# 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 |**Med**ical **L**anguage **E**xtraction and **E**ncoding 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: **H**ealth **I**nformation **T**ext **Ex**traction (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.

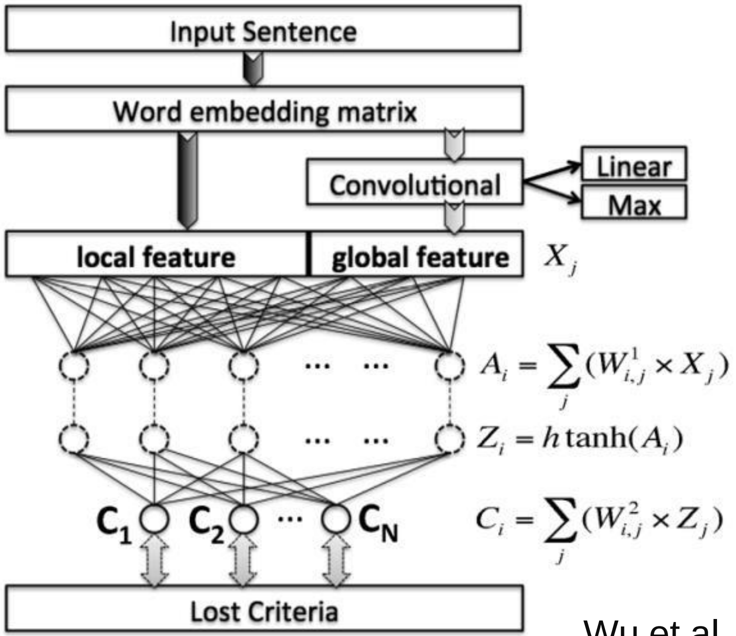
# Challenges

- 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

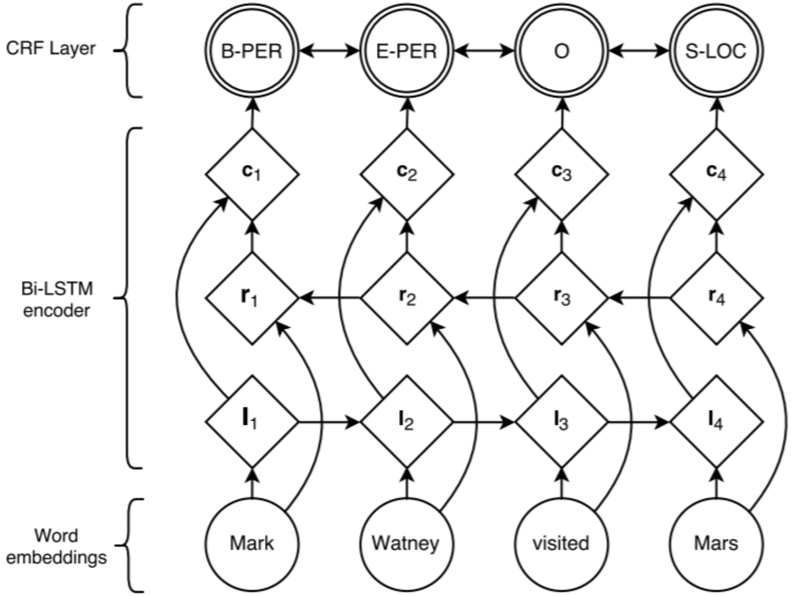
- Raw sequence in sentence
- Improved word embedding
- CNN and RNN

# 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

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

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)

Extracting information from textual documents in the electronic health record: a review of recent research. Meystre SM et. 2008

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 C1997

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

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Lafferty JD et al . 2001

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

Unlocking the Power of Deep PICO Extraction: Step-wise Medical NER Identification Zhang T et al 2020.