any sentence with the acronym "TGA" rather than some form of "T_g" as they are not relevant for our models. The third labeling function is more complicated with various checks for any glass transition mention or related temperature. Examples include regular expressions identifying any instance of a "T_g" mention, as well as combinations of "T_g" and temperature. The function is finalized by labeling anything which remains as "JUNK", while abstaining from several patterns which may be indicative of a glass transition temperature. These include any sentence with a "T_g", "glass" or "transitions" keyword or any sentence which includes both a number and degree symbol, "T_g", "glass", or "poly" keyword. The preprocessors used for the first three

[3] IUPAC: https://iupac.org/polymer-edu/what-are-polymers/

labeling functions convert each sentence to lowercase and remove all punctuation with the exclusion of degree signs. The fourth labeling function is an additional check for any combination of a temperature and variation of "T_g", while excluding "TGA", which is a tool for thermogravimetric analysis. This function is meant to ensure any sentence which does not meet these requirements is designated as "JUNK". The final two functions are a continuation of this strategy to verify that no instances of glass transition mentions were missed and that any sentence without a relevant keyword is avoided. The three final functions use preprocessors to convert text to lowercase and the final labeling function removes all numeric characters for easier identification of significant abbreviations. For this model, we substituted the default discriminative model in Snorkel with a Support Vector Machine (SVM) classifier with an RBF kernel after experimenting with several classifiers and kernels.

Temperature Model

The temperature model is used to broadly identify any sentence with a number and degree sign. However, there are instances of overlap with the T_g model involved to lessen the number of positive cases: four labeling functions are used to identify any sentence with a temperature, while including constraints to specifically identify glass transition temperatures. Since the Snorkel model learns from overlap and differences in coverage from the labeling functions, these restrictions increase the confidence of classifying specific types of embedding. The first identifies the most apparent indicator for temperature through a keyword search for a degree sign (°, °C, °F) and common temperature scales. Kelvin is not included here as the degree symbol is not included in the notation. However, all three of the most common temperature scales will be accounted for in the following functions through regular expressions to identify temperatures, and often their association to a T_g mention. The first of these identifies general variations of "T_g" along with either a degree sign and "F" or "C", or a single "K" with a leading whitespace as punctuation was removed. This function further identifies "T_g" mentions with any numbers which are not included in abbreviations or words. Finally, any number followed by space and degree sign, as well as relevant temperature scales is identified as a temperature. The final two functions are similar and repetitive with one dedicated to checking for a numerical character mention within each sentence, while the other will attempt to find any indication of a temperature or degree symbol. A single preprocessor was used for the first three to remove any punctuation that does not include a degree or equal sign as these are beneficial to identifying temperature. Here, we used a TF-IDF vectorization and an SVM model with an RBF kernel.

Polymer Model

The polymer model consists of five labeling functions, largely focused on identifying significant keywords and abbreviations relating to polymer names. The first three functions are keyword functions which identify various forms of polymer names or chains of polymers. These functions use preprocessors to make all text lowercase and check the use of parentheses in each sentence. This ensures there are no mismatched parentheses, and any additional whitespace is removed. The final labeling functions use regular expressions to identify polymer abbreviations and names respectively with certain overlap. There are many instances of polymer references which do not follow any standardized form; however, these abbreviations often begin with "P", whereas polymer names can begin with "poly" or

"poly(". These patterns are included in the regular expressions, while tolerating a certain level of deviation. The polymer model utilized TF-IDF vectorization and an SVM model with a polynomial kernel.

4.5 Combination of models

The previously described models were developed to discover three principal entities in the target relation, or polymer-T_g pairs in scientific text. These include glass transition or T_g mentions, temperatures, and polymer names or abbreviations. Various methods of combining the output sentences from the models were considered. Our primary objective was efficient identification of relevant sentences containing the most valuable contextual detail as it relates to polymer discovery within scientific research. It could be expected that comparing the output labels of three separate models would result in a comprehensive list of sentences containing all relevant information (illustrated in Figure 3). The T_g and temperature models were originally combined into a single model as they are reasonably analogous and potentially repetitive. However, we ultimately discovered that separating the two models allowed for more precise results and fewer extraneous sentences. The results when applying only two models (T_g/temperature, and polymer model) were compared to those of the three-model framework, one for each entity (T_g, temperature, and polymer names). We purposefully prioritized recall over precision as extracting essential information was our primary consideration and anticipated that false positives allowed for more context extraction. We reframed the IE task as a scientific text summarization problem, rather than one purely for concise extraction of facts. Of course, this approach would also increase the overall number of sentences required for human review.

