

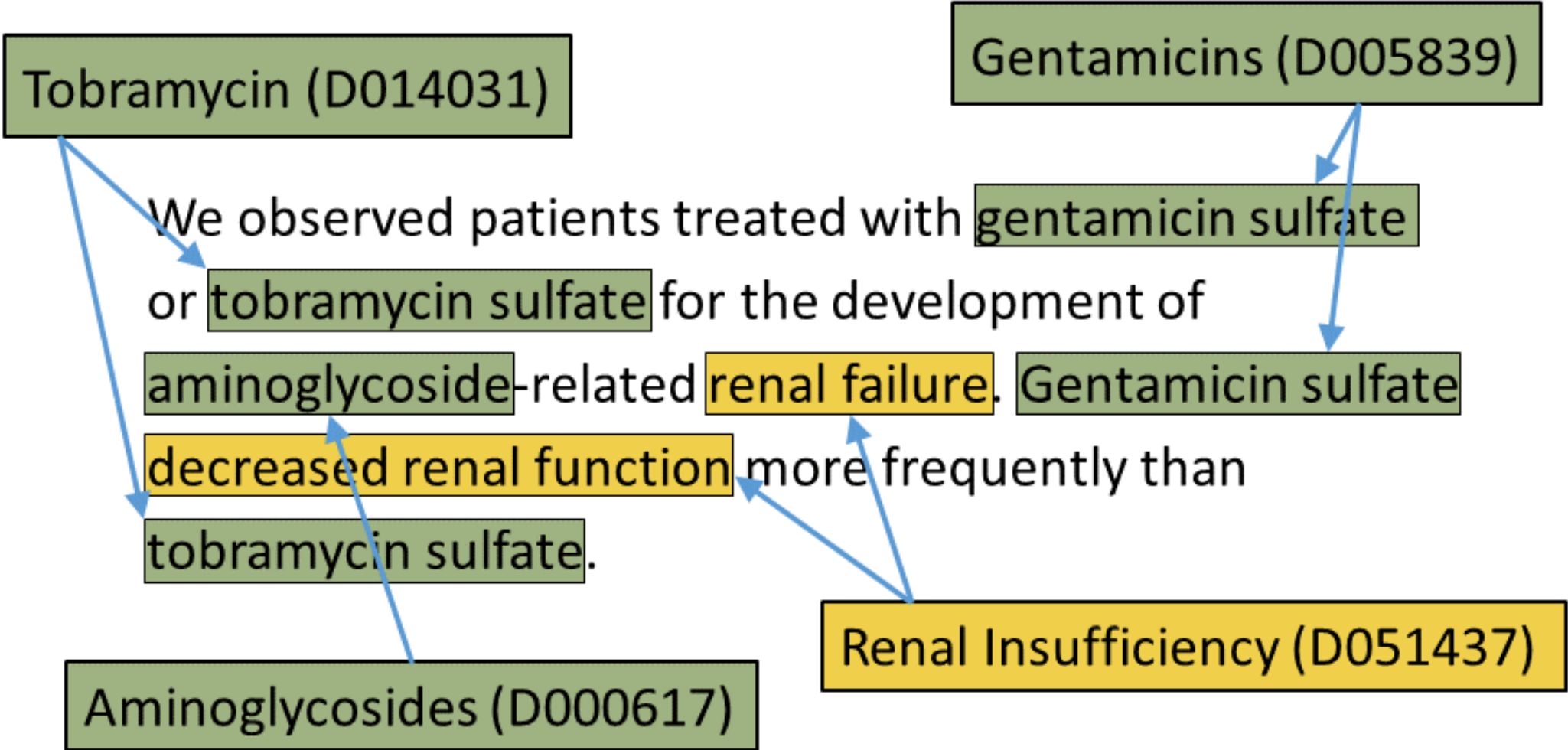
Biomedical Named Entity Recognition and Information Extraction with PubTator

Robert Leaman & Shankai Yan

May 10, 2019



Named Entities: Recognition and Normalization



Challenge: name variation

Pattern	Disease Examples
Neoclassical	Nephropathy
Eponyms	Schwartz-Jampel syndrome
Anatomy	breast cancer
Symptoms	cat-eye syndrome
Causative agent	staph infection
Biomolecular etiology	G6PD deficiency
Heredity	X-linked agammaglobulinemia
Traditional	pica, founder

Pattern	Gene Examples
Phenotype appearance	White swiss cheese
Function	Heat shock protein 60 Calmodulin suppressor of p53
Pop culture	Sonic hedgehog I'm Not Dead Yet ken and barbie
Creative	Cheap date

