Semi-Automated Construction of Food Composition Knowledge Base

Jason Youn,[*1,2,3] **Fangzhou Li,**[*1,2,3] **Ilias Tagkopoulos**[1,2,3]

[1] Department of Computer Science, University of California, Davis, CA 95616, USA.
[2]Genome Center, University of California, Davis, CA 95616, USA.
[3]USDA/NSF AI Institute for Next Generation Food Systems (AIFS), University of California, Davis, CA 95616, USA.
jyoun, fzli, itagkopoulos@ucdavis.edu

Abstract

A food composition knowledge base, which stores the essential phyto-, micro-, and macro-nutrients of foods is useful for both research and industrial applications. Although many existing knowledge bases attempt to curate such information, they are often limited by time-consuming manual curation processes. Outside of the food science domain, natural language processing methods that utilize pre-trained language models have recently shown promising results for extracting knowledge from unstructured text. In this work, we propose a semi-automated framework for constructing a knowledge base of food composition from the scientific literature available online. To this end, we utilize a pre-trained BioBERT language model in an active learning setup that allows the optimal use of limited training data. Our work demonstrates how human-in-the-loop models are a step toward AI-assisted food systems that scale well to the ever-increasing big data.

Introduction

Constructing high-quality knowledge bases of foods is crucial for various needs like understanding the impact of dietary intake on human health and enabling personalized diet recommendations (Jacobs and Tapsell 2007; Young 2003). There exist multiple knowledge bases that attempt to curate food composition information such as FoodData Central (406,999 foods and ∼300 key nutrients) (USDA 2019), FooDB (797 foods and 15,750 chemicals) (TMIC 2017), KNApSAcK (24,704 plant species and 62,647 metabolites) (Shinbo et al. 2006), and Phenol-Explorer (459 foods and 501 polyphenols) (Rothwell et al. 2013). In an effort to curate the relationship between chemicals and human health, knowledge bases like CTD (Davis et al. 2022) and KEGG (Kanehisa and Goto 2000) curate information about the interactions among chemicals, genes, and/or disease entities, while other resources, including ChemFont (Wishart et al. 2022) and GO (Ashburner et al. 2000), are dedicated to creating an ontology of chemicals. Yet, existing approaches to creating and expanding these knowledge bases are often bottlenecked by the need for time-consuming manual annotation processes that often require the expertise of domain experts.

[*]These authors contributed equally.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

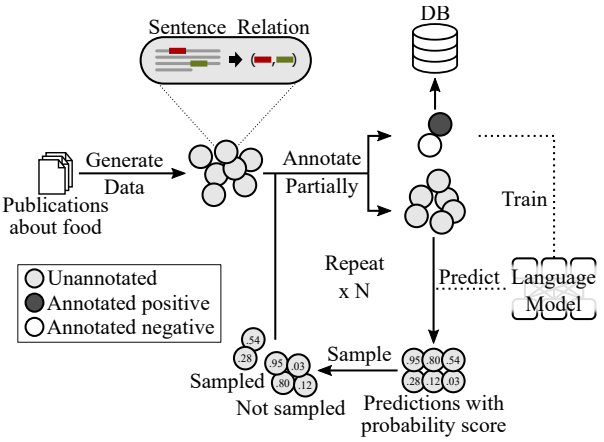

Figure 1: Overarching pipeline for semi-automated construction of a food knowledge base from online publications using language models with active learning. From the sentences in the literature that mention both foods (red) and chemicals (green), we extract relations where food and chemical entities are connected by the *contains* relation. We then use active learning with a language model to partially annotate and train over $N = 10$ rounds.

While knowledge bases have traditionally been curated manually from text data (Kotova and Pisarev 2019), recent approaches utilize deep learning-based state-of-the-art relation extraction (RE) models for constructing knowledge bases (Jiang et al. 2020). RE is a task in natural language processing (NLP) that extracts semantic relations between entities in natural language sentences (Bach and Badaskar 2007) (e.g., given a sentence 'Joe Biden is the president of United States', an RE model extracts a relation of 'isPresidentOf' between the entities 'Joe Biden' and 'United States'). However, these deep learning-based approaches often require many labeled training data (LeCun, Bengio, and Hinton 2015), which is often not feasible in fields like food science where the data annotation procedure is expensive. To address such issue, active learning methods, which use the model to choose the data that can most efficiently improve its performance and therefore reduce the amount of data needed for achieving the desired training outcome (Set-

tles 2009), have been widely used in various fields such as natural language processing (Shen et al. 2004; Longpre et al. 2022; Rotman and Reichart 2022), computer vision (Coleman et al. 2022; Chen et al. 2022; Yu et al. 2021), and studies in biology (Wang et al. 2020). Furthermore, recent works suggest that semi-automatic models are desirable for tasks related to knowledge bases as machines, though faster than human annotators, often yield low recall (Wang, Guo, and Chen 2022; Zhuang et al. 2017).

