Exploiting Curated Databases to Train Relation Extraction Models for Gene-Disease Associations*

(Discussion Paper)

Stefano Marchesin[a], Gianmaria Silvello[a]

[a]*Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Padova, Via Gradenigo 6/b, 35131, Padova, Italy*

Abstract

Databases are pivotal to advancing biomedical science. Nevertheless, most of them are populated and updated by human experts with a great deal of effort. Biomedical Relation Extraction (BioRE) aims to shift these expensive and time-consuming processes to machines. Among its different applications, the discovery of Gene-Disease Associations (GDAs) is one of the most pressing challenges. Despite this, few resources have been devoted to training – and evaluating – models for GDA extraction. Besides, such resources are limited in size, preventing models from scaling effectively to large amounts of data. To overcome this limitation, we have exploited the DisGeNET database to build a large-scale, semi-automatically annotated dataset for GDA extraction: TBGA. TBGA is generated from more than 700K publications and consists of over 200K instances and 100K gene-disease pairs. We have evaluated state-of-the-art models for GDA extraction on TBGA, showing that it is a challenging dataset for the task. The dataset and models are publicly available to foster the development of state-of-the-art BioRE models for GDA extraction.

Keywords

Weak Supervision, Relation Extraction, Gene-Disease Association

1. Introduction

Curated databases, such as UniProt [2], DrugBank [3], or CTD [4], are pivotal to the development of biomedical science. Such databases are usually populated and updated with a great deal of effort by human experts [5], thus slowing down the biological knowledge discovery process. To overcome this limitation, the Biomedical Information Extraction (BioIE) field aims to shift population and curation processes to machines by developing effective computational tools that automatically extract meaningful facts from the vast unstructured scientific literature [6, 7]. Once extracted, machine-readable facts can be fed to downstream tasks to ease biological knowledge discovery. Among the various tasks, the discovery of Gene-Disease Associations (GDAs) is one of the most pressing challenges to advance precision medicine and drug discovery [8], as it helps to understand the genetic causes of diseases [9]. Thus, the automatic extraction and

* The full paper has been originally published in BMC Bioinformatics [1]

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems (SEBD 2022), June 19–22, 2022, Pisa, Italy

stefano.marchesin@unipd.it (S. Marchesin); gianmaria.silvello@unipd.it (G. Silvello)

0000-0003-0362-5893 (S. Marchesin); 0000-0003-4970-4554 (G. Silvello)

CEUR Workshop Proceedings (CEUR-WS.org)

curation of GDAs is key to advance precision medicine research and provide knowledge to assist disease diagnostics, drug discovery, and therapeutic decision-making.

Most datasets used to train and evaluate Relation Extraction (RE) models for GDA extraction are hand-labeled corpora [10, 11, 12]. However, hand-labeling data is an expensive process requiring large amounts of time to expert biologists and, therefore, all of these datasets are limited in size. To address this limitation, distant supervision has been proposed [13]. Under the distant supervision paradigm, all the sentences mentioning the same pair of entities are labeled by the corresponding relation stored within a source database. The assumption is that if two entities participate in a relation, at least one sentence mentioning them conveys that relation. As a consequence, distant supervision generates a large number of false positives, since not all sentences express the relation between the considered entities. To counter false positives, the RE task under distant supervision can be modeled as a Multi-Instance Learning (MIL) problem [14, 15, 16, 17]. With MIL, the sentences containing two entities connected by a given relation are collected into bags labeled with such relation. Grouping sentences into bags reduces noise, as a bag of sentences is more likely to express a relation than a single sentence. Thus, distant supervision alleviates manual annotation efforts, and MIL increases the robustness of RE models to noise.

Since the advent of distant supervision, several datasets for RE have been developed under this paradigm for news and biomedical science domains [13, 18, 7]. Among biomedical ones, the most relevant datasets are BioRel [18], a large-scale dataset for domain-general Biomedical Relation Extraction (BioRE), and DTI [7], a large-scale dataset developed to extract Drug-Target Interactions (DTIs). In the wake of such efforts, we created TBGA: a novel large-scale, semi-automatically annotated dataset for GDA extraction based on DisGeNET. We chose DisGeNET as source database since it is one of the most comprehensive databases for GDAs [19], integrating several expert-curated resources.