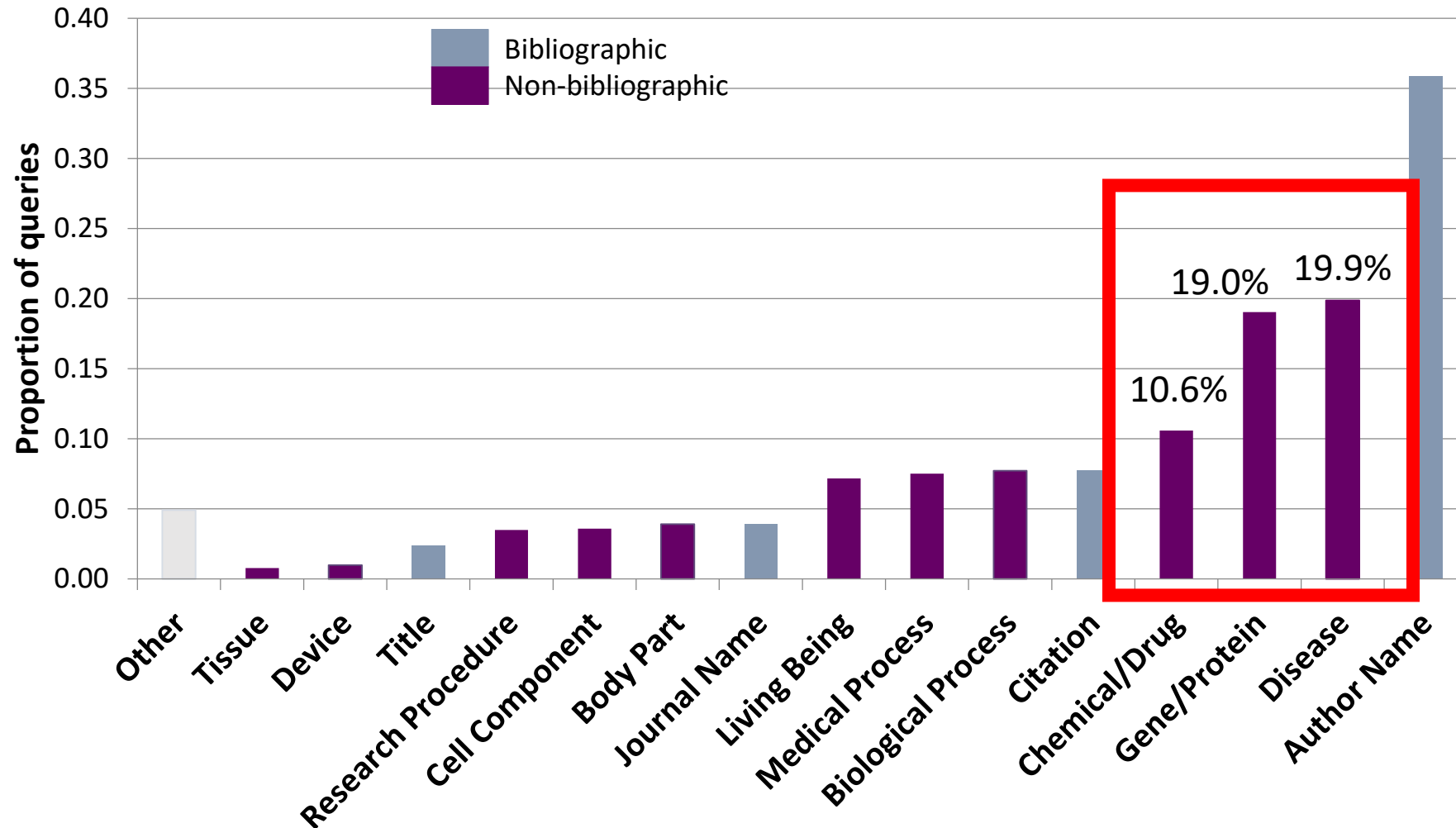
Challenge: phrase variation

Mention Text	Concept name (MeSH/OMIM ID)
bipolar affective disorder	Bipolar disorder (D001714)
immunodeficiency disease	Immunological deficiency syndrome (D007153)
colon carcinoma	Colon cancer (D003110)
anaemia	Anemia (D000740)
pharungitis [sic]	Pharyngitis (D10612)
oral cleft	Cleft lip (D002971)
asthmatic	Asthma (D001249)
absence of functional C7	C7 deficiency (OMIM:610102)
widening of the vestibular aqueduct	Dilated vestibular aqueduct (OMIM:600791)

Challenge: ambiguity

Mention Text	Analysis
THE	English article or gene name?
White	Color or gene name?
founder	Horse disease or creator?
HD	HD gene or Huntington Disease?
P50	Human: NFKB1, CD40, or ARHGEF7?
kaliotoxin	Polypeptide: protein or chemical?
Zinc finger protein	Not anatomy, maybe not zinc
Acute Coronary Syndrome	“Acute” part of name, not modifier

Most searched topics in PubMed



Neveol, Dogan, Lu, Semi-automatic semantic annotation of PubMed queries:
A study on quality, efficiency, satisfaction, *Journal of Biomedical Informatics*, 2010

Key entity types

Disease

- diabetes mellitus; DM; type 2 diabetes

Genomic variation

- c.77A>C; c.77A->C; A77C; AC

Gene/Protein

- TP53; tumor protein p53; p53; BCC7; LFS1

Species

- Arabidopsis thaliana; thale-cress; AT

Chemical/Drug

- Aspirin; 2-(Acetyloxy)benzoic Acid; Acetysal

Cell line

- HEK293; 293 cells; human embryonic kidney 293

Our NER tools

Disease	<ul style="list-style-type: none">• TaggerOne: 83.70%
Genomic variation	<ul style="list-style-type: none">• tmVar 2.0: 86.24%
Gene/Protein	<ul style="list-style-type: none">• GNormPlus: 86.70%
Species	<ul style="list-style-type: none">• SR4GN 86.00%
Chemical/Drug	<ul style="list-style-type: none">• TaggerOne: 89.50%
Cell line	<ul style="list-style-type: none">• TaggerOne: 83.10%

- Freely available & open source
- High Performance
- Novel NLP techniques
- BioC format compatible for improved interoperability

