Natural Language Processing for Healthcare: Text Representation, Multitask Learning, and Applications

Shaoxiong Ji

shaoxiong.ji@helsinki.fi

LT Research Seminar 2023-01-26

Outline

1. Introduction
2. Effective Text Representation
3. High-dimensional and Imbalanced Medical Codes
4. Multitask Learning
5. Conclusion

NLP in Healthcare Domain

Texts in Healthcare Domain

- clinical notes (reports, discharge summaries, etc.)
- biomedical literature
- discussions in medical forum
- social posts about health

Healthcare Applications

- Medical Code Prediction
- Patient Outcome Prediction
- Adverse Drug Event Detection

Medical Code Prediction

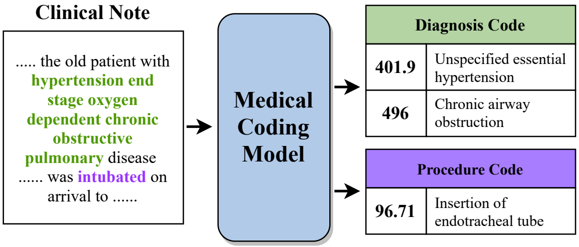

Clinical Note
..... the old patient with hypertension end stage oxygen dependent chronic obstructive pulmonary disease ...... was intubated on arrival to ......
Medical Coding Model
Diagnosis Code
401.9
Unspecified essential hypertension
496
Chronic airway obstruction
Procedure Code
96.71
Insertion of endotracheal tube

Figure: An example of medical coding with ICD codes

- Standard translation of written patient descriptions
- Standardized treatment alignment; insurance reimbursement
- Extreme multi-label multi-class classification

Patient Outcome Prediction

- diagnosis
- length of hospital stay
- development of heart failure
- life expectancy
- the likelihood of readmission after discharge

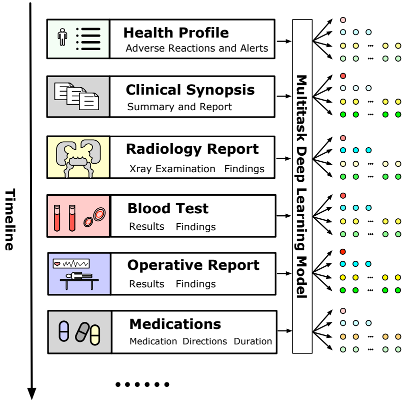


Figure: An example of multitask patient outcome prediction based on sequential inputs of clinical notes in the electronic health record.

Adverse Drug Event Detection

- Post-marketing medication safety surveillance
- Spontaneous reporting systems
- Data from social media, biomedical articles, and medical forums
- NLP models for automated ADE detection

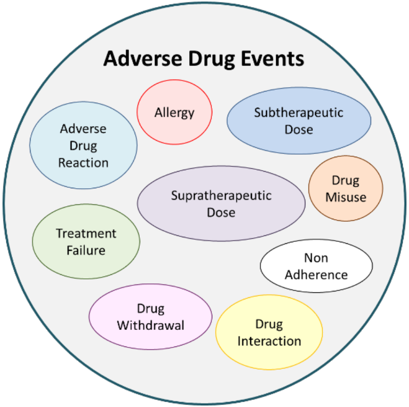


Figure: Adverse drug events. Source: PharmaNet, British Columbia, Canada

Today's Talk

- Effective text representation with deep neural networks,
- Learning with high-dimensional and imbalanced medical codes,
- Solving multiple prediction tasks in healthcare