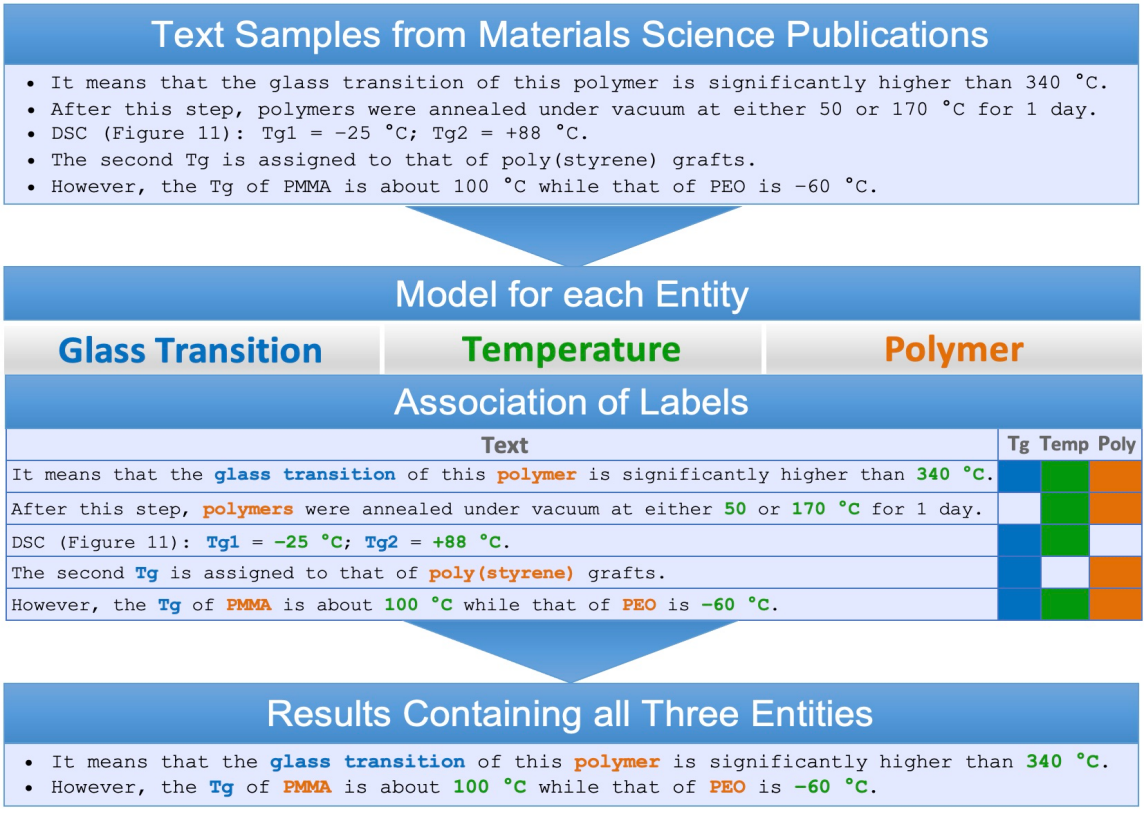

Figure 3: Three Model Ensemble System

5. Results & Discussion

Predictably, the model findings varied for each entity, notably between polymer names and glass transition or temperature references. Polymer naming conventions are variable and often challenging to identify reliably while simultaneously avoiding false positives. Temperature and "T_g" mentions are more precise and consistent resulting in fewer overall erroneous sentences. Detecting relevant sentences to avoid omitting crucial information while minimizing the number of false positives is vital to this effort. However, as previously mentioned, extracting supplementary context on or surrounding the target relation is often valuable when summarizing polymer research. This allows for a human to quickly review and determine which sentences are relevant or meaningful to the application, including context information about the target scientific fact.

5.1 Classification results

In this Section we report on each of the models individually, as well as different combinations of models with the final results for each in Table I and Table II. The original ground truth labels for the T_g and temperature models included only sentences with both a temperature and glass transition mention which expectedly resulted in substantially low precision. We then used labels from a prior ensemble method (ELSIE [25]) to evaluate the individual T_g and temperature models against more general labels using general T_g and temperature labeling functions as opposed to only those contained in the extracted relations.

Glass Transition (Tg) Model

The glass transition model demonstrated an average recall of 89.68% with a precision of only 74.75% against ELSIE T_g labels, and while some of the false positives were unnecessary, many also provided details later determined to be useful (See Table I and Section C). This result is consistent and marginally higher than previous results using rule-based methods. Indeed, previous work has shown that a rule-based method to identify T_g alone achieved 88% precision and 71% recall [13]. Looking for "T_g =" for example will allow for high precision and retrieve many directly reported T_g's. However, it will also likely miss some temperatures. In the same study, the precision and recall drastically decrease when matching the T_g to the correct polymer name (38% precision and 31% recall) [13] illustrating the difficulty in correctly identifying polymer names. When considering the more constraining original ground truth labels (only sentences containing both T_g and temperature), our model achieved nearly 97% recall at the cost of lower precision (21.1%) as shown in Table II. The glass transition model aimed to identify glass transition temperature with an emphasis on primarily including T_g mentions regardless of temperature where necessary. Many false positives are a direct result of this distinction. Another reason for the low precision is that there may be T_g and temperature mentions not related to an extracted pair in the limited ground truth. Finally, the low precision can also be attributed to the variability and scarcity of glass transition mentions, leading to increased difficulty and fewer examples to train the model.