All numbers are F1 scores

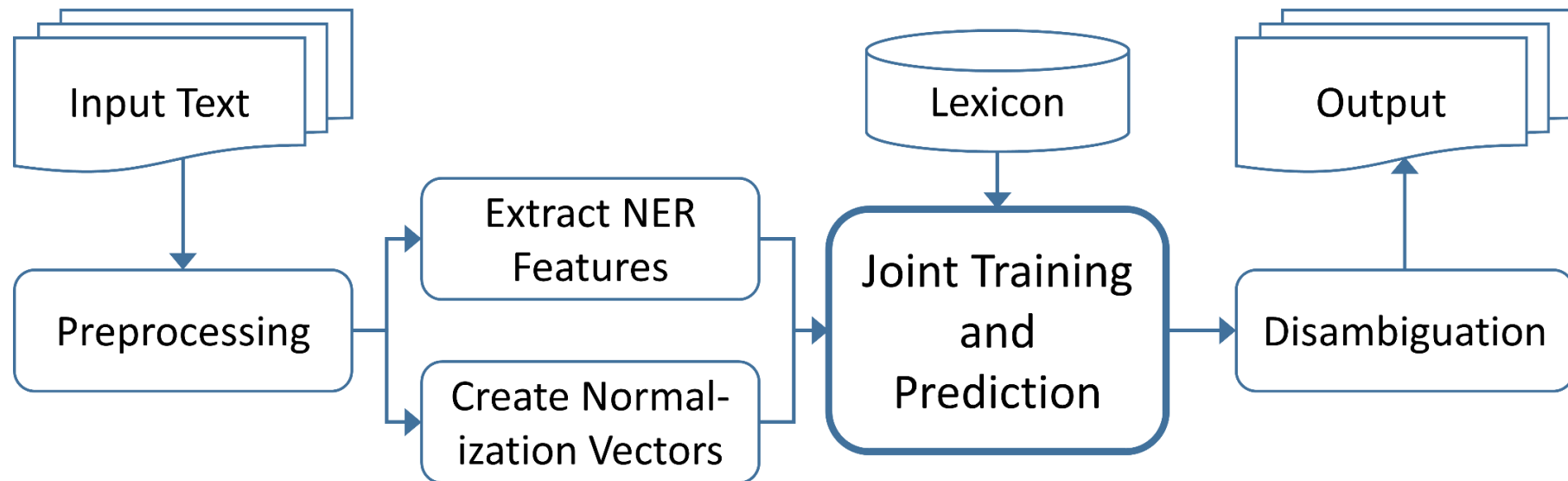
Fundamental methods

- Dictionary based
 - Straightforward, efficient
 - Difficult to find new entities or different variations
- Rule based
 - Can find new entities
 - Rules created manually
 - Adaptation requires system modification
- Machine learning based
 - Can find new entities
 - Learns from examples; needs training data
 - Adaptation requires new training data

Most systems
are hybrids

TaggerOne: joint NER and normalization

- Hypothesis: simultaneous normalization improves NER performance
- NER: rich feature approach
- Normalization score used as a feature in NER scoring

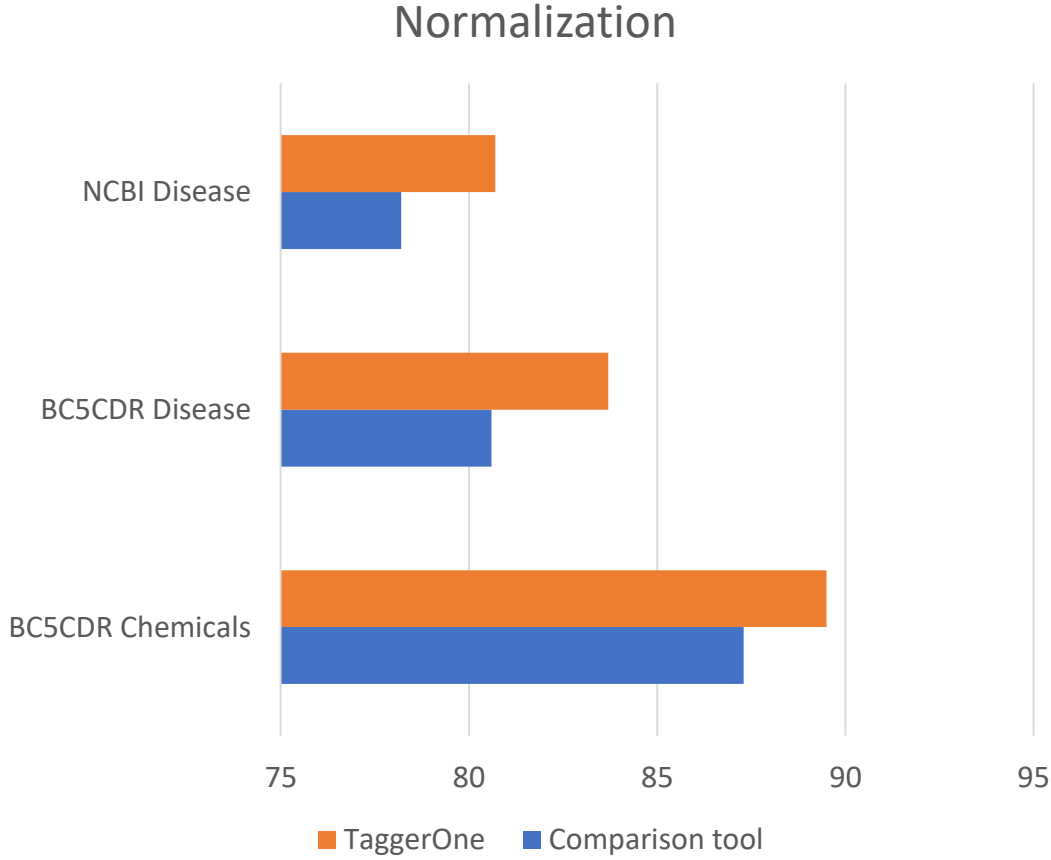
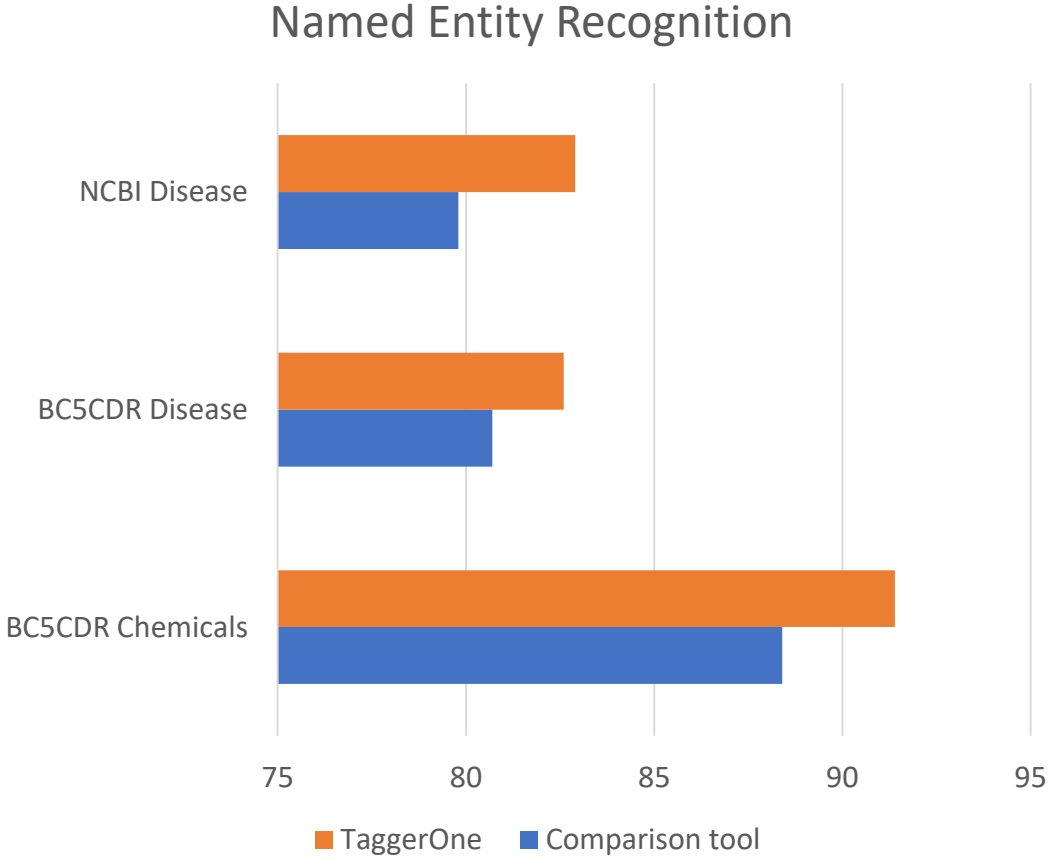