Effective Text Representation

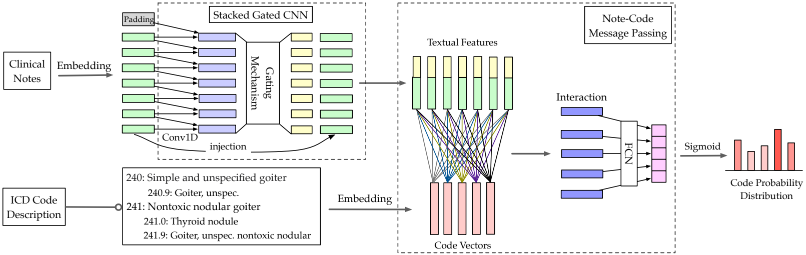
Challenges

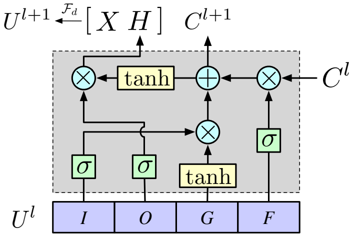
- Complex diagnosis information: professional medical vocabulary and noise e.g., non-standard synonyms and misspellings
- Lengthy documents: from hundreds to thousands of tokens
- Relational learning for adverse drug reactions

Solutions for effective text representation learning:

- Effective convolutional neural networks (CNNs)
- Improved BERT-baed (hierarchical) models
- Contextualized graph embeddings

Effective CNN Encoding: Gated CNN





GatedCNN-NCI in Findings of ACL 2021 (Ji et al., 2021b)

Effective CNN Encoding: Results

Table: Results on MIMIC-III dataset with top-50 ICD codes. “-” indicates no results reported in the original paper.

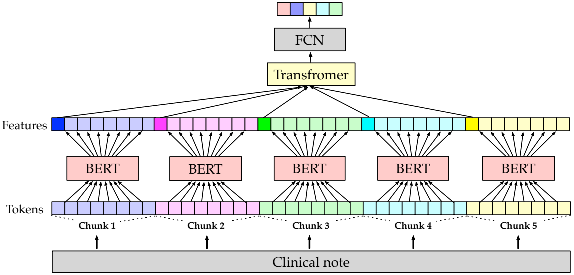
Model	AUC-ROC		F1		
	Macro	Micro	Macro	Micro	P@5
C-MemNN (Prakash et al., 2017)	83.3	-	-	-	42.0
Attentive LSTM (Shi et al., 2017)	-	90.0	-	53.2	-
CAML (Mullenbach et al., 2018)	87.5	90.9	53.2	61.4	60.9
MultiResCNN (Li and Yu, 2020)	89.9±0.4	92.8±0.2	60.6±1.1	67.0±0.3	64.1±0.1
HyperCore (Cao et al., 2020)	89.5±0.3	92.9±0.2	60.9±0.1	66.3±0.1	63.2±0.2
GatedCNN-NCI (ours)	**91.5**±0.3	**93.8**±0.1	**62.9**±0.5	**68.6**±0.1	**65.3**±0.1

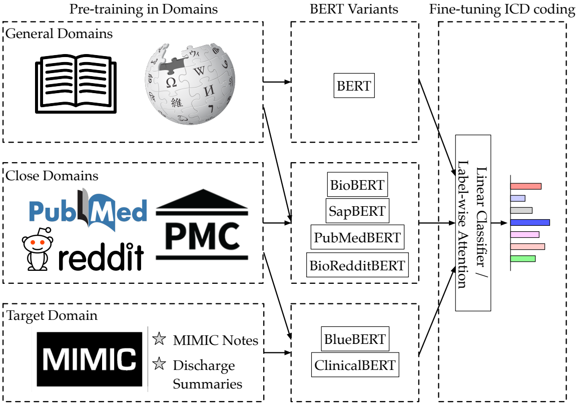
Effective CNN Encoding: Parameters

Table: Number of trainable parameters

Model	num. params.
CAML (Mullenbach et al., 2018)	6.2M
DCAN (Ji et al., 2020)	8.7M
MultiResCNN (Li and Yu, 2020)	11.9M
ClinicalBERT (Alsentzer et al., 2019)	113.8M
GatedCNN-NCI	7.6M

Hierarchical Encoding with Language Models





Published in Computers in Biology and Medicine, Ji et al. (2021a)

Results

Table: Results of PLM fine-tuning with `BERT-hier` + LAN in various domains on MIMIC-III dataset with top-50 and full ICD codes. Clinical notes are truncated at length of 2500.

