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**
    
  
• **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”
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

Phase 1

EMRs

Text Extractor

Sentences

Preprocessing

Word2Vector + words + Part of Speech

Phase 2

NER model

Constructing a Multiclass Classification Model via CNN

Samples

Representing
Methodology

• Word2Vec (2013 Google)
  – A new word representation
  – Reduce dimensions of data representation
  – Overcome challenges of data sparseness
  – …
Methodology

- Word2Vec for EMRs analysis

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>0.50</td>
<td>0.82</td>
<td>0.46</td>
</tr>
<tr>
<td>patient</td>
<td>0.11</td>
<td>0.30</td>
<td>0.33</td>
</tr>
<tr>
<td>complains</td>
<td>0.15</td>
<td>0.22</td>
<td>0.54</td>
</tr>
<tr>
<td>of</td>
<td>0.25</td>
<td>0.23</td>
<td>0.41</td>
</tr>
<tr>
<td>feeling</td>
<td>0.51</td>
<td>0.12</td>
<td>0.84</td>
</tr>
<tr>
<td>agitated</td>
<td>0.75</td>
<td>0.42</td>
<td>0.74</td>
</tr>
</tbody>
</table>
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

Construct Multiclass Classification Model

Samples

Segment Data

Subset 1
Subset 2
Subset 3
Subset m

Training model 1 via CNN
Training model 2 via CNN
Training model 3 via CNN

Model 1
Model 2
Model 3
Model m
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

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Tagging Results</th>
</tr>
</thead>
</table>
## Experimental results

### Data set

<table>
<thead>
<tr>
<th>EMR Type</th>
<th>#Documents</th>
<th>#Sentences</th>
<th>#Characters</th>
<th>#Entities</th>
</tr>
</thead>
</table>
| 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 %)*

<table>
<thead>
<tr>
<th>Model</th>
<th>Disease</th>
<th>Disease Group</th>
<th>Symptom</th>
<th>Test</th>
<th>Treatment</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>44.82</td>
<td>N/A</td>
<td>51.72</td>
<td>65.96</td>
<td>59.00</td>
<td>58.91</td>
</tr>
<tr>
<td>ME</td>
<td>48.32</td>
<td>34.19</td>
<td>56.34</td>
<td>76.10</td>
<td>58.80</td>
<td>65.68</td>
</tr>
<tr>
<td>SVM</td>
<td>57.18</td>
<td>37.22</td>
<td>62.52</td>
<td>80.17</td>
<td>60.48</td>
<td>70.46</td>
</tr>
<tr>
<td>CRF</td>
<td>77.33</td>
<td>48.39</td>
<td>77.83</td>
<td>90.05</td>
<td>77.47</td>
<td>83.94</td>
</tr>
<tr>
<td>Our Model</td>
<td><strong>52.80</strong></td>
<td><strong>40.00</strong></td>
<td><strong>65.76</strong></td>
<td><strong>79.28</strong></td>
<td><strong>53.14</strong></td>
<td><strong>68.60</strong></td>
</tr>
</tbody>
</table>
### Experimental results

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Disease</th>
<th>Disease Group</th>
<th>Symptom</th>
<th>Test</th>
<th>Treatment</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>69.50</td>
<td>N/A</td>
<td>70.09</td>
<td>71.85</td>
<td>41.59</td>
<td>67.49</td>
</tr>
<tr>
<td>ME</td>
<td>71.49</td>
<td>41.15</td>
<td>72.37</td>
<td>77.58</td>
<td>52.93</td>
<td>72.44</td>
</tr>
<tr>
<td>SVM</td>
<td>77.77</td>
<td>21.12</td>
<td>76.92</td>
<td>81.49</td>
<td>56.36</td>
<td>76.45</td>
</tr>
<tr>
<td>CRF</td>
<td>87.24</td>
<td>36.06</td>
<td>87.09</td>
<td>90.31</td>
<td>75.60</td>
<td>87.22</td>
</tr>
<tr>
<td>Our Model</td>
<td><strong>76.19</strong></td>
<td><strong>12.50</strong></td>
<td><strong>76.31</strong></td>
<td><strong>76.65</strong></td>
<td><strong>51.83</strong></td>
<td><strong>73.40</strong></td>
</tr>
</tbody>
</table>
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