TaggerOne: joint NER and normalization

- Normalization: learns mapping from mention text to concept names

	Gentamicin	sulfate	decreased	renal	function	more	frequently	than	tobramycin	sulfate
Markov model	C	C	D	D	D	O	O	O	C	C
Semi-Markov model	C		D			O	O	O	C	

TaggerOne - results



Multiple resources enrich the lexicon



- Different organization, coverage & granularity
- Example: Hodgkin's Lymphoma
 - MeSH: 1 concept
 - OMIM: 3 concepts (inheritance)
 - UMLS: 7 (histopathology & demographics)
 - OrphaNet: 8 (histopathology)
 - Disease Ontology: 49 (histopathology & anatomical site)

Integrating lexical resources

- Method: use agreement between resources to learn the accuracy of each
- Model: predicted accuracy → expected pairwise agreements
- Training: observed agreement → updated accuracy prediction

Vocabulary added	NCBI Disease	BC5 CDR
+ Disease Ontology	+ 0.0%	+ 1.1%
+ MONDO	- 0.5%	+ 1.7%
+ PharmGKB	+ 1.8%	+ 2.3%
+ probable synonyms	+ 3.7%	+ 7.2%

PubTator

<https://www.ncbi.nlm.nih.gov/research/pubtator/>

- Biomedical concept annotations
 - Genes/proteins, Genetic variants, Diseases, Chemicals, Species, Cell lines
 - New deep-learning based disambiguation
- PubMed abstracts & PMC Text Mining subset
 - Immediately available
 - Daily updates
- Web service: freely available, no installation
 - Wei, Chih-Hsuan, Hung-Yu Kao, and Zhiyong Lu. "PubTator: a web-based text mining tool for assisting biocuration." *Nucleic acids research* 41.W1 (2013): W518-W522.
 - Wei, C.H., Allot, A., Leaman, L. and Lu, Z. "PubTator Central: Automated Concept Annotation for Biomedical Full Text Articles" *Nucleic Acids Research*, *In press*.

PubTator

<https://www.ncbi.nlm.nih.gov/research/pubtator/>

- Online interface
 - Search
 - Visualize
 - Create collections
- RESTful service
- bulk FTP download

The screenshot displays the PubTator Central interface. At the top, the search bar contains the ID '30315184'. Below the search bar, there are dropdown menus for 'group by' (set to 'type') and 'sort by' (set to 'freq'). A search bar is also present. The left sidebar shows a list of concepts categorized by GENE, DISEASE, and CHEMICAL. The main content area displays the title 'Endocrine therapy-resistant breast cancer model cells are inhibited by soybean glyceollin I through Eleanor non-coding RNA' and the abstract text. The abstract text is color-coded to match the concepts in the sidebar. On the right, there is a 'BioConcepts' panel with checkboxes for GENE, DISEASE, CHEMICAL, MUTATION, SPECIES, and CELLLINE. Below this is a table of contents with links to various sections of the paper. A 'BioC XML' button is visible below the title.

PubTator_{Central} 30315184 API TUTORIAL

group by type sort by freq

Search...

GENE
ER (73)
RP18 (7)
GAPDH (5)
ANNEXIN V (5)
MTOR (3)
more

DISEASE
BREAST CANCER (21)
TUMOR (9)
DEATH (3)
INFECTION (2)
BREAST AND OVARIAN TUMORIGENESIS (2)
more

CHEMICAL
GLYCEOLLIN I (95)
ESTROGEN (48)
RESVERATROL (47)

Endocrine therapy-resistant **breast cancer** model cells are inhibited by **soybean glyceollin I** through Eleanor non-coding RNA

PMID30315184 PMC6185934 YAMAMOTO T, SAKAMOTO C ... NAKAO M Sci Rep 2018

full-text

BioC XML

Long-term **estrogen** deprivation (LTED) of an **estrogen receptor (ER) alpha**-positive **breast cancer** cell line recapitulates **cancer** cells that have acquired **estrogen**-independent cell proliferation and endocrine therapy resistance. Previously, we have shown that a cluster of non-coding RNAs, Eleanors (**ESR1** locus enhancing and activating non-coding RNAs) formed RNA cloud and upregulated the **ESR1** gene in the nuclei of LTED cells. Eleanors were inhibited by **resveratrol** through **ER**. Here we prepared another **polyphenol**, **glyceollin I** from stressed **soybeans**, and identified it as a major inhibitor of the Eleanor RNA cloud and **ESR1** mRNA transcription. The inhibition was independent of **ER**, unlike one by **resveratrol**. This was consistent with a distinct tertiary structure of **glyceollin I** for **ER** binding. **Glyceollin I** preferentially inhibited the growth of LTED cells and induced apoptosis. Our results suggest that **glyceollin I** has a novel role in LTED cell inhibition through Eleanors. In other words, LTED cells or endocrine therapy-resistant **breast cancer**

BioConcepts

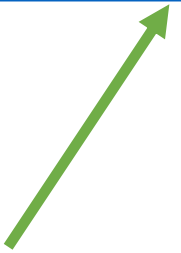
- GENE
- DISEASE
- CHEMICAL
- MUTATION
- SPECIES
- CELLLINE

TITLE
INTRODUCTION
RESULTS
DISCUSSION
METHODS
ELECTRONIC SUPPLEMENTARY M.
ELECTRONIC SUPPLEMENTARY M.
AUTHOR CONTRIBUTIONS
COMPETING INTERESTS