	MIMIC-III Top-50 Codes					MIMIC-III Full Codes					
Model	AUC-ROC		F1		P@5	AUC-ROC		F1		P@8	P@15
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro		
BERT-base	82.7	86.3	40.8	50.8	52.2	82.2	96.6	5.8	44.1	63.3	48.1
BlueBERT	**89.4**	**92.0**	61.0	65.6	62.8	84.4	97.5	5.1	42.5	62.6	47.3
BioBERT full text	88.8	*91.7*	60.4	66.0	*63.1*	85.2	97.4	**6.4**	**47.0**	*65.8*	*50.7*
BioRedditBERT	87.1	89.6	59.4	64.8	62.4	*86.5*	*98.0*	3.0	40.6	62.4	47.8
PubMedBERT full text	88.6	90.8	**63.3**	**68.1**	**64.4**	**87.4**	**98.1**	4.3	44.5	65.2	50.4
SapBERT full text	88.5	90.8	*62.2*	*66.7*	*63.1*	86.4	97.7	*6.2*	*46.8*	**68.5**	**53.0**
ClinicalBERT all notes	*89.2*	91.6	59.5	64.8	62.0	84.7	97.4	6.0	46.6	65.1	49.9

PubMedBERT, trained entirely from scratch on biomedical article corpora, performs better than other pretrained models from other domains.

Results

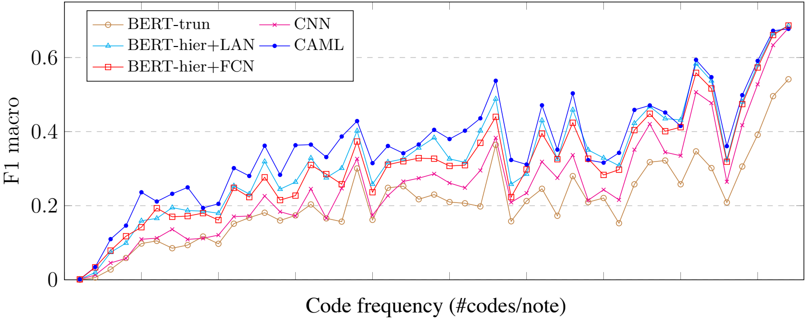


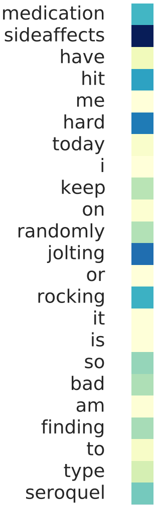
Figure: Comparing BERT-based models with convolution-based approaches. F1 scores of different models on the MIMIC-III full code dataset (8,922 labels). Code frequency groups are sorted in ascending order from left to right.

More advanced hierarchical embedding method and model ensemble has shown promising performance (Zhang and Jankowski, 2022).

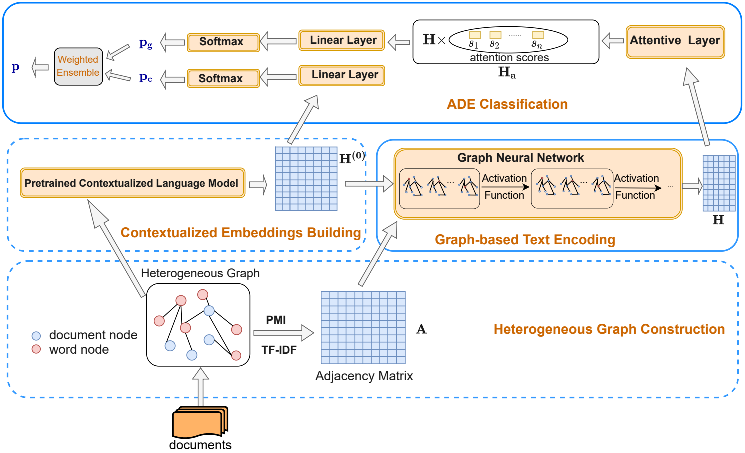
Contextualized Graph Embeddings

Adverse Drug Event Detection

- Feature learning for critical text mentions
- Relational learning
- Effective ADE detection



