A Multiclass Classification Method Based on Deep Learning for Named Entity Recognition in Electronic Medical Records

Xishuang Dong *, Lijun Qian, Yi Guan,

Lei Huang, Qiubin Yu, Jinfeng Yang

***Corresponding author, presenter**

Postdoc, Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT)

Department of Electrical and Computer Engineering

Prairie View A&M University

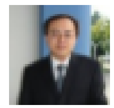
Email: dongxishuang@gmail.com

Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT)

Sponsored by US DOD

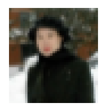
People of CREDIT

- Lijun Qian, Ph.D.
  PI and Director
  Professor of Electrical and Computer Engineering

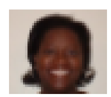


- Lei Huang, Ph.D.
  Co-PI and Associate Director for Research
  Assistant Professor of Computer Science

- John Fuller, Ph.D.
  Co-PI and Associate Director for Education & Outreach
  Professor of Electrical and Computer Engineering



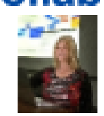
- Xiangfang Li, Ph.D.
  Co-PI
  Assistant Professor of Electrical and Computer Engineering



- Pamela Obiomon, Ph.D.
  Co-PI
  Associate Professor & Interim Department Head of Electrical and Computer Engineering

- Yonggao Yang, Ph.D.
  Co-PI
  Professor & Department Head of Computer Science

Collaborator:



- Dr. Barbara Chapman
  Professor, Stony Brook University

Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT)

- **Mission:** perform big data research for mission-critical applications and provide training to students and professionals
- **Research Thrusts:**
  - System architecture design for a military cloud computing system
  - Secure and robust big data aggregation and storage
  - Novel machine learning algorithms designed for big high-dimensional dataset
  - Visualization of massive military datasets interactively
- **Location:** Prairie View A&M University of the Texas A&M University System located near Houston, Texas
- **Sponsor:** US DOD OSD/AFRL
- Contact: Lijun Qian (liqian@pvamu.edu)

Research Interests

- **Natural Language Processing (NLP)**
  - **Sentiment Analysis on Texts**
    - **Best Results in TREC 2010 Blog Track: Faceted Blog Distillation**
    - **Dong, X.**; Zou, Q. & Guan, Y. Set-Similarity Joins Based Semi-supervised Sentiment Analysis, ICONIP 2012, 2012, 176-183
  - **Machine Learning for NLP**
    - Yang, J.; Guan, Y.; **Dong, X.** & He, B. Representing Words As Lymphocytes, Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press, 2014, 3146-3147.
- **Current focus**
  - **Convolutional neural network based big data analysis**
  - **E.g. Seismic data, Electricity Load data, NLP**

Outline

- **Named entity recognition in electronic medical records**
- **Methodology**
- **Experimental results**
- **Discussion**
- **Conclusion and future work**

Outline

- **Named entity recognition in electronic medical records**
- **Methodology**
- **Experimental results**
- **Discussion**
- **Conclusion and future work**

