AI for Note-Based Information Extraction, Curation, (and Analytics)

Dan Ruan
Department of Radiation Oncology & Department of Bioengineering
University of California, Los Angeles
Why notes

• Note as the feeder to clinical data warehouse (registries)

• Beyond DB data elements
  • Quality research with user-defined and empirically induced search patterns, contextual understanding of treatment rationale, drug use, symptom surveillance
  • Meaningful use stage 2 – documentation relates to “transition of care” “data portability” and “clinical summary”
  • Summarization: patient-level aggregate (chart review), group aggregate
General pipeline description

• Preprocessing: delineation, normalization, tokenization, POS tagging, noun-phrase chunking, name entity recognition (NER)

• Data warehousing: utilization of domain-specific concept hierarchy to classify entities into categories

• Association mining and predictive models
Name entity recognition (NER) annotator

- Dictionary and rule based
- Machine learning models (CRF, SSVM)
- Deep learning
Rule-base systems: most established

- MedLEE |Medical Language Extraction and Encoding System (Columbia University, proprietary) was designed to process radiology reports, later extended to other domains, and tested for transferability to another institution. MedLEE discovers clinical concepts along with a set of modifiers.

- HITEx: Health Information Text Extraction (Brigham and Women's Hospital and Harvard Medical School), open-source clinical NLP system incorporated within i2b2 toolset
Rule-base systems: most established

• MetaMap: (National Library of Medicine), for processing biomedical scholarly articles, providing mappings to the UMLS Metathesaurus concepts

• BioTeKS and MedKAT: IBM’s biomedical domain NLP.

• Intended or cancer biomedical informatics (university of Pittsburg) cancer tissue information extraction system, uses NCI enterprise vocabulary system and MetaMap for information extraction from surgical pathology reports.
State of the art ML-based NER systems

- CRF: i2b2 2010 clinical concept extraction challenge winner.
  - Human designed features with bag of words model
  - Semi-supervised Markov model.
- SSVM model: i2b2 best after-challenge performance
  - Distributional word representation by random indexing
  - Generalizes SVM for structured output labels.
CRFs

• Formulate the clinical NER as a sequence labeling problem: tagging annotated entities

• Features at various levels
  • Context words and n-grams
  • Linguistic features and document levels (section association)
  • Derived knowledge features from general clinical NLP systems (MedLEE, MetaMap, KnowledgeMap), and dictionary from UMLS.
Challenges

• Feature representation may not have sufficient connection to concept.

• Feature engineering of important tokens and token combinations (loc + disorder).

• CRF’s local word windows may not have sufficient long-term dependency.
More recent deep learning work

• Raw sequence in sentence
• Improved word embedding
• CNN and RNN
DL - CNN approach

Wu et al. Stud Health Technol Inform 2015
DL – RNN approach

Concatenation of the left and right representations from Bi-LSTM

Lample et al Arxiv 2016
## Performance comparison of ML+DL

<table>
<thead>
<tr>
<th>Approach</th>
<th>Feature</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRFs baselines</td>
<td>Word</td>
<td>82.32</td>
<td>72.92</td>
<td>77.33</td>
</tr>
<tr>
<td></td>
<td>Word+Linguistic+Discourse</td>
<td>83.25</td>
<td>76.75</td>
<td>79.87</td>
</tr>
<tr>
<td></td>
<td>Word+Linguistic+Discourse+ MedLEE+KnowledgeMap+DST</td>
<td>86.52</td>
<td>81.04</td>
<td>83.60</td>
</tr>
<tr>
<td>SSVMs by Tang et al. (Current best)</td>
<td>All features in CRF baselines +Brown clustering + Random indexing</td>
<td>87.38</td>
<td>84.31</td>
<td>85.82</td>
</tr>
<tr>
<td>Semi-Markov (Best in challenge)</td>
<td>Word+context+sentence+section+cTAKES+MetaMap+ConText+Brown clustering</td>
<td>86.88</td>
<td>83.64</td>
<td>85.23</td>
</tr>
<tr>
<td>CNN</td>
<td>Word embedding</td>
<td>84.91</td>
<td>80.73</td>
<td>82.77</td>
</tr>
<tr>
<td>RNN</td>
<td>Word embedding</td>
<td>85.33</td>
<td>86.56</td>
<td>85.94</td>
</tr>
</tbody>
</table>

Wu et al, AMIA 2017
Curation (free?)

- Negation and context annotators e.g., NegEx algorithm
- Normalization (variation of string -> single embedding)
- Contextual embedding (polysemy -> distinct embedding)
- Mapping to concept
Collaborative effort

- Annotation platform that supports collaborative hierarchical entity typing
  - facilitates knowledge elicitation by allowing the creation and continuous refinement of concept hierarchies during annotation
  - minimize not only annotation time but the time it takes for project creators to set up and distribute projects to annotators

- Customizable, modularized open-source computational methodology for analytical knowledge discovery (t2kg)

Natural language processing: an introduction. Nadkarni PM et al 2011

An overview of MetaMap: historical perspective and recent advances. Aronson AR et al 2010

Towards a comprehensive medical language processing system: methods and issues. Friedman C 1997

The KnowledgeMap project: development of a concept-based medical school curriculum database. Denny JC et al 2003

Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. de Bruijn B et al. 2011

A hybrid system for temporal information extraction from clinical text. Tang B et al 2013


Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Savova GK et al 2010