✉

PubTator: RESTful API

[https://www.ncbi.nlm.nih.gov/research/pubtator-api/publications/export/\[Format\]?\[Type\]=\[Identifiers\]&concepts=\[Bioconcepts\]](https://www.ncbi.nlm.nih.gov/research/pubtator-api/publications/export/[Format]?[Type]=[Identifiers]&concepts=[Bioconcepts])



28483577|t|Formoterol and fluticasone propionate combination improves histone deacetylation and anti-inflammatory activities in bronchial epithelial cells exposed to cigarette smoke.
 28483577|a|The addition of long-acting beta2-agonists (LABAs) to corticosteroids improves asthma control. Cigarette smoke exposure increasing oxidative stress may negatively affect corticosteroid responses. The anti-inflammatory effects of formoterol (FO) and fluticasone propionate (FP) in human bronchial epithelial cells exposed to cigarette smoke extracts (CSE) are unknown. The present study provides compelling evidences that FO combined with FP may contribute to revert some processes related to steroid resistance induced by oxidative stress due to cigarette smoke exposure increasing the anti-inflammatory effects of FP.

Formats:

- pubtator
- biocxml
- biocjson

List of PMIDs or PMCIDs:

- pmids=28483577
- pmcids=PMC6207735
- pmids=28483577,28483577,28483577

List of concept types:

- gene, disease, chemical, species, mutation, cellline (optional)

28483577	911	916	HDAC3	Gene	8841
28483577	931	936	HDAC2	Gene	3066
28483577	1009	1013	IL-8	Gene	3576
28483577	1015	1020	TNF-a	Gene	7124
28483577	1022	1027	IL-1b	Gene	3553
28483577	1245	1250	HDAC3	Gene	8841
28483577	1264	1269	HDAC2	Gene	3066

Other tools

- **MetaMap & MetaMap lite: identifies UMLS concepts**

Aronson, Alan R. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001.

Demner-Fushman, Dina, Willie J. Rogers, and Alan R. Aronson. "MetaMap Lite: an evaluation of a new Java implementation of MetaMap." Journal of the American Medical Informatics Association 24.4 (2017): 841-844.

- **cTAKES: framework based on UIMA to build pipeline systems**

Savova, Guergana K., et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." Journal of the American Medical Informatics Association 17.5 (2010): 507-513.

- **Web services: BeCAS and Thalia**

Nunes, Tiago, et al. "BeCAS: biomedical concept recognition services and visualization." Bioinformatics 29.15 (2013): 1915-1916.

Soto, A.J., Przybyła, P. and Ananiadou, S. (2018) Thalia: Semantic search engine for biomedical abstracts. Bioinformatics, bty871

ezTag: interactive annotation <https://eztag.bioqrator.org/>

The screenshot displays the ezTag web interface. At the top, there are navigation tabs for 'ezTag', 'Collections', 'Lexicons', 'Models', and 'Tutorial'. A user ID 'c3a2048ee892' is shown in the top right. Below the navigation is a toolbar with buttons for 'Back', 'BioC Info', 'Download', and 'Demo'. A 'Curatable' toggle switch is also present. The main content area shows a text snippet with several terms highlighted in different colors: 'renal' (pink), 'kidney' (purple), 'urinary' (orange), 'renal tubular epithelium' (green), and 'vasculature' (yellow). To the right of the text is a table of concepts.

Type	Concept ID	Text
Multitissue...	UBERON:000...	vasculature
Organ	UBERON:000...	renal
		kidney
		renal
		renal
		kidney
		renal
Organism...	UBERON:000...	blood
		urinary
		urinary
Tissue	UBERON:000...	urine
		renal tubular epithelium

What and why?

- Information Extraction after NER

Chemical

- Knowledge Summarization

- Digestion of massive information

- Much less costly and less time-consuming

The diagram illustrates a sentence with highlighted words and labels. The sentence is: *Adenine phosphoribosyltransferase plays a role in purine salvage by catalyzing the direct conversion of adenine to adenosine triphosphate.* The words "Adenine phosphoribosyltransferase", "adenine", and "adenosine triphosphate" are highlighted in orange. The word "Adenine" is also highlighted in white. The word "adenine" is labeled as "Gene" below it. The word "adenosine triphosphate" is labeled as "Gene" below it. A purple circle highlights the word "adenine" and the word "adenosine". A question mark is placed to the right of the circle.

What kinds of information do we expect?

- Protein Interaction (e.g. signal transduction)
- Drug Interaction (e.g. side effect using aspirin and warfarin)
- Gene Disease Association (e.g. PARKx and Parkinson's Disease)
- Drug Gene Interaction (e.g. druggable genes)
- Genotype Phenotype Association

Which data resource do we use?

Biomedical Literature

Clinical Notes



Shared Tasks

BioCreative

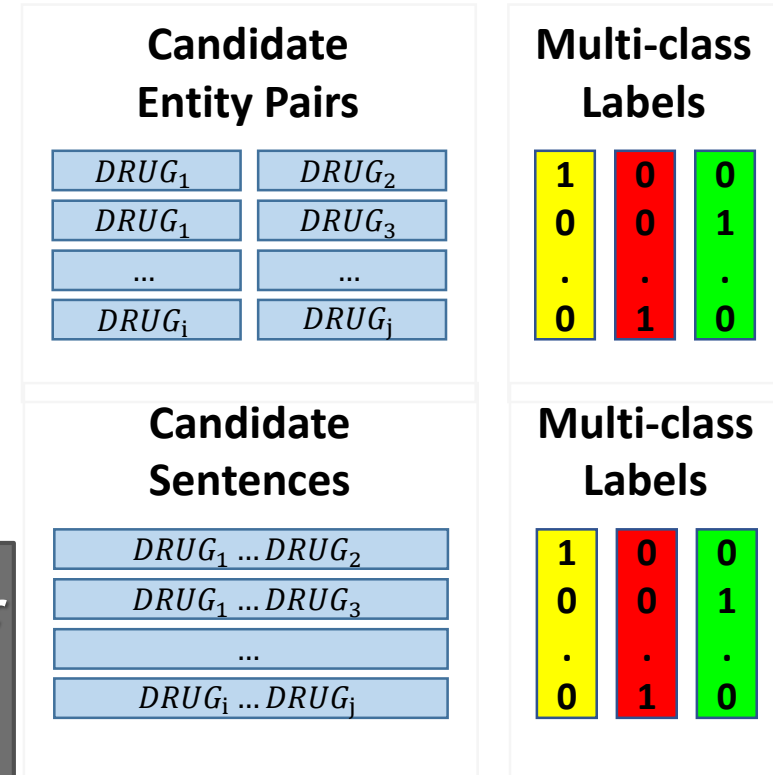
DDIExtraction

BioNLP-ST

i2b2

Problems

- **Pair-wise entities classification**



Fenfluramine may increase slightly the effect of

DRUG₁

antihypertensive drugs, e.g., guanethidine,

DRUG₂

DRUG₃

methyldopa, reserpine.