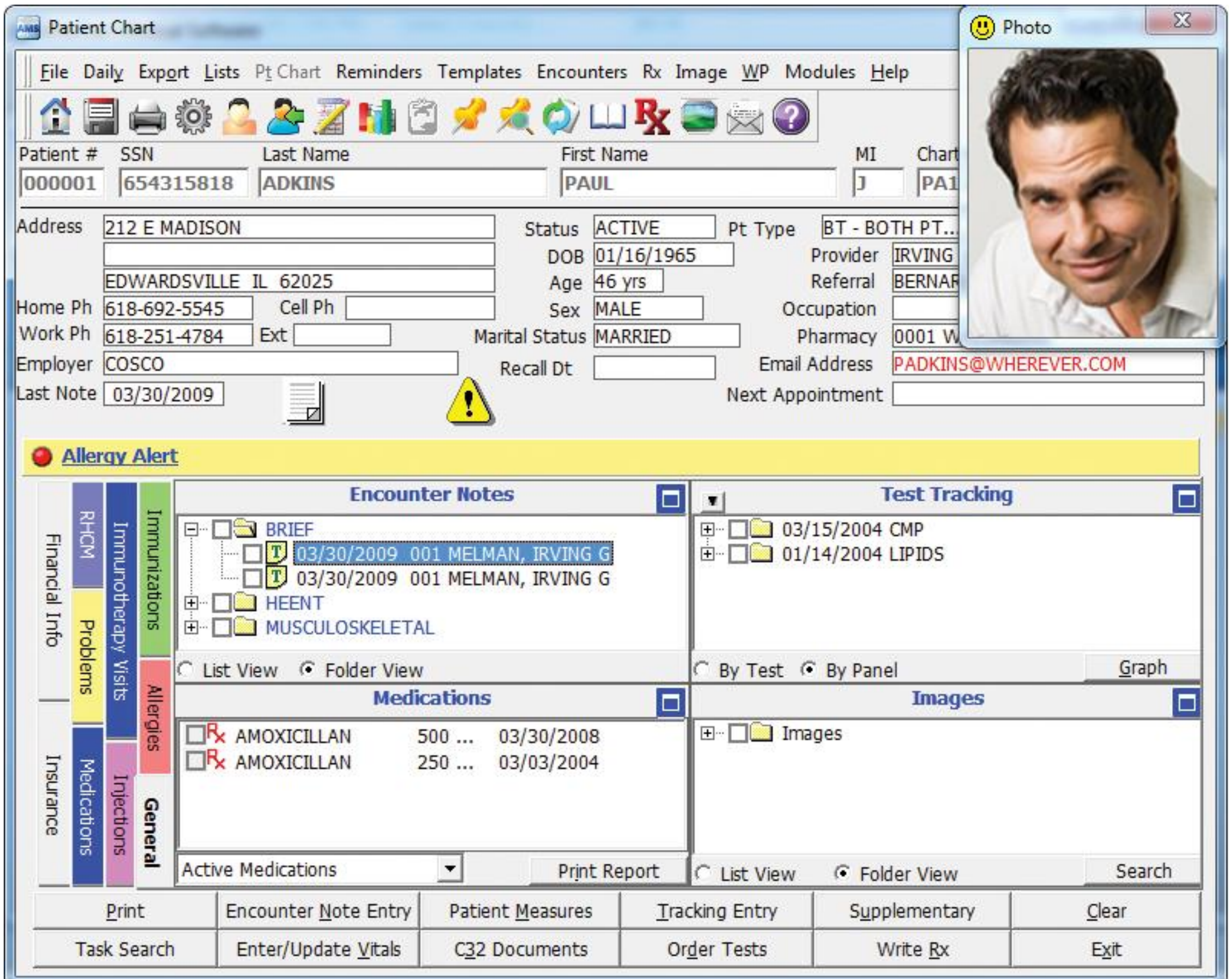
Named entity recognition in electronic medical records

- Electronic medical records (EMRs)
- Named entity recognition
- Previous studies
- Our goals

