Towards hybrid human-machine scientific information extraction

Roselyne Tchoua [a], Aswathy Ajith [a], Zhi Hong [a], Logan Ward [b,c], Kyle Chard [b,c], Debra Audus [d], Shrayesh Patel [e], Juan de Pablo [e] and Ian Foster [a,b,c]

[a] Department of Computer Science, University of Chicago, Chicago, IL, USA, [b]Globus, University of Chicago, Chicago, IL, USA, [c]Data Science and Learning Division, Argonne National Laboratory, Argonne, IL, USA, [d]Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD, USA, [e]Institute for Molecular Engineering, University of Chicago, Chicago, IL, USA

Author Email: roselyne@uchicago.edu

A wealth of valuable research data is locked within the millions of research articles published every year [1]. Extracting pertinent scientific facts (e.g., materials properties, known variants in genomics, population statistics etc.) from those articles has become an unmanageable task for researchers. This problem hinders the advancement of science, making it difficult to build on existing results, avoid unnecessary repetition, and to translate results into applications. Moreover, since these data are often loosely encoded in esoteric scientific articles intended for human consumption, they are, in general, not machine accessible. Thus, it is not often tractable to develop studies that automatically leverage this valuable information.

There has been significant research focus on the identification and extraction of scientific named entities and entity relations, particularly in medicine and biology [2-7]. In some cases, domain experts are tasked with manual extraction of facts—a laborious and painstaking task. More recently, researchers have employed experts to rigorously annotate text to train machine learning information extraction methods. In either case, the process repeats itself as new literature is published or new domain-specific entities are sought. However, in other fields, or fields in which there are no such pool of experts dedicated to the task of manual extraction or annotation, these facts remain buried in publications.

In this talk we describe our efforts to develop human-machine methods for automatically extracting scientific facts from literature. Our overarching goal is to decrease the amount of manual extraction – a tedious, time-consuming, and error-prone process – and automate extraction activities where possible. With the assumption that we cannot entirely automate extraction we instead seek to prioritize human involvement to optimize extraction accuracy. As part of a long-term project to create a dictionary of polymer names and a digital handbook of polymer properties, we have developed three hybrid Information Extraction (IE) systems.

The first is χDB [8,9]. In this project we developed a hybrid model to extract a complex materials property: the Flory-Huggins interaction (χ) parameter, a measure of miscibility between two entities—typically a polymer and either another polymer or a solvent. This property is not only important in the design of new materials, as it determines how compatible – or incompatible – two materials are, it is also particularly challenging to extract due to the fact that it is published in heterogeneous data formats (e.g., text, figures, tables) and is represented in several different temperature-dependent expressions. Moreover, identifying and storing the χ values only makes sense if the corresponding polymers, solvents, molecular masses, methods, errors, and other measurement information are also captured. For these reasons, our approach required a considerable amount of manual extraction. Hence, the χDB model consists of an automated Web IE phase followed by a crowdsourced curation phase. Using χDB, we have extracted 263 χ values, a number comparable to that found in in the Physical Properties of Polymers Handbook [10]. We extracted more measured—as opposed to cited— χ values partly because we were able to collect values reported after the 2007 publication of the Handbook (84 of our χ values are from 2010 to 2015). We were able to improve the publication selection process considerably, decreasing the number of reviewed publications that do not contribute to the χ database from 61.5 % to 13.1 %. These results emphasize the poten-

978-1-5386-7933-3/18/$31.00 ©2018 IEEE

tial for using hybrid human-computer approaches to create and maintain a digital database of properties that is more comprehensive and up to date than any survey publication.

The second system we implemented is an IE pipeline for the Glass Transition Temperature (T_g) of polymers [11]. T_g is defined as the temperature at which a polymer transitions from a solid, amorphous, glassy state to a rubbery state as the temperature is increased. As the properties for the two states are drastically different, the glass transition plays a key role in both choosing a polymer for a given application and in the processing of the polymeric material. T_g is more straightforward to extract than χ as it is expressed as one temperature for a single polymer and there are only a handful of well-established methods to measure it. Hence, we explored a more automated approach that combines a general-purpose NLP toolkit to parse text and perform preliminary recognition (ChemDataExtractor [12]); specialized domain-specific models to identify entities and relationships; a ranking system to prioritize crowdsourced tasks; and a crowdsourcing framework to review candidate relationships. We extended ChemDataExtractor to extract T_g and identified 1,442 T_g candidates—text fragments with characteristics suggestive of a T_g value, but often with various irregularities. Subsequent automated and crowdsourcing curation steps then processed these candidates, in some cases confirming and/or completing a polymer–T_g value and in others establishing that no such value is in fact present. Curating the output of the NLP extraction phase required only a half-hour of expert time and a combined six hours of untrained crowds. To date, we have extracted 259 T_g values from 6090 articles. In comparison, the recent edition of the expert-curated Physical Properties of Polymer Handbook [10], last published in 2007, contains about 600 T_g values. The polymer classifier module, used to differentiate between all automatically extracted chemical compounds and polymers, achieved 91.8% precision and 93.2% recall. The polymer proximity search module correctly identified missing polymers for 50.0% of those T_g values without polymers. We crowdsourced the recovery of unrecognized polymer names for an additional 22 polymer–T_g pairs and demonstrated that using untrained crowds for simple, well-defined domain-specific tasks can decrease the need for expert validation by about three fourth (78.6% labels resolved by non-experts using consensus method). A basic prioritization scheme for human-review yielded encouraging results (we found 40% more errors in the first 50 ranked entries than would be expected if entries to be manually inspected were randomly selected). Our IE pipeline findings further emphasizes that the combination of domain-specific automated and crowdsourcing extraction and curation modules is a viable approach for extracting high-quality and accurate polymer–T_g pairs.