In this work, we propose a semi-automated framework for constructing a knowledge base of food composition information using active learning of language models (Figure 1). Training and evaluating 100 runs of the proposed active learning strategy each with a unique random seed shows that language models are able to extract correct relations from the sentence with high confidence (precision $= 0.92 \pm 0.04$, recall $= 0.82 \pm 0.07$, and F1 $= 0.87 \pm 0.04$). We also found that using the proposed active learning sampling strategy accelerates new knowledge discovery by $21.0\% \pm 0.05\%$. All code and instructions on how to reproduce the results can be found in https://github.com/ibpa/SemiAutomatedFoodKBC.

Proposed Method

Data Generation. We downloaded 1,226 food names (both commonly used and scientific) from FooDB (TMIC 2017) and used them to query LitSense (Allot et al. 2019), a sentence-level search system provided by National Center for Biotechnology Information (NCBI) for biomedical literature from PubMed and PubMed Central (PMC). We used the search query template *'food_name contains'* as we empirically found that the returned sentences had the most food-chemical relations. In addition to the food and chemical entities already tagged and returned by LitSense for each sentence, we manually generated a list of 70 common food parts and used a rule-based named entity recognition (NER) approach to strictly match the food part entities. Note that since LitSense returns all species in the NCBI taxonomy even if they are not food, we dropped non-food species by keeping only the species in the 1,226 FooDB food names. Therefore, for each sentence $s \in S$ returned by LitSense, three sets of entities F, P, and C for foods, food parts, and chemicals, respectively, were recognized. Finally, we extracted a set of relations R for s as

$$R = \{template(f,p,c) \mid \forall (f,p,c) \in F \times P \times C\} \cup \{template(f,c) \mid \forall (f,c) \in F \times C\}, \quad (1)$$

where $template(\cdot)$ is a function that takes as input entities, with or without food part entity, and outputs a *contains* relation between entities. For example, $template$(apple, vitamin A) = *apple contains vitamin A* and $template$(apple, skin, vitamin A) = *apple skin contains vitamin A*, if there is a food part entity. This process resulted in 85,839 sentence-relation (SR) pairs with 21,313 unique sentences.

Data Annotation. We randomly selected train, validation, and test datasets from these SR pairs, while making sure there were no duplicate sentences or relations across the datasets that could cause bias during training and evaluating (Table 1). During the manual annotation process, two annotators were asked to assign three possible classes *positive*, *negative*, and *skip*. The *positive* class was assigned when there was enough evidence in the sentence to support that the paired relation was true. Otherwise, the *negative* class was assigned. We assigned the class *skip* if the NER by LitSense was not performed correctly, where we eventually discarded the skipped SR pairs. To ensure that the annotation quality was high, we kept only the *positive* and *negative* SR pairs whose annotation results were consensus between the two annotators.

Table 1: Statistics of the dataset used for training and evaluation. Note that the training data of 1,000 sentence-relation(SR) pairs are distributed evenly across N=10 rounds of active learning, while the validation and test set is held-out for consistency.

	train	val	test
# of SR pairs	1,000	300	300
# of positive SR pairs	453	116	129
# of negative SR pairs	547	184	171
# of unique sentences	747	174	157
# of unique relations	1,000	300	300
# of entities	537	175	169
# of food entities	288	95	86
# of chemical entities	249	80	83

Language Model. We used the BioBERT (Lee et al. 2020) language model that is based on the original BERT (Devlin et al. 2018) model but trained using biomedical domain corpora PubMed and PMC, sharing a similar domain of the text returned by LitSense (Allot et al. 2019). We formatted input to the model by concatenating the sentence and relation strings separated by the [SEP] token. We fine-tuned BioBERT using the binary classification scheme, with the grid search of 8 possible hyperparameter combinations (learning rate = $\{2 \times 10^{-5}, 5 \times 10^{-5}\}$, batch size = $\{16, 32\}$, and epochs = $\{3, 4\}$) using the validation set. Finally, the hyperparameter set with the highest precision score was selected for the active learning step that we will discuss in the next section.