Named entity recognition in electronic medical records

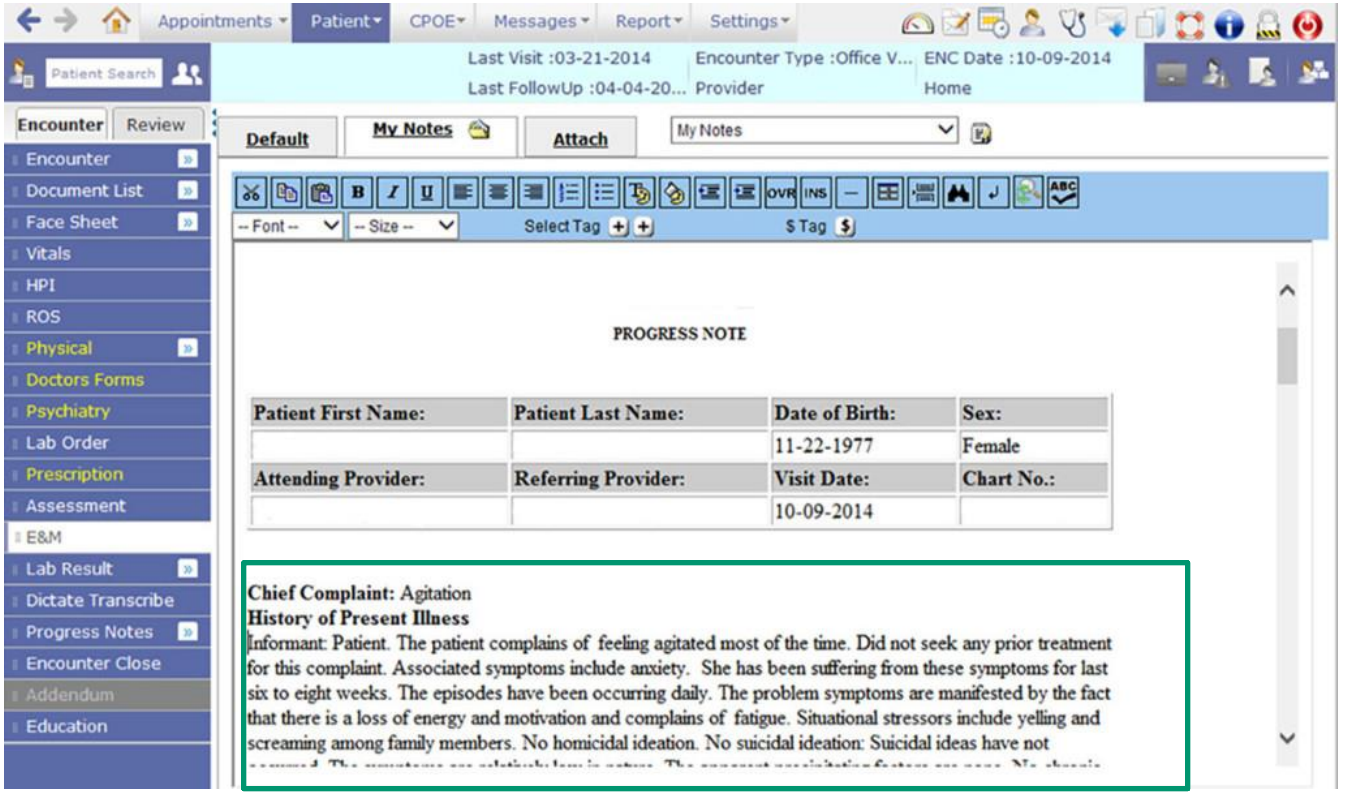
- **Electronic medical records (EMRs)**
  - Semi-structure data
  - Captured by medical staffs using health information systems in clinical activities.
  - Contain words, symbols, charts, graphs, numbers, and images detailing the health conditions of patients.

Named entity recognition in electronic medical records



Named entity recognition in electronic medical records

- Electronic medical records (EMRs)



Named entity recognition in electronic medical records

- **Electronic medical records (EMRs)**
  - Language characteristics
    - massive medical jargons, for example, “cerebral infarction”;
    - test results followed by units or doses such as “100/70 mmHg”;
    - numerous abbreviations such as “CT”;
    - incomplete syntactic components of sentences.

Named entity recognition in electronic medical records

- **Named entity recognition (NER)**
  - A subtask of NLP
  - Seeks to locate and classify named entities in text into pre-defined categories
    - Names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, and so on.

Named entity recognition in electronic medical records

- **Named entity recognition**

https://www.ravn.co.uk/named-entity-recognition-ravn-part-1/

Named entity recognition in electronic medical records

- **Previous studies**
  - **Lexicon-based**
  - **Supervised machine learning-based**
    - **classification**

Named entity recognition in electronic medical records

- **Previous studies**
  - **Supervised machine learning based**
    - **Classification**

“ Fred showed Sue Mengqui Huang’s new painting ”

Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

Named entity recognition in electronic medical records

- **Previous studies**
  - **NER in EMRs**
    - **Seeks to locate and classify named entities in EMRs into pre-defined categories**
    - **Names of drugs, treatments, test, and so forth.**

Named entity recognition in electronic medical records

- **Previous studies**
  - **Most of studies focus on NER in English EMRs**
  - **Deep learning**
    - **Convolutional neural network (CNN)**
  - **Word to Vectors (Word2Vec)**