Driven by the observation in previous work that significant errors were attributed to the challenges identifying polymers and properties, we have explored machine learning-based approaches for automatically identifying polymer names in text. Here, we aim to implement semi-supervised methods that can more easily be used for scientific Named Entity Recognition (NER) problems that are often plagued with a lack of exhaustively annotated corpora required for machine learning. To circumvent the need for a large annotated corpus of polymer names, we used an ensemble of word embedding models [13,14] and domain-specific knowledge to propose candidate polymers (candidates are words that are deemed similar to a known polymer by our models). We assigned the labeling of these candidates (identifying strings that were not actually polymer names) to an expert material scientist. This task is more straightforward than reading and recognizing the entities in documents. Finally, we trained a semi-supervised named entity classifier to select actual polymer names from candidates proposed by the word embedding model. Our results are comparable (within 10% precision) of an enhanced, polymer-aware version of ChemDataExtractor. However, our approach requires only minimal human input and it does not rely on an annotated corpus. To explore the generalizability of our approach we also applied it to a completely different NER task: identifying text-based references to social science datasets. Here we achieved precision of 59.6% and recall of 58.7%, again with minimal training data. We are currently exploring an active learning methodology to improve the performance of the classifier. The goal here is to obtain human annotations of automatically labeled candidates near the decision boundary of the classifier to improve its performance.

So far, our hybrid human-machine efforts have extracted 263 values for the Flory-Huggins interaction parameter for 91 unique polymers and 302 values for Glass Transition Temperature for 243 unique polymers. Our verified extracted facts are available at both http://pppdb.uchicago.edu and https://materialsdatafacility.org. Despite significant progress in natural language processing and machine learning, there remains a gap between the current data extraction needs in fields such as materials science and the capabilities of state-of-the-art tools. Our results demonstrate the considerable potential of combining automated and crowdsourcing modules to bridge the gap between data extraction needs and the capabilities of state-of-the-art tools.

References

[1 Ware, Mark, and Michael Mabe. "The STM report: An overview of scientific and scholarly journal publishing." (2015).

[2] Friedman, Carol, et al. "A general natural-language text processor for clinical radiology." *Journal of the American Medical Informatics Association* 1.2 (1994): 161-174.

[3] Friedman, Carol, et al. "Representing information in patient reports using natural language processing and the extensible markup language." *Journal of the American Medical Informatics Association* 6.1 (1999): 76-87.

[4] Savova, Guergana K., et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." *Journal of the American Medical Informatics Association* 17.5 (2010): 507-513

[5] Friedman, Carol, et al. "GENIES: a natural-language processing system for the extraction of molecular path ways from journal articles." *ISMB (supplement of bioinformatics)*. 2001.

[6] Lussier, Y., and C. Friedman. "BiomedLEE: a natural-language processor for extracting and representing phenotypes, underlying molecular mechanisms and their relationships." *ISMB: 2007* (2007).

[7] Krauthammer, Michael, and Goran Nenadic. "Term identification in the biomedical literature." *Journal of bio medical informatics* 37.6 (2004): 512-526.

[8] Tchoua, Roselyne B., et al. "Blending education and polymer science: Semiautomated creation of a thermodynamic property database." *Journal of chemical education* 93.9 (2016): 1561-1568.

[9] Tchoua, Roselyne B., et al. "A hybrid human-computer approach to the extraction of scientific facts from the literature." *Procedia computer science* 80 (2016): 386-397.

[10] Eitouni, Hany B., and Nitash P. Balsara. "Thermodynamics of polymer blends." *Physical Properties of Polymers Handbook*. Springer, New York, NY, 2007. 339-356.

[11] Tchoua, Roselyne B., et al. "Towards a Hybrid Human-Computer Scientific Information Extraction Pipe line." *eScience (e-Science), 2017 IEEE 13th International Conference on*. IEEE, 2017.

[12] Swain, Matthew C., and Jacqueline M. Cole. "ChemDataExtractor: A toolkit for automated extraction of Chemical information from the scientific literature." *Journal of chemical information and modeling* 56.10 (2016): 1894-1904

[13] T. Mikolov, K. Chen et al., "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[14] T. Mikolov, I. Sutskever et al., "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111–3119.