Temperature Model

The temperature model averaged a recall of 91.09% with a notably higher precision of 97.79% (See Table I). The labels used to calculate these metrics differ from the T_g model as identifying a temperature involves a broader range of data. While the temperature model still favors glass transition temperatures in particular, it does tolerate more exceptions than the T_g model. These labels were obtained from prior research [25] as well and allowed for more precise performance evaluation. As shown in Table II, when considering the more restrictive original ground truth labels, this model achieves an even higher recall than the T_g model at 98.7%. This model however has a notably lower precision of 11.1%, which is not surprising due to the greater prevalence of temperature in polymer related articles. The increased frequency and association of temperatures with variables other than glass transition (i.e., melting point) contribute to the lower precision, but the high recall is optimal in this case as omitting information is considerably more detrimental.

Polymer Model

The polymer model achieved a recall of 99.43% with few polymer references missed (See Figure 3). However, the precision was low at 29.32% and largely suffered due to the challenge of identifying complex polymer names and the scarcity of entities in text, while also avoiding similar chemical elements or acronyms. As previously mentioned, this was the case in previous studies resulting in additional crowdsourcing efforts to retrieve polymer names [13].

	Recall	**Precision**	**F1-score**	**Test Accuracy**
Glass Transition Model *& Margin of Error*	89.68% ± 1.37%	74.75% ± 2.15%	0.81 ± 0.01	99.16% ± 0.08%
Temperature Model *& Margin of Error*	91.09% ± 0.85%	97.79% ± 0.7%	0.94 ± 0.005	99.43% ± 0.04%
Polymer Model *& Margin of Error*	99.43% ± 0.17%	29.32% ± 0.38%	0.45 ± 0.005	60.68% ± 0.32%

Table 1: Individual model results using ELSIE Tg & temperature labels (average of 30 trials)

	Recall	**Precision**	**F1-score**	**Test Accuracy**
Glass Transition Model *& Margin of Error*	96.89% ± 1.71%	21.12% ± 1.39%	0.34 ± 0.02	98.06% ± 0.1%
Temperature Model *& Margin of Error*	98.66% ± 1.29%	11.07% ± 0.95%	0.2 ± 0.015	95.72% ± 0.14%

Table 2: Tg and temperature model results using Tg/Temp ground truth (average of 30 trials)

5.2 Combining models

Once each sentence is determined to contain a reference to glass transition, temperature, or polymer name, the final labels for all three are compared to determine the overlap, that is, any sentence which contains all three entities. Theoretically, this should provide a succinct summary of the text containing all sentences with the pertinent information. Here, we discuss the overall results of combining the model output and the text summarization application of this combination.

The following results shown in Table III and IV below are achieved through the comparison of all three entity models. Due to the combination of output for three models and to account for variability in results, we show the aggregated results of 30 trials. The sentences identified include those which were simultaneously predicted by the glass transition model, polymer model, and temperature in our test set, which resulted in 94.7% recall and 58.9% precision on average over 30 trials. This includes fewer than three false negatives out of 10,000 total sentences, meaning at most, only three relevant sentences from ~33 scientific papers were missed. The number of false positives is considerably higher with an average of 28 total sentences of 10,000; however, many of these false positives were deemed contextually valuable. The total sentences returned, including true positive and false positive consist of ~68 of 10,000 total. The summarization therefore reduces the text by 99.3%.

Throughout testing, we noticed the T_g and temperature models alone demonstrated a comparable performance to the final model association. This is not surprising as the polymer model had significantly lower precision. The recall when using only the temperature and T_g models was ~98% up from ~95% when including the polymer labels, with a slightly lower precision at ~55% down from ~59%. This outcome is valuable as it allows for the inclusion of crucial sentences which were previously missed within the initial label comparison due to the polymer model. Using the labels for T_g and temperature demonstrated one false negative on average out of 10,000 total sentences (or 33 published articles) with only 5–6 additional false positives from the prior results. Due to the potential consequences of these false negatives (missed information), and our priority to ensure valuable information is seldomly missed, these results are significant.

