Knowledge Triplets Derivation from Scientific Publications via Dual-Graph Resonance

Kai Zhang[1], Pengcheng Li[4], Kaisong Song[2], Yangyang Kang[2],Xurui Li[2],
Xuhong Zhang[3], Xiaozhong Liu[1]

[1]Computer Science, Worcester Polytechnic Institute
[2]Damo Academy, Alibaba Group
[3]Computer Science, Indiana University
[4]Information Science, Hubei University of Technology

LREC-COLING 2024

Agenda

- Motivation
- Challenges
- Methodology
- Results and Discussion
- Conclusion

Thanks for the music accompaniment!

Motivation

- Existing dilemmas
  - A tremendous number
  - Increasing annually
  - Laborious for researchers

Fiscal Year	Articles Available	Avg. Unique Sessions / Weekday
FY23	9,407,149	3.8 million
FY22	8,386,512	3.2 million
FY21	7,383,455	3.1 million
FY20	6,471,176	3.4 million
FY19	5,725,819	2.6 million
FY18	5,107,590	2.4 million

Can NLP-powered machines help alleviate the laborious task for researchers?

Existing work

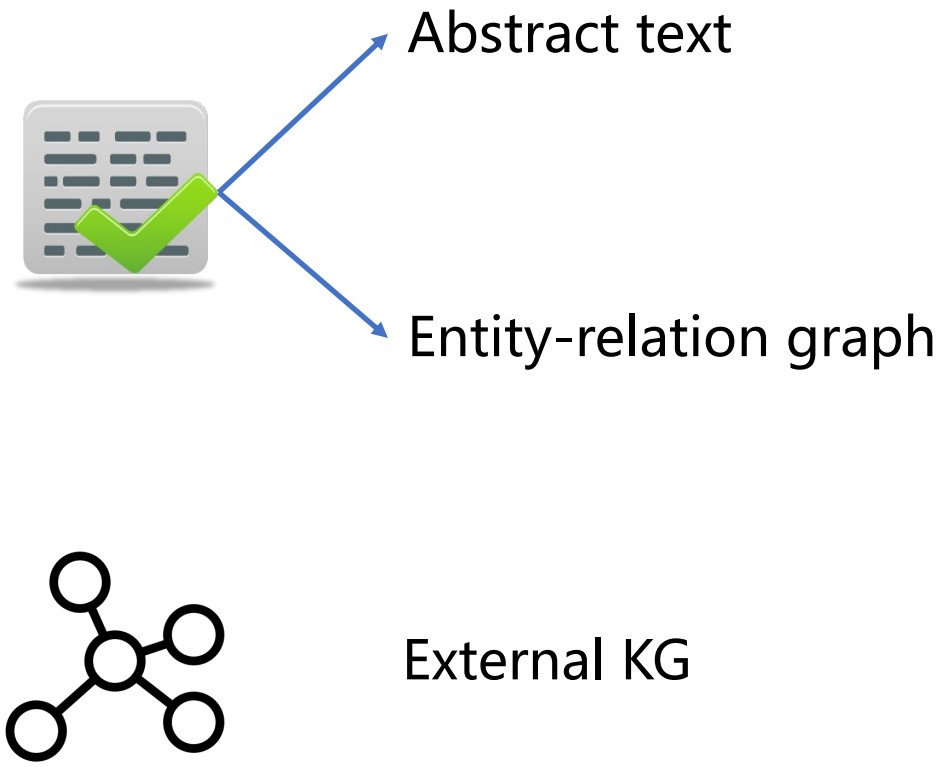
- Extractive method
  - Limited search space
- Generative method
  - Trustworthiness concerns
- Graph method
  - Lack guidance

Biomedical document: The skin is … and causes cancerous process, … that aggravates … skin cancer … .
Method: Extractive method **Triples**: <skin, cancerous, skin cancer> ✓
Method: Generative method **Triples**: <skin, cancerous, skin cancer> ✓ <skin, cancerous, throat cancer> ✗
Method: Knowledge Graph (KG) method **Triples**: <melanoma, belongs to, skin cancer> ✓
Method: Generative method + KG **Triples**: <skin, cancerous, skin cancer> ✓ <skin, cancerous, melanoma> ✓

Taking the advantages of those methods?