Contextualized Graph Embeddings



CGEM (Gao et al., 2022) in ECML-PKDD 2022

Results

Table: Results of TwiMed datasets

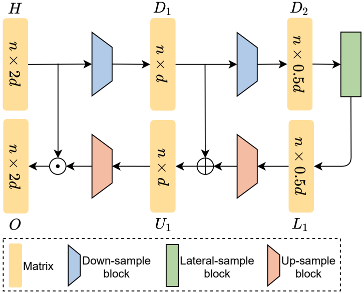
Datasets	Metrics	HTR-MSA	CNN-T	MSAM	IAN	ATL	Ours
TwiMed-Pub	P (%)	75.0	81.3	85.8	87.8	81.5	**88.4**
	R (%)	66.0	63.9	**85.2**	73.8	67.0	85.0
	F1 (%)	70.2	71.6	85.3	79.2	73.4	**86.7**
TwiMed-Twitter	P (%)	60.7	61.8	74.8	83.6	63.7	**84.2**
	R (%)	61.7	60.0	**85.6**	81.3	63.4	83.7
	F1 (%)	61.2	60.9	79.9	82.4	63.5	**83.9**

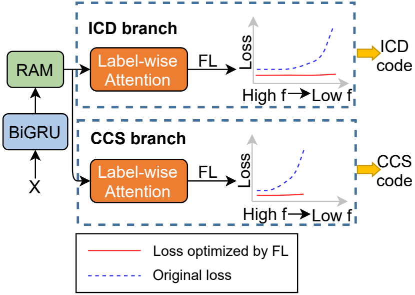
Our graph-based model outperforms other baselines in most cases.

High-dimensional and Imbalanced Medical Codes

Medical code prediction

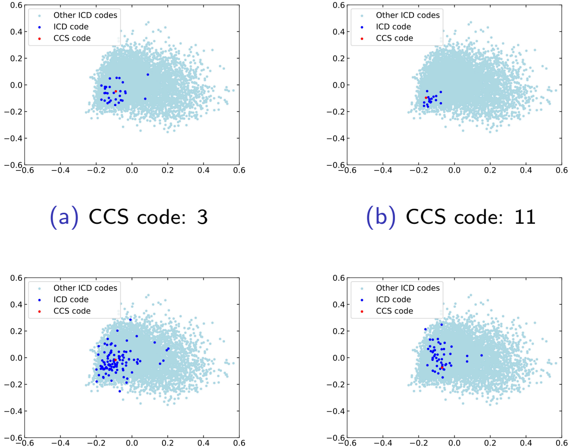
- High-dimensional label space
- Imbalanced labels
- Different disease classification systems
- Multitask learning with different granularities





MT-RAM in ECML-PKDD 2021 (Sun et al., 2021) and MARN in ACM TIST (Sun et al., 2023)

Can multitask learning connect different medical coding systems?



(a) CCS code: 3

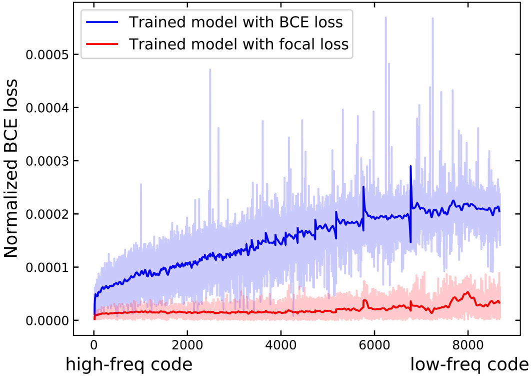
(b) CCS code: 11

(c) CCS code: 195

(d) CCS code: 223

- The embeddings of representative significant CCS codes and their corresponding ICD codes.
- The relevant ICD codes are clustered around the respective significant CCS code.
- MARN learns representations that capture informative relationships between the codes.