Then, we trained and tested several state-of-the-art RE models on TBGA to create a large and realistic benchmark for GDA extraction. We built models using OpenNRE [20], an open and extensible toolkit for Neural Relation Extraction (NRE). The choice of OpenNRE eases the re-use of the dataset and the models developed for this work to future researchers. Finally, we publicly released TBGA on Zenodo,[1] whereas we stored source code and scripts to train and test RE models in a publicly available GitHub repository.[2]

2. Dataset

TBGA is the first large-scale, semi-automatically annotated dataset for GDA extraction. The dataset consists of three text files, corresponding to train, validation, and test sets, plus an additional JSON file containing the mapping between relation names and IDs. Each record in train, validation, or test files corresponds to a single GDA extracted from a sentence, and it is represented as a JSON object with the following attributes:

- `text`: sentence from which the GDA was extracted.

[1] https://doi.org/10.5281/zenodo.5911097

[2] https://github.com/GDAMining/gda-extraction/

- relation: relation name associated with the given GDA.
- h: JSON object representing the gene entity, composed of:
  - id: NCBI Entrez ID associated with the gene entity.
  - name: NCBI official gene symbol associated with the gene entity.
  - pos: list consisting of starting position and length of the gene mention within text.
- t: JSON object representing the disease entity, composed of:
  - id: UMLS Concept Unique Identifier (CUI) associated with the disease entity.
  - name: UMLS preferred term associated with the disease entity.
  - pos: list consisting of starting position and length of the disease mention within text.

If a sentence contains multiple gene-disease pairs, the corresponding GDAs are split into separate data records.

Overall, TBGA contains over 200,000 instances and 100,000 bags. Table 1 reports per-relation statistics for the dataset. Notice the large number of Not Associated (NA) instances. Regarding gene and disease statistics, the most frequent genes are tumor suppressor genes, such as TP53 and CDKN2A, and (proto-)oncogenes, like EGFR and BRAF. Among the most frequent diseases, we have neoplasms such as breast carcinoma, lung adenocarcinoma, and prostate carcinoma. As a consequence, the most frequent GDAs are gene-cancer associations.

Table 1
Per-relation statistics for TBGA. Statistics are reported separately for each data split.

Granularity	Split	Therapeutic	Biomarker	Genomic Alterations	NA
Sentence-level	Train	3,139	20,145	32,831	122,149
	Validation	402	2,279	2,306	15,206
	Test	384	2,315	2,209	15,608
Bag-level	Train	2,218	13,372	12,759	56,698
	Validation	331	2,019	1,147	6,994
	Test	308	2,068	1,122	6,996

3. Experimental Setup

3.1. Datasets

We performed experiments on three different datasets: TBGA, DTI, and BioRel. We used TBGA as a benchmark to evaluate RE models for GDA extraction under the MIL setting. On the other hand, we used DTI and BioRel only to validate the soundness of our implementation of the baseline models.

3.2. Evaluation Measures

We evaluated RE models using the Area Under the Precision-Recall Curve (AUPRC). AUPRC is a popular measure to evaluate distantly-supervised RE models, which has been adopted by OpenNRE [20] and used in several works, such as [7, 18]. For experiments on TBGA, we also computed Precision at k items (P@k).

3.3. Aggregation Strategies

We adopted two different sentence aggregation strategies to use RE models under the MIL setting: average-based (AVE) and attention-based (ATT) [21]. The average-based aggregation assumes that all sentences within the same bag contribute equally to the bag-level representation. In other words, the bag representation is the average of all its sentence representations. On the other hand, the attention-based aggregation represents each bag as a weighted sum of its sentence representations, where the attention weights are dynamically adjusted for each sentence.

3.4. Relation Extraction Models

We considered the main state-of-the-art RE models to perform experiments: CNN [22], PCNN [23], BiGRU [24, 18, 7], BiGRU-ATT [25, 7], and BERE [7]. All models use pre-trained word embeddings to initialize word representations. On the other hand, Position Features (PFs), Position Indicators (PIs), and unknown words are initialized using the normal distribution, whereas blank words are initialized with zeros.

We adopted pre-trained BioWordVec [26] embeddings to perform experiments on TBGA. Two versions of pre-trained BioWordVec embeddings are available: "Bio_embedding_intrinsic" and "Bio_embedding_extrinsic". We chose the "Bio_embedding_extrinsic" version as it is the most suitable for BioRE. As for the experiments on DTI and BioRel, we adopted the pre-trained word embeddings used in the original works [7, 18] – that is, the word embeddings from Pyysalo et al. [27] for DTI, and the "Bio_embedding_extrinsic" version of BioWordVec for BioRel.

For TBGA experiments, we used grid search to determine the best combination between optimizer and learning rate. As combinations, we tested Stochastic Gradient Descent (SGD) with learning rate among {0.1, 0.2, 0.3, 0.4, 0.5} and Adam [28] with learning rate set to 0.0001. For all RE models, we set the rest of the hyper-parameters empirically.

For DTI and BioRel experiments, we relied on the hyper-parameter settings reported in the original works [7, 18].

