

Bootstrapping Gene Expression-Cancer Knowledge Bases with Limited Human Annotations^{*}

Stefano Marchesin¹, Laura Menotti¹, Fabio Giachelle¹, Gianmaria Silvello¹ and Omar Alonso^{2,**}

¹Department of Information Engineering, University of Padua, Padua, Italy

²Amazon, Palo Alto, California, USA

Abstract

We introduce the Collaborative Oriented Relation Extraction (CORE) system for Knowledge Base Construction, based on the combination of Relation Extraction (RE) methods and domain experts feedback. CORE features a seamless, transparent, and modular architecture that suits large-scale processing. Via active learning, the CORE system bootstraps Knowledge Bases (KBs) and then employs RE methods to scale to large text corpora. We employ CORE to build one of the largest KBs focusing on fine-grained gene expression- cancer associations, fundamental to complement and validate experimental data for precision medicine and cancer research. We conducted comprehensive experiments showing the robustness of the approach and highlighting the scalability of CORE to large text corpora with limited manual annotations.

Keywords

Knowledge Base Construction, Relation Extraction, Active Learning, Distant Supervision

1. Introduction


In 2020 there were about 19.2 million cancer cases worldwide and the World Health Organization estimates a 33% overall increase by 2040.¹ With this growing global burden, cancer prevention is one of the century's most pressing public health challenges, and data-driven research is crucial in assisting the development of medical solutions to address it. In this regard, microarray and next-generation sequencing technologies providing raw data about gene expression-cancer interactions [2, 3] are essential to guide diagnosis, assess prognosis, or predict therapy response [4]. Although these data are invaluable to the advancement of cancer research, they cannot be steadily used as is, as they require further processing and validation by experts. In most cases, the outcome of this research process is described in a scientific peer-reviewed publication. Hence, scientific literature is an authoritative data source that can be exploited to complement and validate such experimental data. However, the manual extraction of knowledge (e.g., scientific facts) from domain-specific literature is expensive and time-consuming [5, 6, 7]. In recent years, thanks to the advancement of Machine Learning (ML)

32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

^{*}Extended abstract of [1].

^{**}Work done prior to joining Amazon.

✉ stefano.marchesin@unipd.it (S. Marchesin)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://gco.iarc.fr/tomorrow/en/dataviz/bubbles?sexes=0&mode=population>

methods, automated techniques for Knowledge Base Construction (KBC) have flourished and empowered large-scale construction and curation of Knowledge Bases (KBs) [8, 9, 10]. Nevertheless, the two main components of KBC systems – i.e., Named Entity Recognition and Disambiguation (NERD) and Relation Extraction (RE) – both require expensive and often unavailable labeled data for training.

Thus, alternative solutions have been proposed to address this limitation, such as distant supervision [11, 12] and active learning [13, 14]. Distant supervision and active learning are complementary and often used together [15, 16] to bootstrap KBC systems and generate high-quality datasets for NERD and RE. Therefore, in this work, we use both paradigms to build a modular, pluggable, transparent, and scalable KBC system for cancer research that focuses on the discovery of “gene expression-cancer” associations. Specifically, we present the Collaborative Oriented Relation Extraction (CORE) system [1], a KBC system based on the combination of automated ML-based methods and domain experts feedback. CORE features a seamless, transparent and modular architecture, where the different components can be easily plugged-in. CORE also employs active learning to bootstrap a KB focusing on gene expression-cancer associations. To this end, CORE exploits the fine-grained aspects involved in gene expression-cancer associations to perform iterative tests that measure the reliability of the data to be stored in the KB and return small, selected samples to domain experts for annotation. The high-quality data generated by this process is then used as reinforcement to re-train the ML models from scratch. Active learning makes the CORE system suited to iterative KB versioning. Therefore, with the data annotated by domain experts, re-trained ML models are deployed to build subsequent versions of the KB.