			Out of 10,000 Sentences			
	Recall	**Precision**	**FN**	**FP**	**Total Sentences**	**% Reduction**
T_g & Temp Model *& Margin of Error*	97.76% ± 1.5%	54.94% ± 3.5%	1.08 ± 0.94	33.72 ± 3.63	74.98 ± 7.14	99.25% ± 0.07%
T_g Temp & Poly Model *& Margin of Error*	94.71% ± 2.1%	58.93% ± 3.5%	2.46 ± 0.71	28.02 ± 3.39	67.90 ± 7.15	99.32% ± 0.07%

Table 3: Final label results determined by relation of individual model labels (average of 30 trials). Two model ensemble (Tg & Temp) vs. three model ensemble (Tg, Temp, & Poly)

5.3 Discussion of Errors and Text Summarization

Text summarization of research publications in specialized domains is often challenging and commonly involves errors which are unavoidable. However, this ultimately allows room for interpretation as the importance of each sentence can vary depending upon the intended use. Our primary goal within materials research is to summarize each article through the extraction of information related to polymers and their corresponding glass transition temperatures. These instances are relatively infrequent but immensely valuable, therefore several combinations of models and their results were explored.

5.3.1 Error Analysis

In this section we discuss the classification errors through the lens of text summarization. We aim to demonstrate that while the combination of models lack precision, this method retrieves the most valuable information while providing context related to the target. The false negatives include sentences which have been overlooked by the models and are demonstrated by the two trials in Figure 4 & 5. The two false negatives between these two trials represent the only sentences missed throughout the total 30 trials. The first (in Figure 6) is largely due to the absence of a temperature or T_g mention as the sentence contains a polymer name and abbreviation, while the associated details are located within the sentence following.

False Negatives:

- The azo-polymer material, poly[4‘-[[2-(acryloyloxy)ethyl]ethylamino]-4-nitroazobenzene], often referred to as poly(disperse red 1 acrylate) (hereafter pdr1a), was synthesized as previously reported.

False Positives:

- *According to DSC data, the synthesized polymers did not show any glass transition up to 300 °C.*
- *Here, a PS96k film (75 nm thick) was thermally preannealed at 105 °C for 72 h (just above the bulk Tg of PS) prior to annealing in a saturated environment of 80:20 wt toluene/ethanol mixture (Figures 8a-c).*
- *Using the empirical rule that 1% of toluene lowers the Tg by 5 °C, the increase in chain mobility was dramatic, leading to a Tg below room temperature.*
- *The LCST transitions were broad, extending over a 10–15 °C range, and the low molecular weight of these chains may explain the observed transition broadening.*
- *Differential scanning calorimetric (DSC) analysis of the obtained polymer indicated that the glass transition temperature (Tg) was −58 °C (Figure S1, Supporting Information).*

Figure 4: Single trial error output (one False Negative and five False Positives)

The false negative sentence was correctly labeled by the polymer model but without additional components of the relation, the final label was not achieved. The following single trial example (Figure 7) demonstrates a similar instance where a temperature and T_g are mentioned but there is no indication of a polymer. The false positives within these two trials are all sentences we would consider to be valuable contextual information. Each contains a reference a glass transition temperature and while polymer is not always included, this further supports a tolerance and selectiveness of errors. The false negatives between these two (Figures 6 & 7) indicate the only two missed sentences throughout the 30 total trials.

False Negatives:

- *A small piece of rectangular shape was cut out of the amorphous strip and stretched by about 5 times the original length above the hot plate around the glass transition temperature (97 °C).*

False Positives:

- *Of importance, the formed polymer is a typical semicrystalline thermoplastic, possessing a Tg of 15.6 °C and a Tm of 96.7 °C.*
- *No large difference is observed between the two different types of networks: the main relaxation at high temperature, Tα, taken as the maximum of tan δ peak and associated with the glass transition, and the sub-Tg relaxation near −80 °C, associated with the change of conformations of cyclohexyl groups in the monomer units31 occur at the same temperatures.*
- *Here, a PS96k film (75 nm thick) was thermally preannealed at 105 °C for 72 h (just above the bulk Tg of PS) prior to annealing in a saturated environment of 80:20 wt toluene/ethanol mixture (Figures 8a−c).*
- *Similarly, from dielectric measurements in an immiscible PC/ABS system, a low frequency interfacial polarization at the boundaries between the occluded conductive ABS component above its Tg (109 °C) and the glassy PC matrix (Tg of 146 °C) was observed.*

Figure 5: Single trial error output (one False Negative and four False Positives

Figure 6 illustrates further examples of context considered to be valuable. Note that in these false positives, the model presumably identified all three entities even when mistakenly labeling a chemical as a polymer or a standard name like "polymer" as a polymer name. We propose that these false positives spread out through documents contain important information as to the polymer and/or the glass transition of materials that experts may be scanning for in publications. The first sentence does not contain a precise T_g but identifies the method by which it was measured. Regarding the second instance, T_g's are generally discussed in the paper pointing to the type of polymers synthesized, rather than mentioning an exact T_g. Once again, it also mentions the method of measurement. The third does in fact mention a T_g precisely, however, DSC is not a polymer acronym but stands for "differential scanning calorimetry" which is yet another technique used to measure polymer properties. The fourth and fifth instance are similar examples where T_g was included but a distinct polymer was not mentioned. While this qualitatively supports our hypothesis, we can present these results to experts to determine their relevancy (including whether the context information is valuable) and re-evaluate precision.