Named entity recognition in electronic medical records

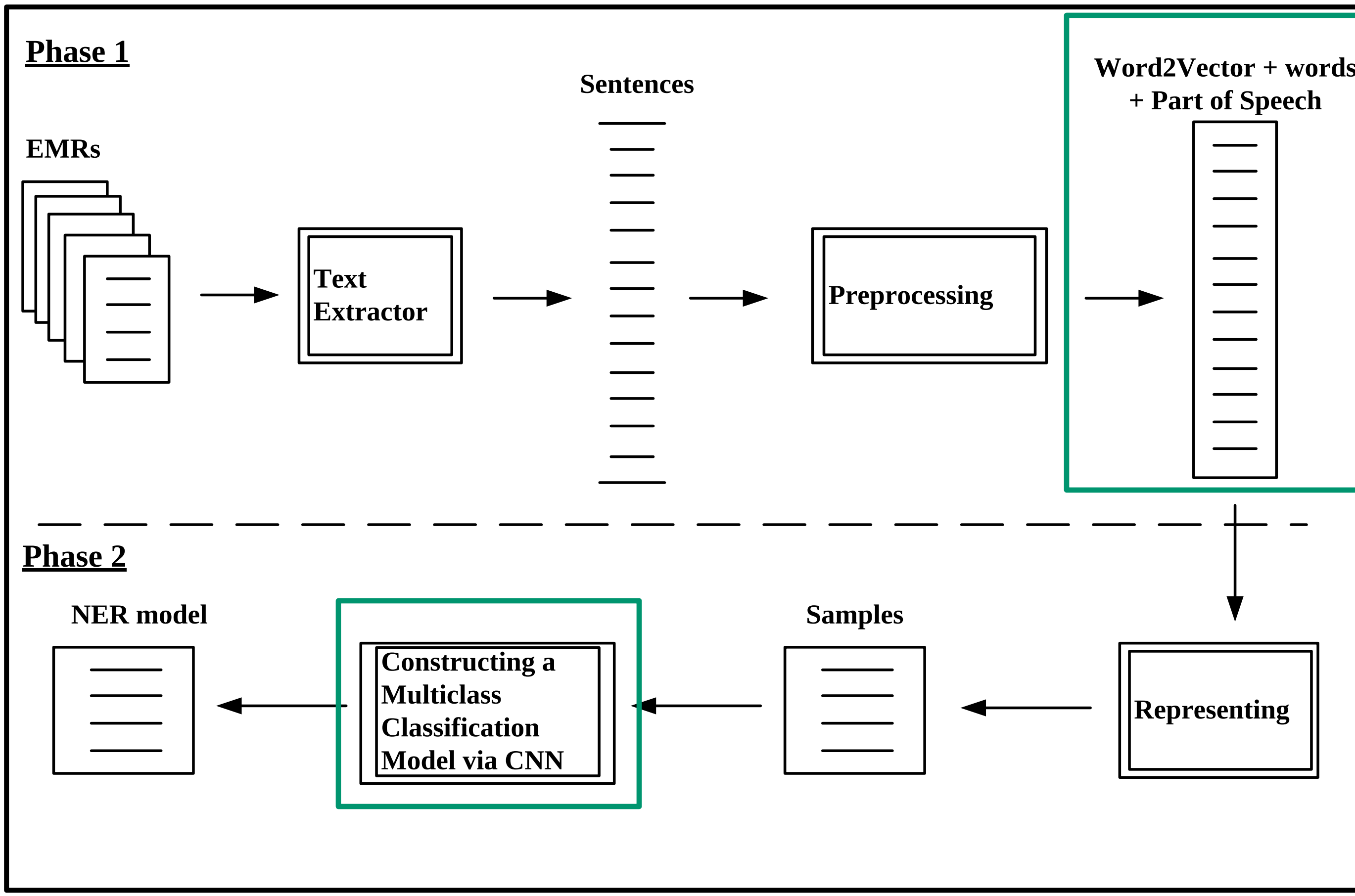
- **Our goals**
  - **Construct a model for accomplishing NER in Chinese EMRs**
  - **Using advantages of CNN and Word2Vec**

Outline

- Named entity recognition in electronic medical records
- **Methodology**
- Experimental results
- Discussion
- Conclusion and future work

Methodology

- **Framework**



Methodology

- **Word2Vec (2013 Google)**
  - **A new word representation**
  - **Reduce dimensions of data representation**
  - **Overcome challenges of data sparseness**
  - **…**