To show the robustness of the proposed approach, we conducted extensive analyses that highlight how CORE scales to large text corpora with little human annotations. Moreover, to evaluate the system effectiveness against the state-of-the-art, we performed a knowledge base completion task showing that CORE achieves top performances. The KB derived by CORE storing fine-grained facts about gene expression-cancer associations is available at <https://zenodo.org/records/7577127>. The KB can also be accessed via CoreKB [17], a web search platform available at <https://gda.dei.unipd.it>.

The rest of the article is as follows: Section 2 reports on related work; Section 3 outlines the CORE system; Section 4 presents the experiments; Section 5 concludes the paper.

2. Related Work

To date, there are a handful of knowledge resources containing data about gene expression-cancer associations [18, 19, 20, 21, 22, 23]. Most of these resources only contain experimental data obtained through microarray and next-generation sequencing technologies [18, 19, 20, 21]. Whereas few of them, such as BioXpress [22] and OncoMX [23], also integrate knowledge extracted from the biomedical literature and rely on pattern matching techniques to extract relationships [24]. Thus, there is the opportunity to develop more adaptive RE methods that can broaden the reach of KBC systems to heterogeneous large-scale text corpora.

Beside resources based on experimental studies, there also exist a few literature-based resources [25, 26, 27, 28] such as CoMAGC [25] and OncoSearch [26]. They focus on gene expression-cancer associations, modeling the different, fine-grained aspects involved between gene expression and cancer. Although relevant, CoMAGC only consists of 821 sentences on prostate, breast, and ovarian cancers while OncoSearch is currently not maintained. On the other hand, more general and large-scale resources on gene-disease associations – i.e., DisGeNET [27] and LHGDN [28] – store coarse-grained information expressing the existence of an association between gene expression and cancer, which is often insufficient to model such complex, faceted relationships effectively.

Hence, there is a need for KBC systems that can scale to large text corpora and stay up to date while generating fine-grained information about gene expression-cancer associations. These fine-grained associations are essential to complement and validate experimental data, fundamental for advancing cancer research.

3. The CORE System

Preliminaries. Let us consider a directed graph $G = (V, E)$, where $E \subseteq \{(v_1, v_2) \mid (v_1, v_2) \in V \times V\}$ is the set of edges connecting ordered pairs of vertices. Given an edge $e = (v_1, v_2) \in E$, we call v_1 the source vertex and v_2 the target vertex. In our context, the nodes of G are entities and the edges are the relationships between them.

Definition 1 (Aspect). We call aspect an attribute of a relationship between a pair of entities. An aspect has a name and a domain $\text{dom} = \{a_{i1}, \dots, a_{in}\}$, where $a_{ij} \in A_i$ is the j^{th} aspect value of A_i . Given an aspect A , the function $\text{Dom}(A) = \text{dom}$ returns its domain.

When it is clear from the context, the aspect value $a_{ij} \in A_i$ is simply referred to as a_j .

Example 1. Let us consider the context of gene-cancer associations, where there are three aspects describing a possible relationship (e) between gene (v_1) and cancer (v_2): the Change of Gene Expression (CGE), the Change of Cancer Status (CCS), and the Gene-Cancer Interaction (GCI). Following Definition 1, CGE, CCS, and GCI are the names of the aspects with the following domains: $\text{Dom}(\text{CGE}) = \{\text{up}, \text{down}, \text{notinf}\}$, $\text{Dom}(\text{CCS}) = \{\text{progression}, \text{regression}, \text{notinf}\}$, and $\text{Dom}(\text{GCI}) = \{\text{causality}, \text{correlation}, \text{notinf}\}$. A detailed description of these aspect domains can be found in the original paper [1].

Definition 2 (Multi-Aspect Relationship). Given a graph $G(V, E)$ and a set of aspects $\mathcal{A} = \{A_i\}_{i=1}^n$, then a tuple of aspect values (a_{1j}, \dots, a_{nj}) associated with $e = (v_1, v_2) \in E$ defines a multi-aspect relationship between v_1 and v_2 .

Definition 3 (Signature Function). Given a set of aspects $\mathcal{A} = \{A_i\}_{i=1}^n$ and an alphabet Σ , we define $s : \prod_{i=1}^n A_i \rightarrow S \subseteq \Sigma^*$; $s((a_{1j}, \dots, a_{nj})) \mapsto \text{type}$ as the signature function that maps a multi-aspect relationship to a type from S , called the signature set.