Does the model optimized with focal loss balance the learning between low- and high-frequency codes?

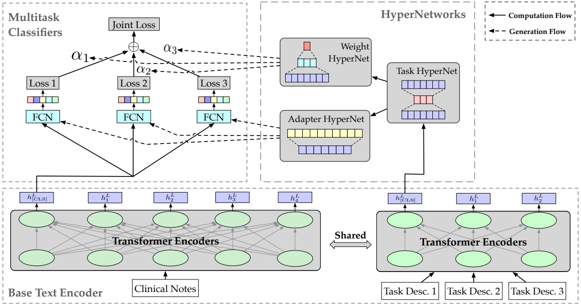


Normalized binary cross entropy loss loss of each ICD code, with x-axis sorted by code frequency. The high-frequency codes are on the left, the low-frequency codes on the right.

MARN optimized with focal loss can balance the learning of high- and low-frequency codes.

Task-conditioned Multitask Learning

- joint training of multiple tasks
- a hypernetwork for task-specific parameter generation to share information among different tasks
- regularize the objective function via a parameterized task weighting scheme that effectively balances the learning process among multiple tasks



MT-Hyper (Ji and Marttinen, 2023) in EACL 2023

Results

Table: Patient outcome prediction from early notes with average score ± standard deviation reported.

Task	Method	Progressive		Ultimate	
		F1	AUC-ROC	F1	AUC-ROC
Readmission	MT-LSTM	51.69 ± 3.72	53.80 ± 1.99	52.36 ± 3.62	56.17 ± 2.23
	MT-BERT	55.52 ± 3.92	56.75 ± 2.56	56.02 ± 4.34	60.85 ± 1.61
	MT-RAM	56.00 ± 3.09	57.26 ± 1.94	56.51 ± 2.80	62.78 ± 1.39
	MT-Hyper	**56.38** ± 2.30	**57.57** ± 1.27	**57.67** ± 2.21	62.41 ± 1.61
Diagnosis	MT-LSTM	8.50 ± 3.39	63.00 ± 0.95	8.69 ± 3.79	62.26 ± 0.91
	MT-BERT	11.63 ± 4.83	64.13 ± 1.30	11.80 ± 5.02	63.24 ± 1.30
	MT-RAM	14.06 ± 1.21	68.58 ± 1.60	14.39 ± 1.41	67.84 ± 1.64
	MT-Hyper	**19.56** ± 1.33	**72.98** ± 0.45	**20.47** ± 1.38	**73.21** ± 0.43
LOS	MT-LSTM	30.19 ± 2.70	75.73 ± 0.92	30.54 ± 2.68	76.50 ± 0.87
	MT-BERT	26.40 ± 4.34	72.60 ± 2.81	27.25 ± 4.08	73.71 ± 2.66
	MT-RAM	26.20 ± 3.08	73.15 ± 3.44	27.08 ± 2.80	74.04 ± 3.31
	MT-Hyper	**33.18** ± 0.91	71.84 ± 1.95	**33.28** ± 1.03	72.23 ± 2.18
Average	MT-LSTM	30.12 ± 3.27	64.18 ± 1.29	30.53 ± 3.36	64.97 ± 1.34
	MT-BERT	31.18 ± 4.36	64.49 ± 2.22	31.69 ± 4.48	65.93 ± 1.86
	MT-RAM	32.09 ± 2.46	66.33 ± 2.33	32.66 ± 2.34	68.22 ± 2.11
	MT-Hyper	**36.37** ± 1.51	**67.47** ± 1.23	**37.14** ± 1.54	**69.28** ± 1.41

Results

Table: Results (AUC-ROC) of zero-shot diagnosis prediction on unseen diagnosis results with average score ± standard deviation reported.