Intuition

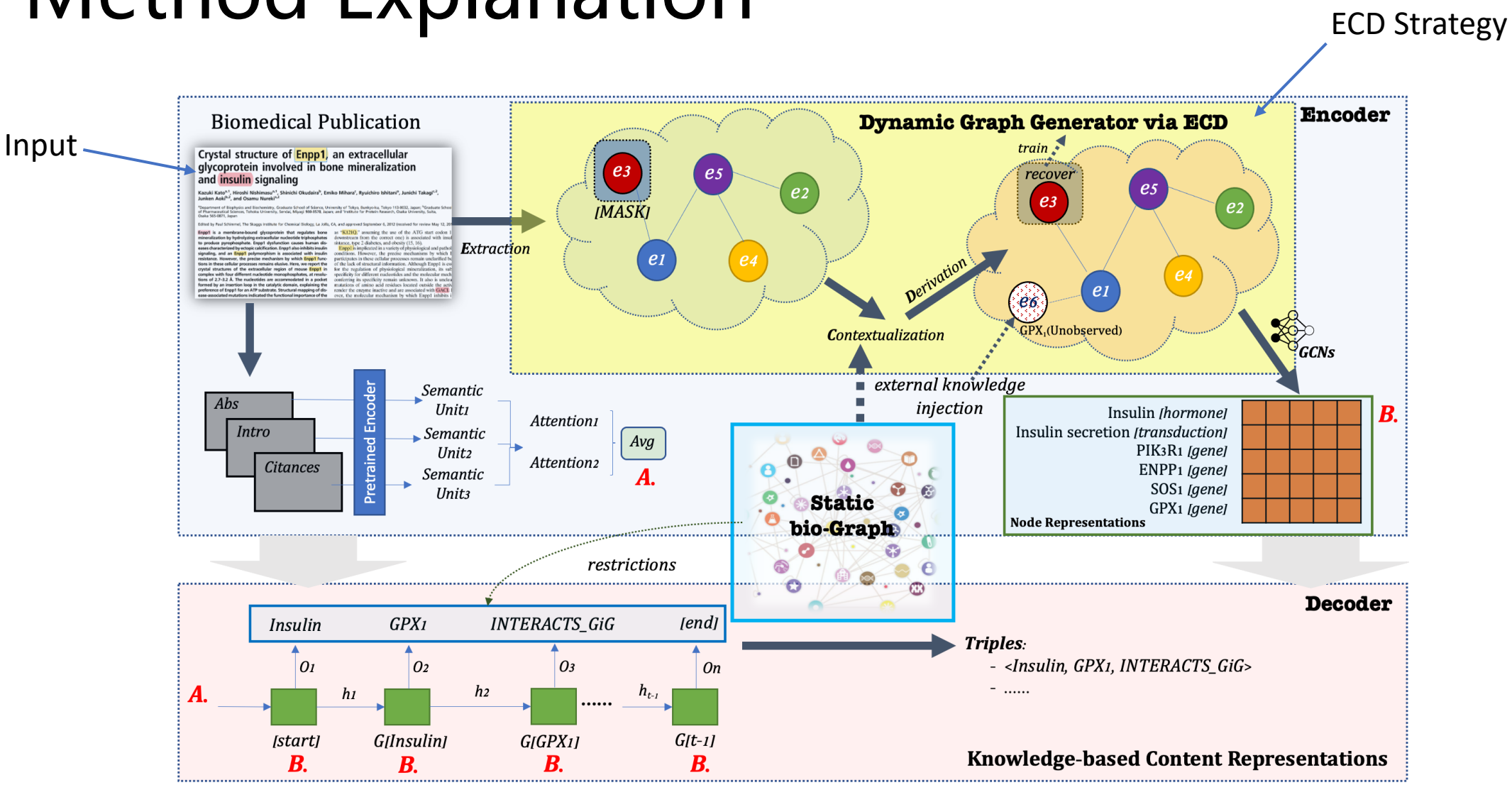
- Semantic text
  - Semantic information
- Entity-relation graph
  - Path information
- External Knowledge Graph
  - Unobserved information



To this end

DGRN: Dual-Graph Resonance Network

Method Explanation



Experimental Setup: Datasets & Tasks

- Task – Triple Extraction
  - Input: A single publication
  - Output: Triplets (e.g., <entity1, entity2, relation>)
- Dataset – Bio-Sci
  - Publications from PMC
  - Triplets from KG

Experimental Setup: Baselines

Model Type	Model
Extractive	HRL (Takanobu et al., 2019)
	CASREL (Wei et al., 2020)
Generative	CopyRE (Zeng et al., 2018)
	CopyMTL (Zeng et al., 2020a)
Graph-based	GAIN (Zeng et al., 2020b)
	AGGCN (Guo et al., 2019)
KG	KeBioLM (Yuan et al., 2021)
	UmlsBert (Michalopoulos et al., 2021)
Pretrained	KECI-BioBert (Lai et al., 2021)
	TEMPGEN-BioBert (Huang et al., 2021a)
	GAIN-BioBert (Zeng et al., 2020a)

Comparative Study

Model	Model Type	P	R	F1
HRL	Extractive	61.17	21.81	32.16
CASREL	Extractive	71.12	32.94	45.03
CopyRE	Generative	54.73	25.72	34.99
CopyMTL	Generative	56.91	29.64	38.98
GAIN	Graph	56.01	21.43	31.00
AGGCN	Graph	61.43	33.91	43.70
KeBioLM	KG	61.18	32.88	42.77
UmlsBert	KG	59.61	29.17	39.17
KECI-BioBert	Extractive	61.93	**40.81**	49.20
TEMPGEN-BioBert	Generative	63.13	33.71	43.95
GAIN-BioBert	Graph	56.61	21.87	31.55
DGRN	Generative+Graph	**73.77***	39.69*	**51.61***

Table 2: Comparison among different models. Models above the double line do not use pre-trained model. Superscript * indicates statistical significance at $p < 0.05$ level compared to the best performance of baselines.