Active Learning. We used the pool-based active learning strategy as defined by Settles (Settles 2009), where we split the total training pool of 1,000 annotated SR pairs (Table 1) equally into 10 active learning rounds. For the first round, we randomly sampled 100 training SR pairs from the pool. We trained the language models using these 100 SR pairs and then predicted the probability of being a *positive* class, denoted as p, for the remaining 900 SR pairs. From round 2, we selected the data using the *uncertainty* sampling scheme, where we sampled the SR pairs closest to the model decision boundary. The uncertainty score for each SR pair was calculated as $min(1-p, p)$, and then the *uncertainty* scheme would sample 100 SR pairs with the highest uncertainty scores. These newly sampled data were added to the previous training data and used to train the next rounds.

Table 2: Performance metrics for 100 runs of active learning (uncertainty sampling) compared to the random sampling.

	Rounds									
Metric	1	2	3	4	5	6	7	8	9	10
Active Learning (Uncertain)										
Precision	.87±.09	.86±.05	.87±.05	.88±.05	.89±.05	.89±.04	.91±.04	.91±.04	.92±.04	.92±.04
Recall	.51±.29	.72±.13	.76±.10	.78±.10	.76±.12	.78±.09	.81±.07	.80±.08	.81±.08	.82±.07
F1	.57±.28	.77±.08	.81±.07	.82±.01	.81±.07	.83±.06	.85±.05	.85±.05	.86±.04	.87±.04
Accuracy	.75±.10	.83±.04	.85±.04	.86±.04	.85±.04	.86±.04	.88±.03	.88±.03	.89±.03	.89±.03
Specificity	.93±.07	.91±.04	.91±.04	.92±.04	.93±.04	.93±.03	.94±.03	.94±.03	.95±.03	.94±.03
Random										
Precision	.88±.07	.88±.05	.88±.04	.88±.04	.90±.04	.90±.04	.91±.04	.91±.04	.92±.04	.92±.04
Recall	.49±.28	.70±.17	.76±.10	.78±.11	.79±.09	.79±.08	.81±.07	.82±.08	.81±.07	.82±.06
F1	.56±.27	.77±.12	.81±.06	.82±.06	.84±.05	.84±.04	.86±.04	.86±.04	.86±.04	.87±.04
Accuracy	.74±.10	.83±.06	.85±.03	.86±.03	.87±.03	.87±.03	.88±.03	.89±.03	.89±.03	.89±.03
Specificity	.93±.06	.92±.05	.92±.04	.92±.04	.93±.03	.93±.03	.94±.03	.94±.03	.95±.03	.95±.03

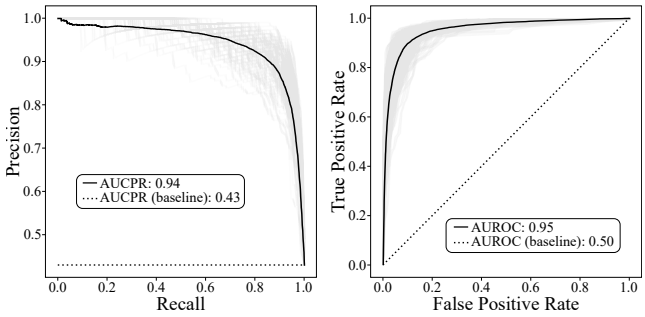

Figure 2: Precision-recall (left) and receiver operating characteristic (right) curves of the models at round 10 with active learning. Gray lines indicate the curves for 100 runs, and the black lines are plotted by concatenating the model probabilities of the test data from these 100 runs. The baseline refers to a classifier that always predicts the minority class (positive).

We repeated this process until round 10 when all training data were sampled and consumed by the model. We refer to this whole process as a single run of active learning, and we repeated this active learning run 100 times to provide statistical significance for our results.

Experiments and Results

Settings. We used 2 $\times$ Nvidia RTX A5000 GPUs for the training and inference of the BioBERT language models, one GPU for each of the active learning strategies, (i.e., uncertainty sampling and random sampling.) The entire procedure (100 runs $\times$ 10 rounds for each active learning strategy) took approximately 90 hours, where around 85% of the time was used for model training and the rest for data sampling and evaluation. The model pipeline was implemented with PyTorch (Paszke et al. 2019) and HuggingFace's transformer library (Wolf et al. 2019).