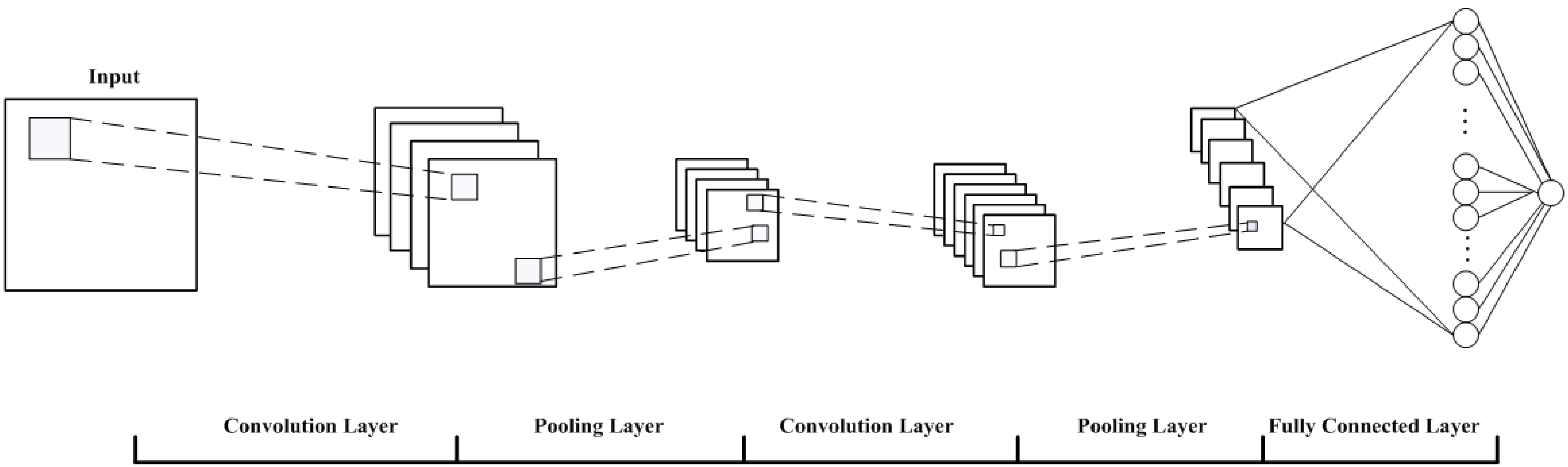
Methodology

- **Word2Vec for EMRs analysis**

The	0.50, 0.82, 0.46,...,0.37
patient	0.11, 0.30, 0.33,...,0.67
complains	0.15, 0.22, 0.54,...,0.27
of	0.25, 0.23, 0.41,...,0.72
feeling	0.51, 0.12, 0.84,...,0.17
agitated	0.75, 0.42, 0.74,...,0.57