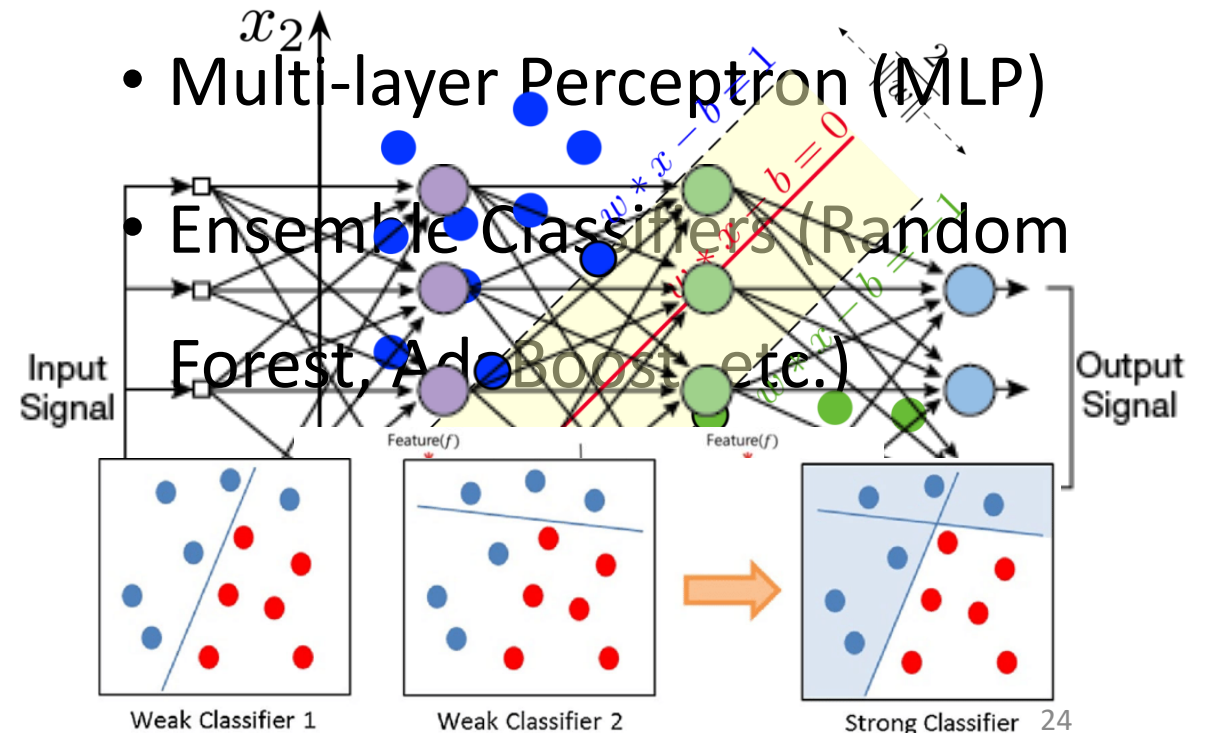
DRUG₄

DRUG₅

Traditional Machine Learning Methods

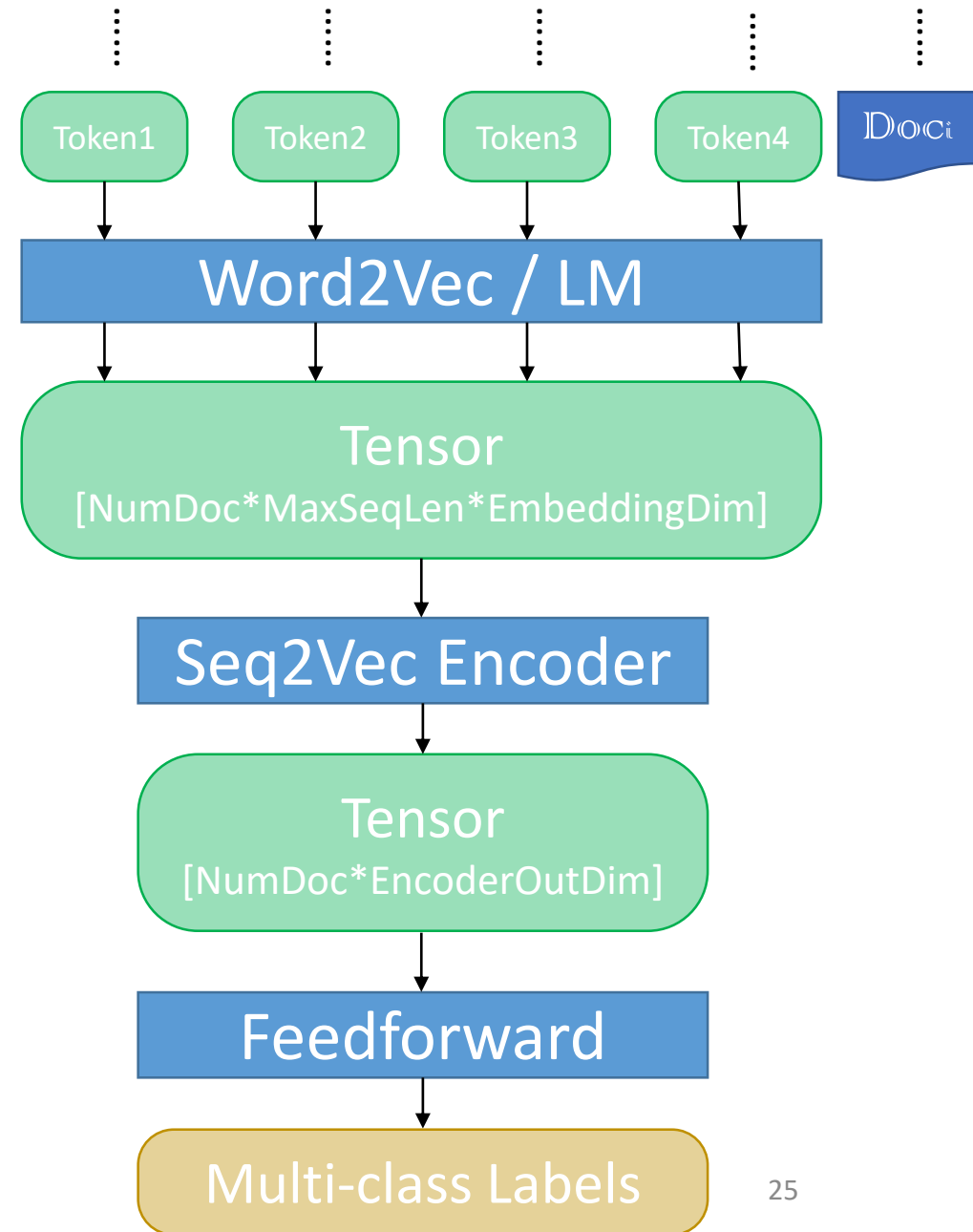
- Handcrafted Features
 - Tokens
 - Part-of-speech (NP, VVP, etc.)
 - Entity type
 - Grammatical function tag (SBJ,OBJ,ADV, etc.)
 - Distance in the parse tree