- *The soft segment glass transition temperature (Tg) and the hard segment melting temperature of the PU nanocomposites were determined via differential scanning calorimetry (DSC) using a TA Instruments Q 1000 series DSC over a temperature range of −90 to 250 °C at a ramp rate of 10 °C min-1.*
- *Although the corresponding copolymers were afforded with perfectly alternating nature and excellent regiochemistry control, only glass-transition temperatures of around 8.5 °C were observed in the differential scanning calorimetry (DSC) curve, demonstrating that the polymers are completely amorphous (see Supporting Information Figure S3).*
- *DSC: Tg = 83 °C.*
- *The lowered Tg of the coil block (−5 °C) led to the formation of the highly ordered MPS hexagonal array structure.*
- *In both of these former cases, the attached polymers were high Tg (100 °C).*

Figure 6: Additional valuable False Positives

5.3.2 Alternate Combination of Models

Through testing various combinations of the ensemble, we discovered that the T_g and temperature models alone proved capable of detecting the most valuable sentences while maintaining a comparable precision to the original three models. While this led to additional false positives, many were often deemed beneficial as valuable information may not have a direct polymer reference. One instance concerns the first false negative case discussed previously (identified in Figure 4). This sentence remained undetected by the three-model ensemble due to the lack of a polymer name or reference. The potential benefit of excluding the polymer model to include these sentences is exemplified in the results of Table III and IV. However, the polymer model can still provide valuable insight when extracting context and removing redundant information in the final text summarization. After labeling all sentences which contain both a temperature and T_g mention, we found extracting the surrounding sentences (two on either side) and checking each for a polymer name improved the summarization and helped achieve 100% recall. Further, if no clear polymer reference was found in the original sentence or those in close proximity, we determined the sentence can reasonably be removed from the final summarization altogether as it does not contain a polymer-T_g pair

To illustrate this strategy, we use the T_g and temperature models as mentioned previously, and extract final labels from their association (where both agree). Those sentences along with the two immediately preceding and following each sentence are then checked for a polymer label using the polymer model. If any of these sentences are labeled as containing a polymer name or abbreviation, they are then added to the text summarization. However, if none of these sentences or the original contain a polymer reference, the original sentence is removed from the final data. It is important to note however, that these sentences could not be shuffled as in the previous trials due to the need for sentences to maintain their original sequence. Documenting sentence indices will be necessary in future work to reconstruct an ordered summary and demonstrates a realistic example of intended use in. Three examples of single document summarizations have been included below to further demonstrate the benefit of prioritizing glass transition and temperature, while subsequently using the polymer model to provide context. The green highlighted portion represents the sentences initially identified by the glass transition and temperature models which are considered gold standard (or true positives as they contain a polymer reference), while the blue highlighted portion denotes those sentences labeled by the T_g and temperature models (considered false positives as they do not contain a polymer reference). All additional nonhighlighted text includes adjacent sentences retroactively identified by the polymer model (as each contains an appropriate abbreviation, name or polymer reference).

The completely saturated homopolymer containing Si(CH3)3 group (APNSi) and a copolymer containing both Si(CH3)3 and n-hexyl side groups (ACPNHSi) were prepared. The polymers have molecular mass of about 300 000 and a high glass transition temperature (Tg > 340 °C). Gas permeation properties of the polymers obtained were studied. It was shown that the properties of the polymers prepared (such as gas permeability, free volume, solubility coefficients) strongly depend on the nature of the side groups and the cis-content in the main chains. All ROMP-type polynorbornenes have relatively flexible chains: the glass transition temperatures in the most cases are in the range 24–140 °C. However, since norbornenes are cycloolefins, they can be also polymerized via the opening of double bonds in the presence of Ni and Pd catalysts. APNSi and ACPNHSi were soluble in aromatic solvents and chloroform. According to DSC data, the synthesized polymers did not show any glass transition up to 300 °C. The TGA scans indicated that APNSi has 5% decomposition in air of 340 °C and in argon of 450 °C (Figure 1). It means that the glass transition of this polymer is significantly higher than 340 °C. The density of APNSi was found to be equal to 0.883 ± 0.001 g/cm3.

Figure 7: Paper summarization (Doc ID ma061215h)

As expected, the thermograms exhibit two glass transition temperatures assigned to both blocks (Figure 11). The Tg values of poly(VDF-co-BDFO) copolymers incorporating 5–10 mol % of BDFO were ranging from −16 to −11 °C, respectively, hence being higher than that of PVDF (Tg = −40 °C) because of the presence of the −C6F12Br grafts. After grafting poly(styrene), the resulting graft copolymers exhibit two Tgs at about −30 and +90 °C. The second Tg is assigned to that of poly(styrene) grafts. As the Tg depends on the molecular weight values, it is not surprising that the PS side segments exhibit a Tg ranging from 80 to 100 °C since their molecular weights are lower than 10 000 g mol-1. For example, the molecular weights of poly(styrene) grafts in the analyzed graft copolymers were estimated at around 8100 g mol-1 and this explains such low Tg value.