The signature function defines a set of mapping rules depending on the domain of interest. In our setting, we refer to the mapping rules described in Table 1. That is, we use the signature function to map multi-aspect gene expression-cancer relationships

Table 1

Inference rules for gene classes. For each combination of CGE, CCS, and GCI, we report the expected gene class. Gene classes refer to the role that a given gene has on a specific disease. The * symbols in Rule 5 mean that CGE and CCS can assume any value between {up, down} and {progression, regression}.

Rule #	CGE	CCS	GCI	Gene Class
1	up	progression	causality	oncogene
2	up	regression	causality	tumor suppressor gene
3	down	regression	causality	oncogene
4	down	progression	causality	tumor suppressor gene
5	*	*	observation	biomarker

to gene prospective roles in cancer. Gene roles allow to distinguish the genes that are responsible for oncogenesis from those that are not; these are essential information for effective for cancer research and therapy design [29].

Definition 4 (Tagging Function). Given an edge $e \in E$ and the signature set S . We define $\sigma : E \rightarrow S; \sigma(e) \mapsto \text{type}$ as the function tagging an edge with a signature type.

The tagging function works on the graph and associates a signature type to an edge. Thus, we use it to label edges with gene prospective roles. In other words, the graph represents gene expression-cancer associations as gene prospective roles in cancer.

Overview. The goal of the CORE system is to harvest facts from text corpora to populate KBs. We model a KB as a directed graph G made up of entities connected by typed relationships. Facts (or statements) are (v_1, e, v_2) triples, where $v_1, v_2 \in V$, $e = (v_1, v_2) \in E$, and $\sigma(e) \in S$.

To obtain facts, CORE collects scientific literature from different sources, identifies sentences containing pairs of entities relevant to the considered task, and extracts aspects from them. Depending on the combination of extracted aspect values, a sentence expresses a specific signature type. Note that, for a given pair of entities, different sentences can express various signature types, as we show in the next example.

Example 2. Let us consider the following sentences taken from the biomedical literature:

- A. **Colorectal cancer** (CRC) growth and progression is frequently driven by RAS pathway activation through upstream growth factor receptor activation or through mutational activation of KRAS or **BRAF**.
- B. Somatic mutations of the BRAF gene, causing constitutive activation of **BRAF**, have been found in various types of human cancers such as malignant melanoma, and **colorectal cancer**.

In both sentences, the following entities are extracted $v_1 = \text{BRAF}$ and $v_2 = \text{Colorectal Cancer}$. Considering the aspects introduced in Example 1, for sentence A we find CGE = up, CCS = progression, and GCI = causality, leading to the signature type $s((\text{up}, \text{progression}, \text{causality})) = \text{oncogene}$. On the other hand, the aspect values of sentence

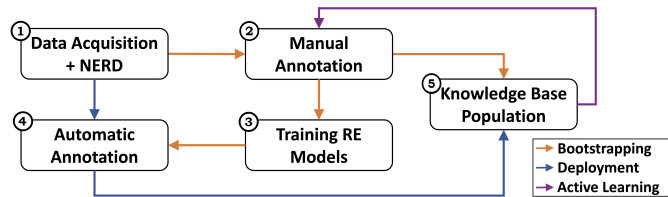


Figure 1: Overview of the CORE system architecture. The system consists of three main processes: bootstrapping (orange), deployment (blue), and active learning (purple).

B are CGE = up, CCS = progression, and GCI = correlation, leading to the signature type $s(\text{up, progression, correlation}) = \text{biomarker}$.

From Example 2, we see that different sentences may lead to different signature types. In the scientific discourse, it is not surprising that there are different viewpoints and that various studies can lead to different conclusions – even in contradiction with each other. Hence, we need to consider this potential uncertainty when facts are extracted from the literature. The CORE system models this inherent uncertainty by assigning the likelihood of being true to each aspect value. This probability is based on the evidence we can extract from the literature. Given a set of sentences concerning the same two entities, the more an aspect value is consistent in the set, the higher the probability for that value to be true. Hence, we define the concepts of Aspect-Probability Set and Multi-Aspect Function.