- Classical ML models
 - Support Vector Machine (SVM)
 - Multi-layer Perceptron (MLP)
 - Ensemble Classifiers (Random Forest, AdaBoost, etc.)



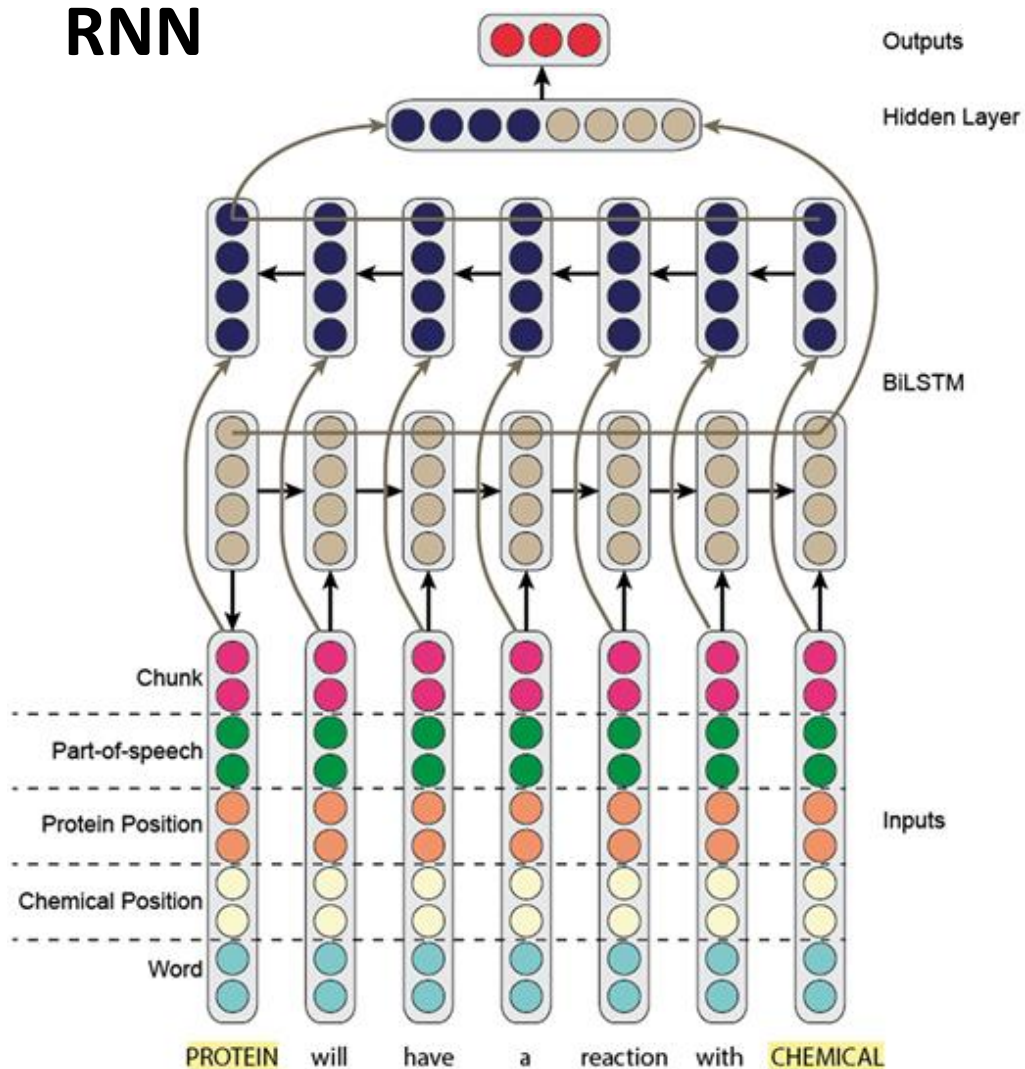
Deep Learning Methods

- Word Embedding (cbow, skipgram, fastText, glove) or Language Model (ELMo, GPT, BERT)
- Sequence to Vector Encoder
 - Bag of Embedding (average or sum)
 - RNN (e.g. LSTM, GRU)
 - CNN
- Classifier:
 - Feedforward Layer
 - Linear Layer