Figure 8: Paper summarization (Doc ID ma061554a)

The added detail and inclusion of referenced figures in text summarization (as seen in Figure 7 & 8) allows for the intended user to access details which may be crucial to their research. Theoretically, this information would be extracted and catalogued per document with the ability to view the originally identified sentences or added context dependent upon the application. The below text summarization (Figure 9) includes previously mentioned false negatives where the entities were split between more than one sentence and was consequently overlooked by the initial model ensemble. With the adjusted methodology, the sentence is correctly identified by the T_g and temperature model and checking the surrounding sentences for polymers returns additional context and our approach achieves 100% recall for this trial. In previous ensemble trials, the single highlighted sentence was identified but not the first which contains the relevant polymer name.

The azo-polymer material, poly[4'-[[2-(acryloyloxy)ethyl]ethylamino]-4-nitro azobenzene], often referred to as poly(disperse red 1 acrylate) (hereafter pdr1a), was synthesized as previously reported. The prepared material was determined to have a molecular weight of 3700 g/mol, and a corresponding Tg in the range 95–97 °C. Samples for patterning were prepared by spin-coating the azo-polymer solutions (pdr1a in anhydrous THF solvent) onto cleaned glass microscope slides.

Figure 9: Paper summarization (Doc ID ma061733s)

The text summarization is potentially instrumental to materials scientists through the considerable reduction of reading time to label or confirm automatic labeling of entities. The average article length from the 36 used in this research is 300 sentences, illustrating the benefit of this reduction to fewer than 10 sentences on average, or more than 96.67% text reduction (Figures 7, 8, & 9). The information which is vital to experts is readily available along with any desired context.

5.4 Limitations

Several limitations exist within text summarization, particularly as it relates to summarizing technical research publications. The benefit to Snorkel includes the ability to produce highly accurate probabilistic labels using efficient modeling techniques which still allow the user to develop complex and specific labeling functions. Leveraging weak supervision allows for the avoidance of extensive hand labeled training data, which can be costly and time consuming [15]. While these models are comparable to the quality provided by substantial hand labeled training data, the tradeoff for increased level of efficiency is often a greater opportunity for errors. However, this compromise is tolerable due to the substantially reduced cost of data and virtually 100% recall.

The models additionally demonstrate lower precision with several false positives, which can often be attributed to the wide variance of polymer names and technical jargon. While many of these false positives were deemed useful and provide supplementary detail, some are considered unnecessary and frequently redundant (as seen in Figure 10). Examples include instances where a temperature may be mentioned but shares no relation to a glass transition or polymer. Various erroneous cases contain the word "transition", a temperature and/or acronym which are easily mistaken for polymer references. The probable cause of these errors are emphasized in bold for each case in Figure 10. Each sentence will contain similar features to entities we aim to identify; hence, further efforts to improve precision remain imperative.

- *The double peak shape and the temperature of the Curie transition were not modified by annealing above the Curie* ***transition*** *(15, 45, 105, and 225 min at* ***125 °C****) as shown in Figure* ***S23****, for films formed from the melt.*
- *The micellar solutions exhibited a rod-to-sphere* ***transition temperature around 60 °C****.*
- *The* ***LCST transitions*** *were broad, extending over a* ***10–15 °C*** *range, and the low molecular weight of these chains may explain the observed* ***transition*** *broadening.*
- *Aqueous solutions of* ***poly(N-isopropylacrylamide) (PNIPAM)*** *exhibit a lower critical solution temperature* ***(LCST) transition at 32 °C****; above this temperature, the solubility of the polymer is vastly reduced.*
- ***DSC*** *over a* ***temperature range of −90 to 250 °C*** *at a ramp rate of 10 °C min-1.*