Active Learning Results. The performance metrics obtained from 100 different runs of the active learning each with 10 rounds and different random seeds are shown in Figure 2. Compared to the first round, precision at the final 10th round increased by 6.2% (0.87 ± 0.09 vs. 0.92 ± 0.04, respectively), recall by 62.0% (0.51 ± 0.29 vs. 0.82 ± 0.07, respectively), and F1 by 50.9% (0.57 ± 0.28 vs. 0.87 ± 0.04, respectively). However, compared to the random sampling active learning strategy which chooses the samples to train on for the subsequent round randomly (gray line and box plots in Figure 2), we did not observe any statistical difference between the performance metrics (p-value > 0.05) except for precision at round 2 (0.88 ± 0.05 vs. 0.86 ± 0.05, respectively, p-value $= 3.4 \times 10^{-2}$), recall at round 5 (0.79 ± 0.09 vs. 0.76 ± 0.12, respectively, p-value $= 2.9 \times 10^{-2}$), and F1 at round 5 (0.84 ± 0.05 vs. 0.81 ± 0.07, respectively, p-value $= 6.8 \times 10^{-3}$). The performance metrics start to show an insignificant difference in their average values compared to the final round at round 8 for precision (0.91 ± 0.04 vs. 0.92 ± 0.04, respectively, p-value $= 1.6 \times 10^{-1}$), round 7 for recall (0.81 ± 0.07 vs. 0.82 ± 0.07, respectively, p-value $= 2.5 \times 10^{-1}$), and round 9 for F1(0.86 ± 0.04 vs. 0.87 ± 0.04, respectively, p-value $= 3.2 \times 10^{-1}$). Note that all p-values were calculated using the two-sided t-test. The models at the final round that was trained using the complete training data have AUCPR $= 0.94$ and AUROC $= 0.95$ as shown in Figure 2.

Although both the random sampling strategy and the uncertainty active learning strategy discover the same set of 453 positives in the training pool with 1,000 SR pairs (Table 1) at the final round, the uncertainty active learning strategy discovers positive data $21.0\% \pm 0.05\%$ faster on average during the intermediate rounds (rounds 2 through 9) compared to the random sampling strategy that discovers new knowledge in each round of training linearly (Figure 3).

Conclusion

In this work, we presented a semi-automated framework for constructing a food composition knowledge base from

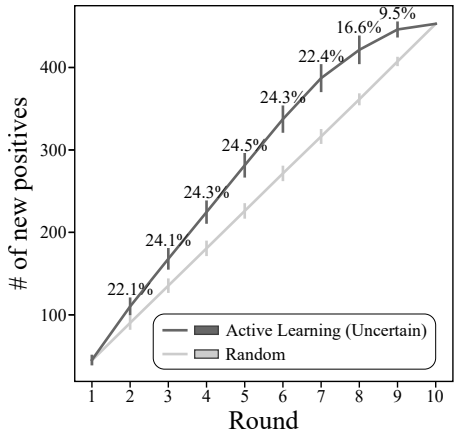

Figure 3: Rate of discovery of new knowledge within the training data compared between the uncertainty active learning sampling strategy and the random sampling strategy. Percentage value denotes the increase of relative abundance of positive data for each round from random to uncertainty active learning method.

scientific literature using pre-trained language models supported by active learning. Although the active learning sampling strategy that selects the uncertain samples around the probability of 0.5 has not shown statistically significant predictive performance improvement over the random sampling approach, we found that the uncertainty sampling approach was able to find new positive data faster, therefore leading to the creation of knowledge base in an accelerated manner than the random sampling. In future work, we plan to test additional sampling strategies like disproportionate stratified sampling and one that only samples data with the highest probability scores. We also plan to train the model with a bigger dataset, create a knowledge graph of food-chemical information enriched with ontological relationships like taxonomy and chemical classification, and perform link prediction (Yao, Mao, and Luo 2019; Youn and Tagkopoulos 2022) on the knowledge graph to discover novel food-chemical relations.

References

Allot, A.; Chen, Q.; Kim, S.; Vera Alvarez, R.; Comeau, D. C.; Wilbur, W. J.; and Lu, Z. 2019. LitSense: making sense of biomedical literature at sentence level. *Nucleic acids research*, 47(W1): W594–W599.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1): 25–29.

Bach, N.; and Badaskar, S. 2007. A review of relation extraction. *Literature review for Language and Statistics II*, 2: 1–15.

Chen, Z.; Zhang, J.; Wang, P.; Chen, J.; and Li, J. 2022. When Active Learning Meets Implicit Semantic Data Augmentation. In *European Conference on Computer Vision*, 56–72. Springer.

Coleman, C.; Chou, E.; Katz-Samuels, J.; Culatana, S.; Bailis, P.; Berg, A. C.; Nowak, R.; Sumbaly, R.; Zaharia, M.; and Yalniz, I. Z. 2022. Similarity Search for Efficient Active Learning and Search of Rare Concepts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36: 6402–6410.