Definition 5 (Aspect-Probability Set). Given an aspect $A_i = \{a_j\}_{j=1}^m$ such that each aspect value a_j carries a likelihood $\Pr(a_j)$, we call $AP_i = \{(a_j, \Pr(a_j))\}_{j=1}^m$ its aspect-probability set.

Definition 6 (Multi-Aspect Function). Let $G = (V, E)$ be a directed graph and $\mathcal{AP} = \{AP_i\}_{i=1}^n$ a set of aspect-probability sets. We define $\phi : E \rightarrow \prod_{i=1}^n AP_i$; $\phi(e) \mapsto (\{(a_{1j}, \Pr(a_{1j}))\}_{j=1}^{|A_1|}, \dots, \{(a_{nj}, \Pr(a_{nj}))\}_{j=1}^{|A_n|})$ as the multi-aspect function that, given an edge, returns the n -tuple of aspect-probability sets.

Thus, for each pair of target entities, CORE computes the probabilities for all the aspect values and combines them into tuples of aspect-probability sets – which represent a probability distribution over multi-aspect relationships. In this way, sentences serve as supporting or contradicting evidence that strengthens or weakens the likelihood of a fact. Furthermore, aspect-probability sets drive another essential aspect of CORE: the data-driven, active learning approach used to bootstrap KBs. That is, through reliability tests based on aspect value likelihoods and inference rules, the system tags facts as reliable or unreliable. Part of the sentences associated with the most “highly” unreliable facts is then fed to a human-in-the-loop process that reinforces the RE methods for aspect extraction.

Architecture. Figure 1 presents the system architecture. In the first module (module 1), the texts acquired from the literature are processed and normalized to obtain sentences, from which a NERD component extracts entity pairs. The entity-annotated sentences

undergo two different processes: bootstrapping (orange workflow) and deployment (blue workflow). In the bootstrapping workflow, experts manually annotate multi-aspect relationships between the entities (module 2), producing a set of relation-annotated sentences.

The manual, relation-annotated sentences are then used to train RE methods (module 3) and to populate the KB (module 5). The RE methods are trained to predict the different aspects of multi-aspect relationships. Once trained, RE methods are employed in the deployment workflow to obtain automatic annotations expressing multi-aspect relationships between entities (module 4). Then, automatic, relation-annotated sentences are used to further populate the KB (module 5).

In the last module (module 5), relation-annotated sentences are grouped by entity pairs and used to generate facts. First, a knowledge enrichment component computes probabilities for all the aspect values and combines them into tuples of aspect-probability sets. Then, a reliability testing component uses these probabilities to perform multiple tests that tag facts as either reliable or unreliable. Only facts tagged as reliable are used to populate the KB.

When the deployment workflow is complete, unreliable facts are ranked by ascending reliability score and the top- k automatically annotated sentences associated with them are re-annotated by experts – thus triggering an active learning process that reinforces the RE methods (purple workflow).

Versioning. The active learning workflow makes CORE suited to iterative KB versioning. We define a KB version as the graph $G_j = (V_j, E_j)$ obtained after the j^{th} iteration of the bootstrap and deployment workflows. Once the j^{th} version of the KB has been deployed, the active learning workflow starts by generating the batch of unreliable sentences for bootstrapping the $j^{\text{th}} + 1$ version of the KB. The unreliable sentences are manually annotated and used to increase the size of the datasets to re-train the RE methods from scratch, which then generate a new set of automatic annotations to be included in the $j^{\text{th}} + 1$ KB version. When the bootstrap and deployment workflows end, the $j^{\text{th}} + 1$ version of the KB is re-built from scratch and comprises all the available annotations.