Figure 10: Superfluous False Positives

5.5 Future Work

In future work, we will combine our work with the results from ELSIE-Blob, we expect that overlapping our sentences with sentences extracted using ELSIE-Blob will reduce the false positive and only keep context information that is "close" to sentences and/or blobs containing facts (ELSIE-Blob also achieved 100% recall, with 74% precision, while reducing the text to 6%). Using both systems concurrently can help increase confidence in sentences containing facts and detecting sentences containing context. Comparable to other works, our ultimate goal remains to build a hybrid human-machine pipeline which drastically summarizes and reduces text that subsequently must be processed by machines and/or humans [28, 29, 40, 41]. The context information, such as the method of measurement illustrated in our false positives demonstrates that some information cannot be reliably extracted by machine with minimal oversight. However, an expert can efficiently extract this information when presented with text localization and summarizing the target information. Other improvements may be achieved by experimenting with different vectorizers (e.g., Word2Vec or FastText) [42, 43], as well as different classification models and/or kernels for the SVM classifiers. While SVMs expectedly outperform basic models such as linear regression, we can experiment with neural network classifiers. We additionally plan to populate an accessible tool or database for materials scientists to readily summarize and identify key information in numerous polymer related publications. This would require exploring methods to subsequently label facts within relevant sentences (e.g., identifying important features/words). As demonstrated in Figures 7, 8, and 9, the potential to provide further surrounding context in addition to sentences originally labeled by the ensemble method may prove vital in certain applications. While experts may only expect the explicit polymer name and glass transition metrics, others could require greater detail and further references within each publication. In this instance, several scientific facts are automatically extracted, however, supporting contextual information may require manual extraction. Finally, we envision conducting a similar study for melting points and various other details to retrieve a comprehensive summary of polymers and their properties.

6. Conclusion

In this work, we define text summarization in the context of scientific information extraction. Three weakly supervised models were combined to determine which sentences contain significant scientific facts. In this case, our focus primarily concerns polymer references, temperatures, and glass transition mentions. Each model is trained on a set of approximate rules describing a single entity and evaluated through predicting whether a sentence contains the relevant entity; finally, any sentence predicted or "flagged" by all three models is extracted. This weakly supervised ensemble of models was evaluated through comparison of the extracted sentences to sentences that contain entities previously extracted by humans. We achieve 94.7% recall (i.e., sentences that contain relations annotated by humans), while reducing the text by 99.3% of the original documents. While precision is lower (58.93%), we demonstrate the importance of retrieving additional context for scientific data to be subsequently labeled and extracted by machines and/or humans. We propose that our system is a prerequisite to 1) generating balanced and reduced training data for advanced NLP models, 2) distinguishing between

data that is to be extracted by machines and that is to be reviewed by humans for context information, before finally performing the extraction. We implemented a method which prioritized the results of the T_g and temperature models and examined nearby sentences for polymer references which may have been excluded originally. This procedure achieved 100% recall as properties previously missed due to information spanning multiple sentences were now accounted for; emphasizing the importance of recognizing both precise entities as well as supplementary detail. This in turn supports our hypothesis that polymer relations and their surrounding context (associated polymer names, temperatures, methods of measurement, etc.) can be identified through the intersection of individual model labels. Further, a single publication can be reduced to include only the most valuable sentences relating to the polymer entities defined in each model. We intend to further explore the generalizability of these methods among other polymer properties as well as broader scientific relations.

References

[1] Ware, M., & Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing.

[2] Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2018). A survey on open information extraction. arXiv preprint arXiv:1806.05599.

[3] Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2020). Information extraction meets the semantic web: a survey. Semantic Web, 11(2), 255-335.

[4] Lim, C. G., Jeong, Y. S., & Choi, H. J. (2019). Survey of temporal information extraction. Journal of Information Processing Systems, 15(4), 931-956.

[5] Roit, P., Klein, A., Stepanov, D., Mamou, J., Michael, J., Stanovsky, G., ... & Dagan, I. (2019). Controlled crowdsourcing for high-quality QA-SRL annotation. arXiv preprint arXiv:1911.03243.

[6] Dhillon, J., Gupta, V., Govil, R., Varshney, B., & Sinha, A. (2019, March). Crowdsourcing of Hate Speech for Detecting Abusive Behavior on Social Media. In 2019 International Conference on Signal Processing and Communication (ICSC) (pp. 41-46). IEEE.

[7] Hong, Z., Ward, L., Chard, K., Blaiszik, B., & Foster, I. (2021). Challenges and Advances in Information Extraction from Scientific Literature: a Review. JOM, 73(11), 3383-3400.

[8] National Science and Technology Council (US). (2011). Materials genome initiative for global competitiveness. Executive Office of the President, National Science and Technology Council.

[9] de Pablo, J. J., Jackson, N. E., Webb, M. A., Chen, L. Q., Moore, J. E., Morgan, D., ... & Zhao, J. C. (2019). New frontiers for the materials genome initiative. npj Computational Materials, 5(1), 1-23.

[10] Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y. J., & Hiszpanski, A. M. (2020). Data-driven materials research enabled by natural language processing and information extraction. Applied Physics Reviews, 7(4), 041317.

[11] Mueller, T., Kusne, A. G., & Ramprasad, R. (2016). Machine learning in materials science: Recent progress and emerging applications. Reviews in computational chemistry, 29, 186-273.

[12] Huang, C. C., & Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics, 17(1), 132-144.

[13] Dogan, R., & Leaman, R. Zhiyong lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. Journal of biomedical informatics, 47.