Davis, A. P.; Wiegers, T. C.; Johnson, R. J.; Sciaky, D.; Wiegers, J.; and Mattingly, C. 2022. Comparative Toxicogenomics Database (CTD): update 2023. *Nucleic acids research*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacobs, D. R.; and Tapsell, L. C. 2007. Food, Not Nutrients, Is the Fundamental Unit in Nutrition. *Nutrition Reviews*, 65: 439–450.

Jiang, H.; Bao, Q.; Cheng, Q.; Yang, D.; Wang, L.; and Xiao, Y. 2020. Complex relation extraction: Challenges and opportunities. *arXiv preprint arXiv:2012.04821*.

Kanehisa, M.; and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1): 27–30.

Kotova, E. E.; and Pisarev, I. A. 2019. Automated Creation of Knowledge Bases for Intelligent Systems, taking into account Linguistic Uncertainty. In *2019 XXII International Conference on Soft Computing and Measurements (SCM))*, 149–152. IEEE.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

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

Longpre, S.; Reisler, J.; Huang, E. G.; Lu, Y.; Frank, A.; Ramesh, N.; and DuBois, C. 2022. Active Learning Over Multiple Domains in Natural Language Tasks. *arXiv preprint arXiv:2202.00254*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Rothwell, J. A.; Perez-Jimenez, J.; Neveu, V.; Medina-Remon, A.; M'hiri, N.; García-Lobato, P.; Manach, C.; Knox, C.; Eisner, R.; Wishart, D. S.; et al. 2013. Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database*, 2013.

Rotman, G.; and Reichart, R. 2022. Multi-task Active Learning for Pre-trained Transformer-based Models. *arXiv preprint arXiv:2208.05379*.

Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Shen, D.; Zhang, J.; Su, J.; Zhou, G.; and Tan, C. L. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd annual meeting of the*

Association for Computational Linguistics (ACL-04), 589–596.

Shinbo, Y.; Nakamura, Y.; Altaf-Ul-Amin, M.; Asahi, H.; Kurokawa, K.; Arita, M.; Saito, K.; Ohta, D.; Shibata, D.; and Kanaya, S. 2006. KNApSAcK: a comprehensive species-metabolite relationship database. In *Plant metabolomics*, 165–181. Springer.

TMIC. 2017. The Metabolomics Innovation Centre. FooDB Version 1.0. https://foodb.ca.

USDA. 2019. U.S. Department of Agriculture, Agricultural Research Service. FoodData Central. https://fdc.nal.usda.gov.

Wang, J.; Guo, B.; and Chen, L. 2022. Human-in-the-loop Machine Learning: A Macro-Micro Perspective. *arXiv preprint arXiv:2202.10564*.

Wang, X.; Rai, N.; Pereira, B. M. P.; Eetemadi, A.; and Tagkopoulos, I. 2020. Accelerated knowledge discovery from omics data by optimal experimental design. *Nature Communications*, 11.

Wishart, D. S.; Girod, S.; Peters, H.; Oler, E.; Jovel, J.; Budinski, Z.; Milford, R.; Lui, V. W.; Sayeeda, Z.; Mah, R.; et al. 2022. ChemFOnt: the chemical functional ontology resource. *Nucleic Acids Research*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yao, L.; Mao, C.; and Luo, Y. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Youn, J.; and Tagkopoulos, I. 2022. KGLM: Integrating Knowledge Graph Structure in Language Models for Link Prediction. *arXiv preprint arXiv:2211.02744*.

Young, V. R. 2003. Trace Element Biology: The Knowledge Base and its Application for the Nutrition of Individuals and Populations. *The Journal of Nutrition*, 133: 1581S–1587S.

Yu, W.; Zhu, S.; Yang, T.; and Chen, C. 2021. Consistency-based Active Learning for Object Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2022-June: 3950–3959.

Zhuang, Y.; Li, G.; Zhong, Z.; and Feng, J. 2017. Hike: A hybrid human-machine method for entity alignment in large-scale knowledge bases. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1917–1926.

Acknowledgments

We would like to thank the members of the Tagkopoulos lab for their suggestions, and Gabriel Simmons for the initial discussions. This work was supported by the USDA-NIFA AI Institute for Next Generation Food Systems (AIFS), USDA-NIFA award number 2020-67021-32855, and the NIEHS grant P42ES004699 to I.T. All computational analyses were performed and the figures were generated by J.Y. and F.L., and J.Y., F.L., and I.T. contributed to the critical analysis and wrote the paper. I.T. supervised all aspects of the project.