- Generative requires more search space
- Graph model provides topological information
- DGRN performs the best

DRGN Study

- Ablation Study (ECD)

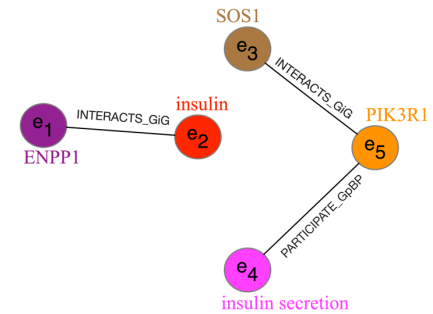
Model	ECD Step	P	R	F1
Extract	1	47.07	23.02	30.92
Extract+Text	1	59.03	38.21	46.39
Extract+Context+Text	1,2	41.17	20.81	27.65
DGRN (Full)	1,2,3	**73.77**	**39.69**	**51.61**

- Masking Study

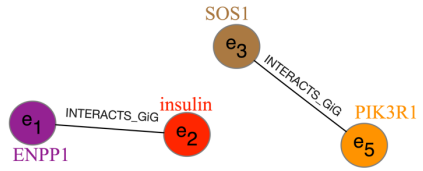
Model	Mask	Precision	Recall	F1-score
DGRN (Full)	**5%**	64.03	33.91	44.34
	10%	**73.77**	**39.69**	**51.61**
	15%	70.91	38.02	49.50
	20%	69.21	37.17	48.37
	25%	60.71	32.03	41.94

Case Study

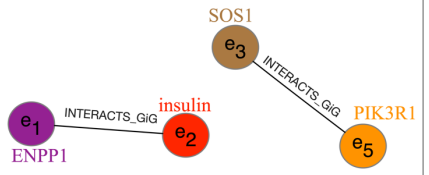
Abstract: PIK3R1 is …disorder. … in both insulin….We examined 152 SOS1…diabetes… .
Introduction: PIK3R1 is ... disease associated... . ...evidence has accumulated ... for the ENPP1 variant of KCNJ11.
Citances: The disorder is characterized ... insulin secretion ... (Weyer et al). …evidence has ... ENPP1 in PIK3R1… (Hani et al).
KG: <GPX6, INTERACTS_GiG, Insulin>, <ENPP1, INTERACTS_GiG, Insulin>, <GPX6, RESEMBLES_GiG, ENPP1> etc.



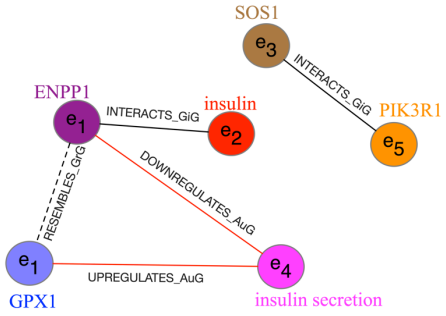
Extractive models (e.g., CasRel)



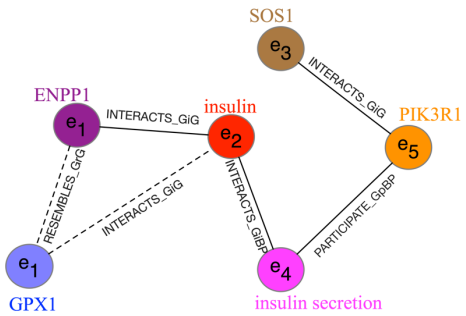
Generative models (e.g., CopyMTL)



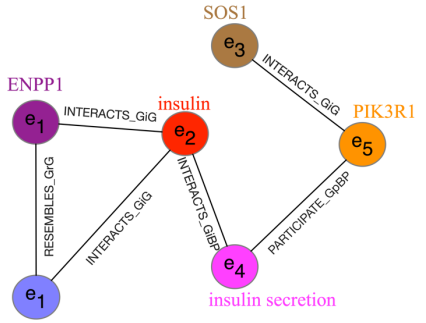
Graph models (e.g., AGGCN)



ChatGPT



DGRN



Ground Truth

Conclusion

- DGRN
  - Shows the superiority of integrating semantic information and topological information
- Next
  - Exploring Scalability and Generalizability

Thanks for listening!

Presenter: Kai Zhang
Worcester Polytechnic Institute
kzhang8@wpi.edu