4. Implementation and Experiments

Knowledge Base Construction. We use different resources to build the KB, which increase with each subsequent iteration of the KB construction process. The considered resources are CoMAGC [25], OncoSearch [26], BioXpress [22], DisGeNET [27], and PubMed.² For CoMAGC, BioXpress, and OncoSearch (KBs 0–3) we revised the available manual annotations to make them compliant with our annotation schema; for DisGeNET (KBs 1–3) we divided its data into two batches to test versioning; and for PubMed (KB3) we only considered the articles citing those stored within KB2. Table 2 reports statistics for the resources used to build each KB version, while Table 3 reports the statistics about each version of the generated KB.

²<https://pubmed.ncbi.nlm.nih.gov>

Table 2

Raw statistics for the KB versions. Rows represent the raw instances considered to build the KB.

		KB0	KB1	KB2	KB3
Manual	CoMAGC (revised)	821	821	821	821
	OncoSearch (revised)	157	157	157	157
	BioXpress (revised)	74	74	74	74
	DisGeNET (batch 1)	-	-	250	250
	DisGeNET (batch 2)	-	-	-	249
Automatic	DisGeNET (batch 1)	-	184,859	184,609	184,609
	DisGeNET (batch 2)	-	-	184,858	184,609
	PubMed (citing papers)	-	-	-	2,841,096
Total		1,052	185,911	370,769	3,211,865

Table 3

Partition and general statistics for each KB version.

		KB0	KB1	KB2	KB3
Partition	Manual	655	585	605	592
	Automatic	-	96,531	95,282	435,283
General	Sentence	655	97,116	95,887	435,875
	Article	411	69,462	65,236	161,449
	Gene	329	9,483	9,981	21,005
	Cancer	98	1,479	1,554	1,665
	Fact	512	71,554	89,999	153,016

First, we can see that the ratio between the sentences stored in the KB (Table 3) and the input ones (Table 2) decreases at each iteration. From the first iteration, where the CORE system uses the 62% of the input sentences to build KB0, we move to the 52% to build KB1, 26% for KB2, and only 14% for KB3. Such a decrease reflects the use of reliability tests and active learning, which make the system more selective and accurate. In particular, active learning leads the CORE system to refine the RE methods at each iteration, thus reducing false positives as well as unreliable facts. Secondly, the large number of different genes and cancers in KB3 highlights the scalability of the approach. In this regard, KB3 contains 21,005 genes, which cover the 70% of the 30,000 estimated genes in the human genome.³ On the other hand, through the integration of DisGeNET data, KBs 1–3 contain most of the (known) cancer types involved in gene expression-cancer associations. Combined, this large number of genes and cancer types leads to more than 150,000 reliable facts. Finally, compared to currently available knowledge resources [22, 23, 26], KB3 represents the largest literature-derived KBs with reliable fine-grained facts about gene expression-cancer associations.

Knowledge Base Completion. We evaluate the effectiveness of the CORE system on a KB completion task, in which we hold out a portion of an existing KB with associated sentences and we assess CORE ability to recover it. To this end, we hold out from

³<https://www.genome.gov/human-genome-project/>

Table 4

CORE system performances on the BioXpress completion task after each (re-)training of the RE methods. We also report DEXTER performance on KB3.

Dataset	Method	Accuracy	Precision	Recall	F1
BioXpress	CORE0	0.9544	0.9601	0.9544	0.9572
	CORE1	0.9703	0.9831	0.9703	0.9766
	CORE2	0.9706	0.9827	0.9706	0.9766
KB3	DEXTER	0.3256	0.6034	0.3256	0.2882

BioXpress [22] the set of 9,636 sentences annotated by DEXTER [24] – a state-of-the-art text-mining system for gene expression-cancer associations based on pattern matching – and we evaluate the CORE system on them. Note that such sentences are not part of those used to train the CORE RE methods. Vice versa, we apply DEXTER on the manually annotated subset of KB3 to evaluate its ability to generalize to heterogeneous sentences, whose syntactic structure can differ from its predefined patterns.

For BioXpress completion, we use the three versions of the CORE system obtained after each (re-)training of the RE methods. Table 4 reports the CORE system performances on the BioXpress completion task after each (re-)training of the RE methods, as well as DEXTER performance on KB3.