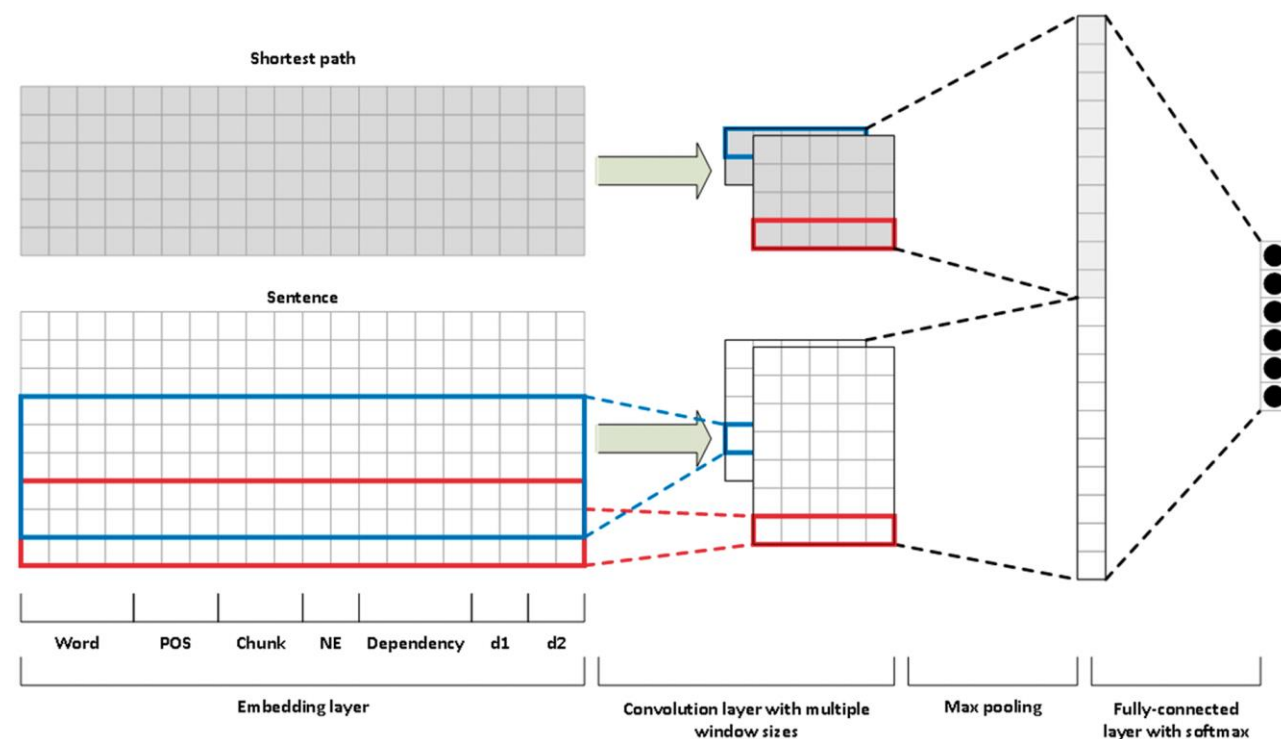


Example for Deep Learning

RNN



CNN



Peng, Yifan, et al. "Extracting chemical-protein relations with ensembles of SVM and deep learning models." *Database* 2018 (2018).

Traditional ML v.s. Deep Learning

Traditional ML

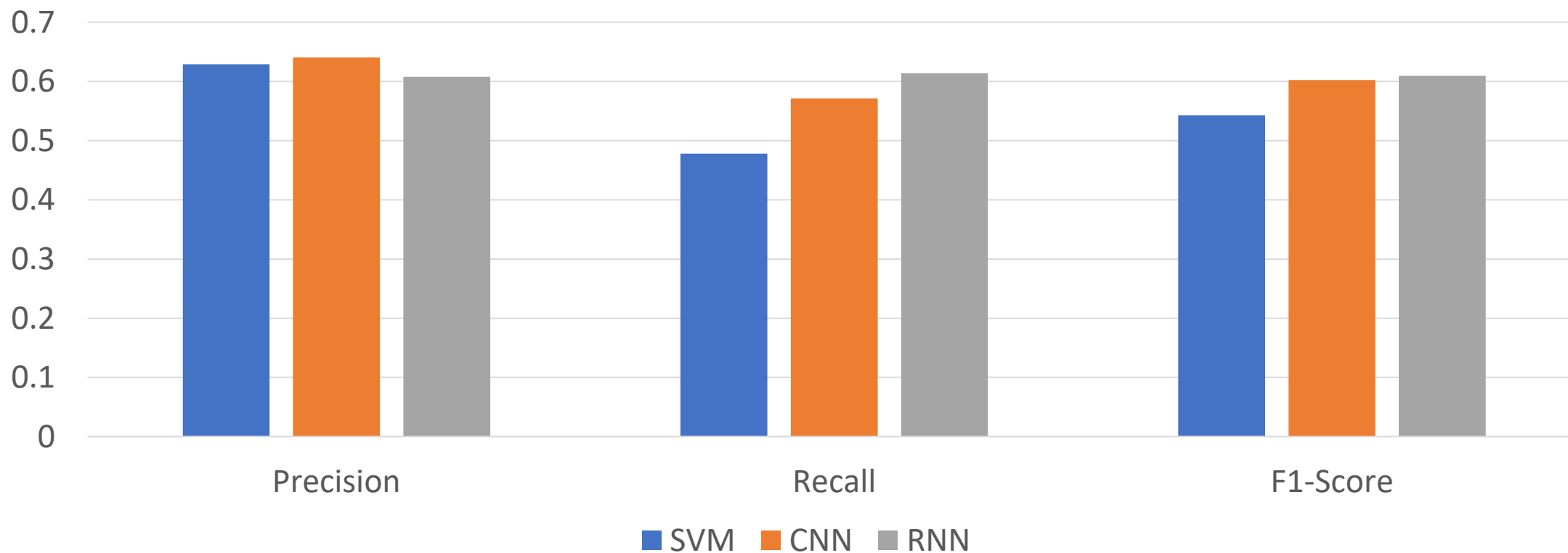
- Hand crafted features
- Simple logic of the methodology
- Computationally efficient (CPU)
- Decent performance

Deep Learning

- Automatic feature extractions
- Complicated architecture
- Require more computations (GPU)
- Improved excellent performance

Traditional ML v.s. Deep Learning

Performance comparison for the ChemProt task at BioCreative VI



Challenges

- Limited Annotations
- Complex Relation Extraction
 - Biomedical event (trigger detection, argument recognition, event prediction)
 - Multiple level event
 - Nesting relationships
- Complex Interaction/Regulation/Association Network

Future Directions

- General relation extraction model
- Clinical relation extraction from electronic health record
- Large-scale complex relation extraction
- Transfer learning

Acknowledgements

- Zhiyong Lu
- Chih-Hsuan Wei
- Alexis Allot
- Rezarta Islamaj
- Dongseop Kwon
- Sun Kim
- Yifan Peng
- Qinyu Cheng



Thank You!