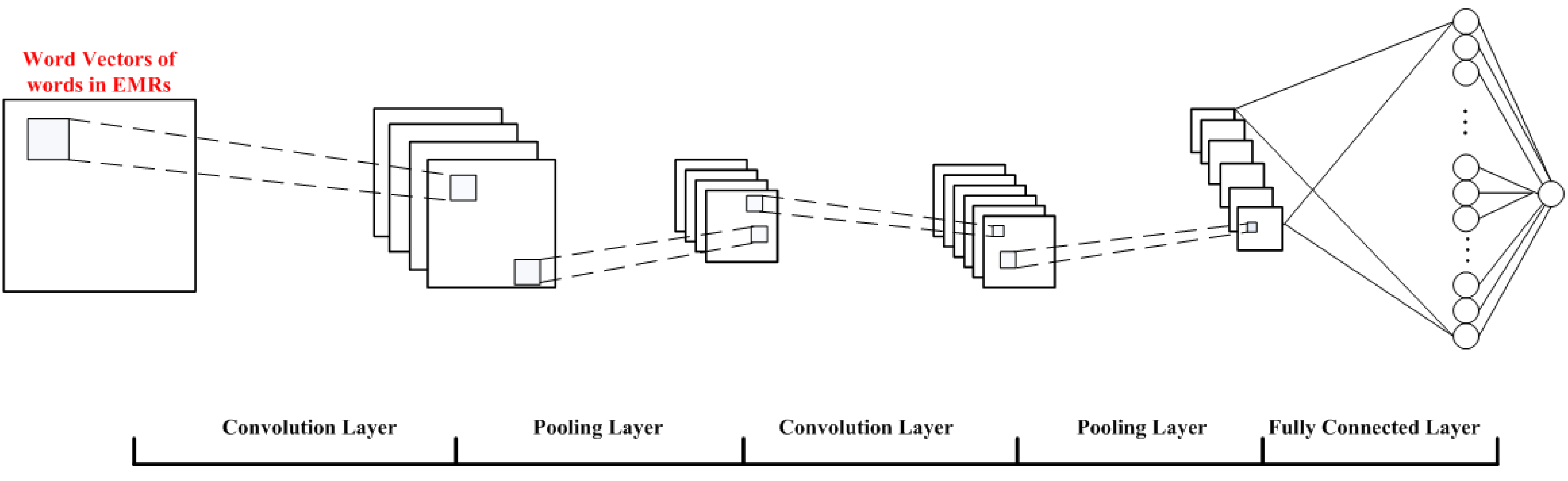
Methodology

- **CNN**



Methodology

- **CNN for NER in EMRs**

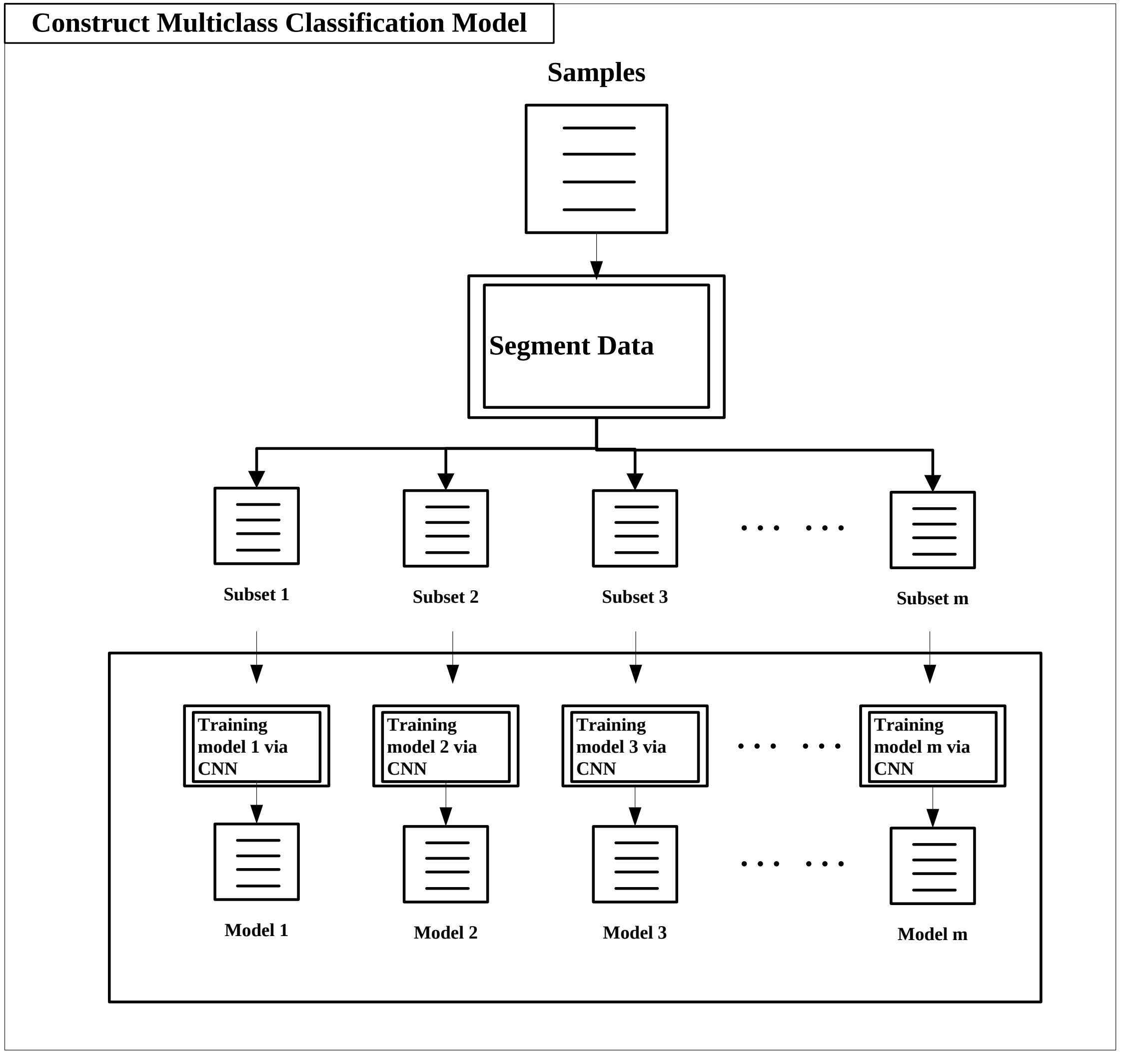


Methodology

- Predefined categories of named entities in EMRs
  - Five categories
    - Disease
    - Symptom
    - Treatment
    - Test
    - Disease Group
  - A multiclass classification problem