We can see that each CORE version consistently achieves performances above 0.95 for each measure. In particular, CORE1 improves over CORE0 by about 2% and reaches a performance plateau, as shown by CORE2 performance. The results highlight the effectiveness of the CORE system in recovering BioXpress using a limited amount of manual annotations to train the RE methods. On the other hand, the poor performance of DEXTER on KB3 highlights a lack of flexibility that hampers its applicability to heterogeneous sentences. To further support this intuition, we observe that between precision and recall it is recall to have the worst performance, with a value of 0.3256. This underlines the expert system nature of DEXTER which, although precise, fails to generalize beyond its set of predefined patterns.

5. Conclusions

In this work we presented CORE, a KBC system based on the combination between automated RE methods and domain experts. The reliability tests and the active learning process make the system suited to iterative KB versioning. We used the CORE to build one of the largest KBs about gene expression-cancer associations. We conducted extensive experiments that (i) highlighted the ability of CORE to scale to large collections of heterogeneous data with limited human annotations and (ii) showed its generalizability and reliability compared to the current state-of-the-art.

Acknowledgments. This work is partially supported by the HEREDITARY Project, as part of the European Union’s Horizon Europe research and innovation programme under grant agreement No GA 101137074.

References

- [1] S. Marchesin, L. Menotti, F. Giachelle, G. Silvello, O. Alonso, Building a Large Gene Expression-Cancer Knowledge Base with Limited Human Annotations, *Database J. Biol. Databases Curation* 2023 (2023). URL: <https://doi.org/10.1093/database/baad061>. doi:10.1093/DATABASE/BAAD061.
- [2] C. Manzoni, D. A. Kia, J. Vandrovцова, J. Hardy, N. W. Wood, P. A. Lewis, R. Ferrari, Genome, Transcriptome and Proteome: the Rise of Omics Data and Their Integration in Biomedical Sciences, *Briefings in Bioinformatics* 19 (2016) 286–302.
- [3] P. Borry, H. B. Bentzen, I. Budin-Ljøsne, M. C. Cornel, H. C. Howard, O. Feeney, L. Jackson, D. Mascalonzi, Á. Mendes, B. Peterlin, B. Riso, M. Shabani, H. Skirton, S. Sterckx, D. Vears, M. Wjst, H. Felzmann, The Challenges of the Expanded Availability of Genomic Information: an Agenda-Setting Paper, *J. Community Genet.* 9 (2018) 103–116.
- [4] B. Neary, J. Zhou, P. Qiu, Identifying Gene Expression Patterns Associated with Drug-Specific Survival in Cancer Patients, *Scientific Reports* 11 (2021) 1–12.
- [5] F. Liu, J. Chen, A. Jagannatha, H. Yu, Learning for Biomedical Information Extraction: Methodological Review of Recent Advances, *CoRR* abs/1606.07993 (2016).
- [6] M. Krallinger, O. Rabal, S. A. Akhondi, M. P. Pérez, J. Santamaría, G. P. Rodríguez, G. Tsatsaronis, A. Intxaurreondo, J. A. Lopez, U. K. Nandal, E. M. van Buel, A. Chandrasekhar, M. Rodenburg, A. Lægneid, M. A. Doornenbal, J. Oyarzábal, A. Lourenço, A. Valencia, Overview of the BioCreative VI chemical-protein interaction Track, in: *Proc. of the sixth BioCreative challenge evaluation workshop*, 2017.
- [7] A. Miranda, F. Mehryary, J. Luoma, S. Pyysalo, A. Valencia, M. Krallinger, Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations, in: *Proc. of the seventh BioCreative challenge evaluation workshop*, 2021.
- [8] G. Weikum, X. L. Dong, S. Razniewski, F. M. Suchanek, Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases, *Found. Trends Databases* 10 (2021) 108–490.
- [9] D. Wright, A. L. Gentile, N. Faux, K. L. Beck, BioAct: Biomedical Knowledge Base Construction using Active Learning, *bioRxiv* (2022).
- [10] P. Ernst, A. Siu, G. Weikum, HighLife: Higher-arity Fact Harvesting, in: *Proc. of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, ACM, 2018*, pp. 1013–1022.
- [11] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: *Proc. of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009) and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, ACL, 2009*, pp. 1003–1011.
- [12] M. Surdeanu, J. Tibshirani, R. Nallapati, C. D. Manning, Multi-instance Multi-

label Learning for Relation Extraction, in: Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea, ACL, 2012, pp. 455–465.