[14] Dogan, R. I., & Lu, Z. (2012, June). An improved corpus of disease mentions in PubMed citations. In BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (pp. 91-99).

[15] Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., & Ré, C. (2016). Data programming: Creating large training sets, quickly. Advances in neural information processing systems, 29.

[16] Tchoua, R. B., Chard, K., Audus, D. J., Ward, L. T., Lequieu, J., De Pablo, J. J., & Foster, I. T. (2017, October). Towards a hybrid human-computer scientific information extraction pipeline. In 2017 IEEE 13th international conference on e-Science (e-Science) (pp. 109-118). IEEE.

[17] Tchoua, R. B., Qin, J., Audus, D. J., Chard, K., Foster, I. T., & de Pablo, J. (2016). Blending education and polymer science: Semiautomated creation of a thermodynamic property database. Journal of chemical education, 93(9), 1561-1568.

[18] Tchoua, R. B., Chard, K., Audus, D., Qin, J., de Pablo, J., & Foster, I. (2016). A hybrid human-computer approach to the extraction of scientific facts from the literature. Procedia computer science, 80, 386-397.

[19] El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, 113679.

[20] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4), 309-317.

[21] Moratanch, N., & Chitrakala, S. (2016, March). A survey on abstractive text summarization. In 2016 International Conference on Circuit, power and computing technologies (ICCPCT) (pp. 1-7). IEEE.

[22] Hou, L., Hu, P., & Bei, C. (2017, November). Abstractive document summarization via neural model with joint attention. In National CCF Conference on Natural Language Processing and Chinese Computing (pp. 329-338). Springer, Cham.

[23] Mei, Q., & Zhai, C. (2008, June). Generating impact-based summaries for scientific literature. In Proceedings of ACL-08: HLT (pp. 816-824).

[24] Qazvinian, V., & Radev, D. R. (2008). Scientific paper summarization using citation summary networks. arXiv preprint arXiv:0807.1560.

[25] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268.

[26] Cohan, A., & Goharian, N. (2018). Scientific document summarization via citation contextualization and scientific discourse. International Journal on Digital Libraries, 19(2), 287-303.

[27] Alampalli Ramu, N., Bandarupalli, M. S., Nekkanti, M. S. S., & Ramesh, G. (2019, September). Summarization of research publications using automatic extraction. In International Conference on Intelligent Data Communication Technologies and Internet of Things (pp. 1-10). Springer, Cham.

[28] Murphy, E., Rasin, A., Furst, J., Raicu, D., & Tchoua, R. (2021, June). Ensemble labeling towards scientific information extraction (ELSIE). In International Conference on Computational Science (pp. 750-764). Springer, Cham.

[29] Murphy, E., Rasin, A., Furst, J., Raicu, D., & Tchoua, R. (2021, September). Ensemble Labeling towards Scientific Information Extraction (ELSIE)—Blob Extraction. In 2021 IEEE 17th International Conference on eScience (eScience) (pp. 11-20). IEEE.

[30] Seifert, C., Granitzer, M., Höfler, P., Mutlu, B., Sabol, V., Schlegel, K., ... & Kern, R. (2013, July). Crowdsourcing fact extraction from scientific literature. In International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data (pp. 160-172). Springer, Berlin, Heidelberg.

[31] Wazny, K. (2018). Applications of crowdsourcing in health: an overview. Journal of global health, 8(1).

[32] Wallace, B. C., Noel-Storr, A., Marshall, I. J., Cohen, A. M., Smalheiser, N. R., & Thomas, J. (2017). Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. Journal of the American Medical Informatics Association, 24(6), 1165-1168.

[33] Olson, G. B., & Kuehmann, C. J. (2014). Materials genomics: from CALPHAD to flight. Scripta Materialia, 70, 25-30.

[34] Mueller, T., Kusne, A. G., & Ramprasad, R. (2016). Machine learning in materials science: Recent progress and emerging applications. Reviews in computational chemistry, 29, 186-273.

[35] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[36] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

[37] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

[38] Audus, D. J., & de Pablo, J. J. (2017). Polymer informatics: Opportunities and challenges. *ACS macro letters*, *6*(10), 1078-1082.

[39] A. Rajaraman and J. D. Ullman, Mining of massive datasets. Cambridge University Press, 2011.

[40] Tchoua, R., Ajith, A., Hong, Z., Ward, L., Chard, K., Audus, D., Patel, S., de Pablo J. & Foster, I. (2019, September). Active learning yields better training data for scientific named entity recognition. In 2019 15th International Conference on eScience (eScience) (pp. 126-135). IEEE.

[41] Tchoua, R., Ajith, A., Hong, Z., Ward, L., Chard, K., Belikov, A., Patel, S., de Pablo J. & Foster, I. (2019, June). Creating training data for scientific named entity recognition with minimal human effort. In International Conference on Computational Science (pp. 398-411). Springer, Cham.

[42] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

[43] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.