Methodology

- **Training models**



Outline

- **Named entity recognition in electronic medical records**
- **Methodology**
- **Experimental results**
- **Discussion**
- **Conclusion and future work**

Experimental results

- **Data set**
- **Results**

Experimental results

- **Data set**
  - **Chinese EMRs from Second Affiliated hospital of Harbin Medical University, Harbin City, Heilongjiang Province, China**

Sentence	**Tagging Results**
患儿既往健康,第1胎, 第1产,青霉素过敏史,生长发育正常,无家族遗传疾病史,按计划免疫接种各种疫苗. (The patient was healthy before, first birth born, allergy history of penicillin, inoculated on schedule with various vaccines planned immunization, developmental history was normal, no hereditary disease family history.)	患儿/O 既往/O 健康/O ,/O 第/B_disease 1/I_disease 胎/I_disease ,/O 第/B_disease 1/I_disease 产/I_disease ,/O 青霉素/B_disease 过敏史/I_disease ,/O 生长/O 发育/O 正常/O ,/O 无/O 家族/B_disease 遗传/I_disease 疾病史/I_disease ,/O 按/O 计划/O 免疫/O 接种/O 各/O 种/O 疫苗/B_treatment ./O

Experimental results

- **Data set**

EMR Type	#Documents	#Sentences	#Characters	#Entities
Discharge Summary	500	27,110	463,918	Disease: 3,554 Symptom: 7,461 Treatment: 2,457 Test: 2,672 Disease Group: 151 Total: 16,295
Progress Note	492	28,375	965,852	Disease: 4,769 Symptom: 11,479 Treatment: 2,785 Test: 4,317 Disease Group: 72 Total: 23,422
Overall	992	55,485	1,429,770	Disease: 8,323 Symptom: 18,940 Treatment: 5,242 Test: 6,989 Disease Group: 223 Total: 39,717

Experimental results

- **Results on Discharge Summary** **(Accuracy %)**

	Entity Type					
Model	**Disease**	**Disease Group**	**Symptom**	**Test**	**Treatment**	**Overall**
NB	44.82	N/A	51.72	65.96	59.00	58.91
ME	48.32	34.19	56.34	76.10	58.80	65.68
SVM	57.18	37.22	62.52	80.17	60.48	70.46
CRF	77.33	48.39	77.83	90.05	77.47	83.94
Our Model	**52.80**	**40.00**	**65.76**	**79.28**	**53.14**	**68.60**

Experimental results

- **Results on Progress Notes (Accuracy %)**

	Entity Type					
Model	**Disease**	**Disease Group**	**Symptom**	**Test**	**Treatment**	**Overall**
NB	69.50	N/A	70.09	71.85	41.59	67.49
ME	71.49	41.15	72.37	77.58	52.93	72.44
SVM	77.77	21.12	76.92	81.49	56.36	76.45
CRF	87.24	36.06	87.09	90.31	75.60	87.22
Our Model	**76.19**	**12.50**	**76.31**	**76.65**	**51.83**	**73.40**

Outline

- **Named entity recognition in electronic medical records**
- **Methodology**
- **Experimental results**
- **Discussion**
- **Conclusion and future work**

Discussion

- We present an effective method to mine NER from Chinese EMRs according to experimental results.
  - Not to pay many attentions to feature selection
- Two deficiencies of our method
  - Cannot model relations between words
  - Consume a mass of computation resources and time for building many of classifiers

Outline

- **Named entity recognition in electronic medical records**
- **Methodology**
- **Experimental results**
- **Discussion**
- **Conclusion and future work**

Conclusion and future work

- **We present an effective multiclass classification method and verify its effectiveness on a corpus consisting of Chinese EMRs.**
- **The method can be used to solve other multiclass classification problems such as image labeling, semantic role labeling of words, and semantic relation classification.**

Conclusion and future work

- **Verify effectiveness of our methods in other applications**
- **Build a dependency parser system to extract dependency syntactic relations.**
- **Automatically annotate EMRs to gain big data for research.**

Thank you!

Q&A