- [13] B. Settles, Active Learning Literature Survey, *Science* 10 (1995) 237–304.
- [14] F. Olsson, A Literature Survey of Active Machine Learning in the Context of Natural Language Processing, SICS Technical Report (2009).
- [15] G. Angeli, J. Tibshirani, J. Wu, C. D. Manning, Combining Distant and Partial Supervision for Relation Extraction, in: Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, ACL, 2014, pp. 1556–1567.
- [16] L. Sterckx, T. Demeester, J. Deleu, C. Develder, Using Active Learning and Semantic Clustering for Noise Reduction in Distant Supervision, in: Proc. of the 4th Workshop on Automated Base Construction at NIPS 2014 (AKBC-2014), 2014, pp. 1–6.
- [17] F. Giachelle, S. Marchesin, G. Silvello, O. Alonso, Searching for Reliable Facts over a Medical Knowledge Base, in: Proc. of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023.
- [18] S. J. Park, B. H. Yoon, S. K. Kim, S. Y. Kim, GENT2: an updated gene expression database for normal and tumor tissues, *BMC Medical Genom.* 12 (2019) 1–8.
- [19] Y. D. Shaul, B. Yuan, P. Thiru, A. Nutter-Upham, S. McCallum, C. Lanzkron, G. W. Bell, D. M. Sabatini, MERAV: a tool for comparing gene expression across human tissues and cell types, *Nucleic Acids Res.* 44 (2016) 560–566.
- [20] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, A. Kasprzyk, International Cancer Genome Consortium Data Portal - a one-stop shop for cancer genomics data, *Database J. Biol. Databases Curation* 2011 (2011).
- [21] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, The Cancer Genome Atlas Pan-Cancer Analysis Project, *Nat. Genet.* 45 (2013) 1113–1120.
- [22] H. Dingerdissen, J. Torcivia-Rodriguez, Y. Hu, T. C. Chang, R. Mazumder, R. Y. Kahsay, BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery, *Nucleic Acids Res.* 46 (2018) D1128–D1136.
- [23] H. M. Dingerdissen, F. Bastian, K. Vijay-Shanker, M. Robinson-Rechavi, A. Bell, N. Gogate, S. Gupta, E. Holmes, R. Kahsay, J. Keeney, H. Kincaid, C. H. King, D. Liu, D. J. Crichton, R. Mazumder, OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data, *JCO Clin. Cancer Inform.* (2020) 210–220.
- [24] S. Gupta, H. Dingerdissen, K. E. Ross, Y. Hu, C. H. Wu, R. Mazumder, K. Vijay-Shanker, DEXTER: disease-expression relation extraction from text, *Database J. Biol. Databases Curation* 2018 (2018) bay045.
- [25] H. J. Lee, S. H. Shim, M. R. Song, H. Lee, J. C. Park, CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations, *BMC Bioinform.* 14 (2013) 323.
- [26] H. J. Lee, T. C. Dang, H. Lee, J. C. Park, OncoSearch: cancer gene search engine

- with literature evidence, *Nucleic Acids Res.* 42 (2014) 416–421.
- [27] J. P. González, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L. I. Furlong, The DisGeNET knowledge platform for disease genomics: 2019 update, *Nucleic Acids Res.* 48 (2020) D845–D855.
- [28] M. Bundschuh, A. Bauer-Mehren, V. Tresp, L. I. Furlong, H. P. Kriegel, Digging for knowledge with information extraction: a case study on human gene-disease associations, in: *Proc. of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010, ACM, 2010*, pp. 1845–1848.
- [29] D. Haber, J. Settleman, Cancer: Drivers and passengers, *Nature* 446 (2007) 145–146.