4. Experimental Results

We report the results for two different experiments. The first experiment aims to validate the soundness of the implementation of the considered RE models. To this end, we trained and tested the RE models on DTI and BioRel datasets, and we compared the AUPRC scores we obtained against those reported in the original works [7, 18]. For this experiment, we only compared the RE models and aggregation strategies that were used in the original works. The

Table 2
Results of the baselines validation on DTI [7] and BioRel [18] datasets. The "–" symbol means that the RE model, for the given aggregation strategy, has not been originally evaluated on the specific dataset.

Model	Strategy	Implementation	DTI	BioRel
CNN	AVE	Reproduced	–	0.800
		Original	–	0.790
	ATT	Reproduced	–	0.790
		Original	–	0.780
PCNN	AVE	Reproduced	0.234	0.860
		Original	0.160	0.820
	ATT	Reproduced	0.408	0.820
		Original	0.359	0.790
BiGRU	AVE	Reproduced	–	0.870
		Original	–	0.800
	ATT	Reproduced	0.379	0.850
		Original	0.390	0.780
BiGRU-ATT	ATT	Reproduced	0.383	–
		Original	0.457	–
BERE	AVE	Reproduced	0.407	–
		Original	0.384	–
	ATT	Reproduced	0.525	–
		Original	0.524	–

second experiment uses TBGA as a benchmark to evaluate RE models for GDA extraction. In this case, we trained and tested all the considered RE models using both aggregation strategies. For each RE model, we reported the AUPRC and P@k scores.

4.1. Baselines Validation

The results of the baselines validation are reported in Table 2. We can observe that the RE models we use from – or implement within – OpenNRE achieve performance higher than or comparable to those reported in DTI and BioRel original works. The only exceptions are BiGRU and BiGRU-ATT on DTI, where the AUPRC scores of our implementations are lower than those reported in the original work. However, Hong et al. [7] report the optimal hyper-parameter settings for BERE, but not for the baselines. Thus, we attribute the negative difference between our implementations and theirs to the lack of information about optimal hyper-parameters. Overall, the results confirm the soundness of our implementations. Therefore, we can consider them as competitive baseline models to use for benchmarking GDA extraction.

4.2. GDA Benchmarking

Table 3 reports the AUPRC and P@k scores of RE models on TBGA. Given the RE models performance, we make the following observations. First, the AUPRC performances achieved by

Table 3
RE models performance on TBGA dataset. For each measure, **bold** values represent the best scores.

Model	Strategy	AUPRC	P@50	P@100	P@250	P@500	P@1000
CNN	AVE	0.422	**0.780**	0.760	0.744	0.696	0.625
	ATT	0.403	**0.780**	0.760	0.788	0.710	0.624
PCNN	AVE	0.426	**0.780**	**0.780**	0.744	0.720	0.664
	ATT	0.404	0.760	0.750	0.744	0.700	0.628
BiGRU	AVE	0.437	0.620	0.720	0.724	0.730	0.678
	ATT	0.423	0.760	0.750	0.748	0.726	0.666
BiGRU-ATT	AVE	0.419	0.740	0.740	0.748	0.694	0.615
	ATT	0.390	0.680	0.760	0.756	0.702	0.631
BERE	AVE	0.419	0.700	0.710	0.720	0.704	0.620
	ATT	**0.445**	**0.780**	**0.780**	**0.800**	**0.764**	**0.709**

RE models on TBGA indicate a high complexity of the GDA extraction task. The task complexity is further supported by the lower performances obtained by top-performing RE models on TBGA compared to DTI and BioRel (cf. Table 2). Secondly, CNN, PCNN, BiGRU, and BiGRU-ATT RE models behave similarly. Among them, BiGRU-ATT has the worst performance. This suggests that replacing BiGRU max pooling layer with an attention layer proves less effective. Overall, the best AUPRC and P@k scores are achieved by BERE when using the attention-based aggregation strategy. This highlights the effectiveness of fully exploiting sentence information from both semantic and syntactic aspects [7]. Thirdly, in terms of AUPRC, the attention-based aggregation proves less effective than the average-based one. On the other hand, attention-based aggregation provides mixed results on P@k measures. Although in contrast with the results obtained in general-domain RE [21], this trend is in line with the results found by Xing et al. [18] on BioRel, where RE models using an average-based aggregation strategy achieve performance comparable to or higher than those using an attention-based one. The only exception is BERE, whose performance using the attention-based aggregation outperforms the one using the average-based strategy. Thus, the obtained results suggest that TBGA is a challenging dataset for GDA extraction.