Dataset	Method	Progressive	Ultimate
Discharge	MT-LSTM	11.00 ± 2.31	9.74 ± 1.55
	MT-BERT	10.47 ± 0.76	9.00 ± 0.36
	MT-RAM	11.40 ± 1.44	10.37 ± 1.25
	MT-Hyper	**64.06** ± 2.02	**68.33** ± 2.76
Early	MT-LSTM	8.65 ± 1.19	8.47 ± 1.00
	MT-BERT	8.24 ± 0.79	8.20 ± 0.74
	MT-RAM	14.25 ± 3.11	14.12 ± 3.12
	MT-Hyper	**62.04** ± 1.13	**63.76** ± 1.03

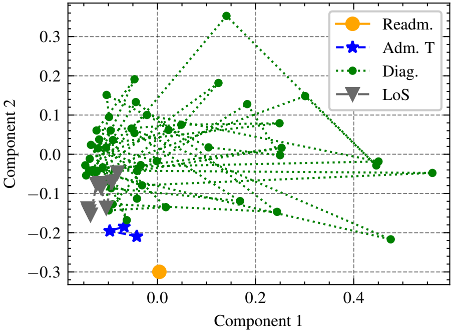


Figure: Visualization of task embeddings with dimension reduced by PCA

Summary

- Effective text representation
  - Effective CNN encoding: GatedCNN-NCI (Ji et al., 2021b)
  - Hierarchical contextualized encoding: BERT-hier (Ji et al., 2021a)
  - Contextualized graph embeddings: CGEM (Gao et al., 2022)
- High-dimensional and imbalanced medical codes
  - MT-RAM (Sun et al., 2021) and MARN (Sun et al., 2023)
  - Multitask learning for high-dimensional codes
  - Focal loss for imbalanced codes
- Multitask learning
  - Hypernetwork-guided multitask learning: MT-Hyper (Ji and Marttinen, 2023)

References I

E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

P. Cao, Y. Chen, K. Liu, J. Zhao, S. Liu, and W. Chong. HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, 2020.

Y. Gao, S. Ji, T. Zhang, P. Tiwari, and P. Marttinen. Contextualized graph embeddings for adverse drug event detection. In *Proceedings of ECML-PKDD*, 2022.

S. Ji and P. Marttinen. Patient outcome and zero-shot diagnosis prediction with hypernetwork-guided multitask learning. In *Proceedings of EACL*, 2023.

S. Ji, E. Cambria, and P. Marttinen. Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 73–78, 2020.

S. Ji, M. Hölttä, and P. Marttinen. Does the Magic of BERT Apply to Medical Code Assignment? A Quantitative Study. *Computers in Biology and Medicine*, 2021a.

S. Ji, S. Pan, and P. Marttinen. Medical code assignment with gated convolution and note-code interaction. In *Findings of ACL-IJCNLP*, 2021b.

F. Li and H. Yu. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8180–8187, 2020.

J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of NAACL-HLT*, pages 1101–1111, 2018.

A. Prakash, S. Zhao, S. A. Hasan, V. Datla, K. Lee, A. Qadir, J. Liu, and O. Farri. Condensed Memory Networks for Clinical Diagnostic Inferencing. In *Proceedings of AAAI*, 2017.

H. Shi, P. Xie, Z. Hu, M. Zhang, and E. P. Xing. Towards Automated ICD Coding Using Deep Learning. *arXiv preprint arXiv:1711.04075*, 2017.

W. Sun, S. Ji, E. Cambria, and P. Marttinen. Multitask Recalibrated Aggregation Network for Medical Code Prediction. In *Proceedings of ECML-PKDD*, 2021.

W. Sun, S. Ji, E. Cambria, and P. Marttinen. Multitask balanced and recalibrated network for medical code prediction. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–20, 2023.

N. Zhang and M. Jankowski. Hierarchical BERT for Medical Document Understanding. *arXiv preprint arXiv:2204.09600*, 2022.