5. Conclusions

We have created TBGA, a large-scale, semi-automatically annotated dataset for GDA extraction. Automatic GDA extraction is one of the most relevant tasks of BioRE. We have used TBGA as a benchmark to evaluate state-of-the-art BioRE models on GDA extraction. The results suggest that TBGA is a challenging dataset for this task and, in general, for BioRE.

Acknowledgments

The work was supported by the EU H2020 ExaMode project, under Grant Agreement no. 825292.

References

[1] S. Marchesin, G. Silvello, TBGA: A Large-Scale Gene-Disease Association Dataset for Biomedical Relation Extraction, BMC Bioinform. (in print) (2022) 1–26.

[2] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL, Nucleic Acids Res. 25 (1997) 31–36.

[3] D. S. Wishart, C. Knox, A. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for *in silico* drug discovery and exploration, Nucleic Acids Res. 34 (2006) 668–672.

[4] C. J. Mattingly, G. T. Colby, J. N. Forrest, J. L. Boyer, The Comparative Toxicogenomics Database (CTD), Environ. Health Perspect. 111 (2003) 793–795.

[5] P. Buneman, J. Cheney, W. C. Tan, S. Vansummeren, Curated Databases, in: Proc. of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2008, June 9-11, 2008, Vancouver, BC, Canada, ACM, 2008, pp. 1–12.

[6] S. Wang, J. Ma, M. K. Yu, F. Zheng, E. W. Huang, J. Han, J. Peng, T. Ideker, Annotating gene sets by mining large literature collections with protein networks, in: Biocomputing 2018: Proc. of the Pacific Symposium, The Big Island of Hawaii, Hawaii, USA, January 3-7, 2018, 2018, pp. 601–613.

[7] L. Hong, J. Lin, S. Li, F. Wan, H. Yang, T. Jiang, D. Zhao, J. Zeng, A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories, Nat. Mach. Intell. 2 (2020) 347–355.

[8] S. Dugger, A. Platt, D. Goldstein, Drug development in the era of precision medicine, Nat. Rev. Drug. Discov. 17 (2018) 183–196.

[9] 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.

[10] E. M. van Mulligen, A. Fourrier-Réglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifirò, J. A. Kors, L. I. Furlong, The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships, J. Biomed. Informatics 45 (2012) 879–884.

[11] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, D. S. Wishart, PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites, Nucleic Acids Res. 36 (2008) 399–405.

[12] 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.

[13] 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.

[14] T. G. Dietterich, R. H. Lathrop, T. Lozano-Pérez, Solving the Multiple Instance Problem with Axis-Parallel Rectangles, Artif. Intell. 89 (1997) 31–71.

[15] S. Riedel, L. Yao, A. McCallum, Modeling Relations and Their Mentions without Labeled Text, in: Proc. of Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, volume 6323 of *LNCS*, Springer, 2010, pp. 148–163.

[16] R. Hoffmann, C. Zhang, X. Ling, L. S. Zettlemoyer, D. S. Weld, Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations, in: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, ACL, 2011, pp. 541–550.

[17] 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.

[18] R. Xing, J. Luo, T. Song, BioRel: towards large-scale biomedical relation extraction, BMC Bioinform. 21-S (2020) 543.

[19] Z. Tanoli, U. Seemab, A. Scherer, K. Wennerberg, J. Tang, M. Vähä-Koskela, Exploration of databases and methods supporting drug repurposing: a comprehensive survey, Briefings Bioinform. 22 (2021) 1656–1678.

[20] X. Han, T. Gao, Y. Yao, D. Ye, Z. Liu, M. Sun, OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction, in: Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, ACL, 2019, pp. 169–174.

[21] Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, Neural Relation Extraction with Selective Attention over Instances, in: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, ACL, 2016, pp. 2124–2133.

[22] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation Classification via Convolutional Deep Neural Network, in: Proc. of COLING 2014, 25th International Conference on Computational Linguistics, Technical Papers, August 23-29, 2014, Dublin, Ireland, ACL, 2014, pp. 2335–2344.

[23] D. Zeng, K. Liu, Y. Chen, J. Zhao, Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks, in: Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, ACL, 2015, pp. 1753–1762.

[24] D. Zhang, D. Wang, Relation Classification via Recurrent Neural Network, CoRR abs/1508.01006 (2015).

[25] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification, in: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers, ACL, 2016, pp. 207–212.

[26] Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, BioWordVec, improving biomedical word embeddings with subword information and MeSH, Sci Data 6 (2019) 1–9.

[27] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, S. Ananiadou, Distributional Semantics Resources for Biomedical Text Processing, Proc. of LBM (2013) 39–44.

[28] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: Proc. of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, 2015, pp. 1–15.