Biomedical Named Entity Recognition

Presenters: Atakan Yüksel &
Batuhan Baykara

What is NER?

Figure 1: An example of NER application on an example text

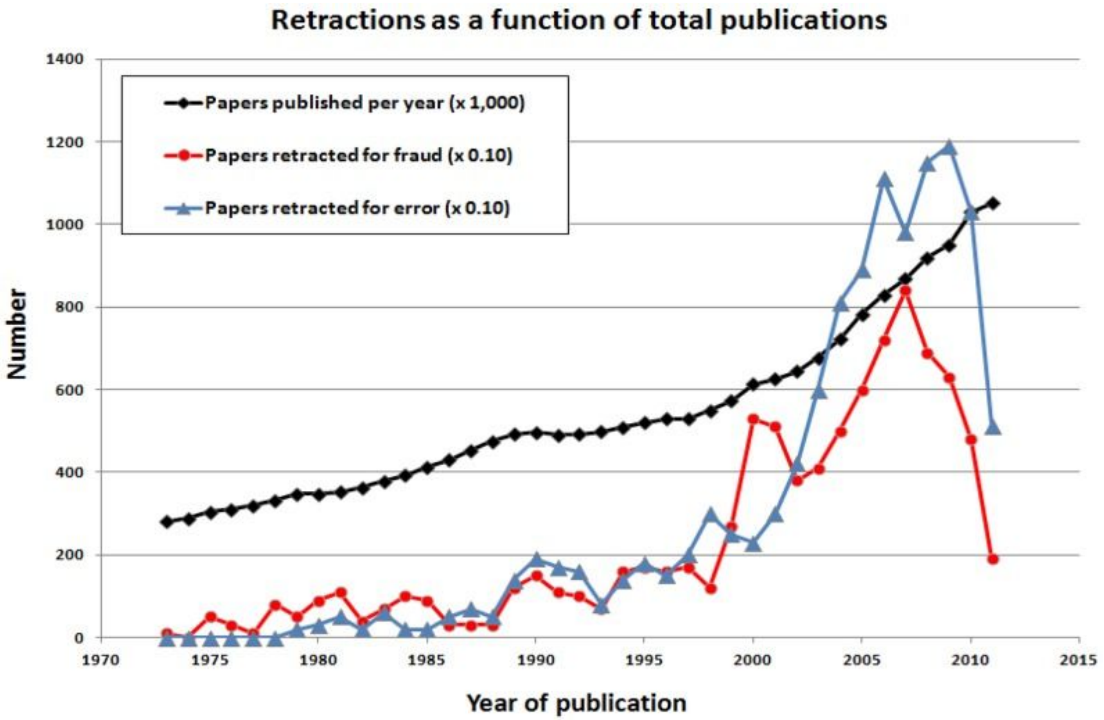
NER In Biomedical

Medline - 20Million papers

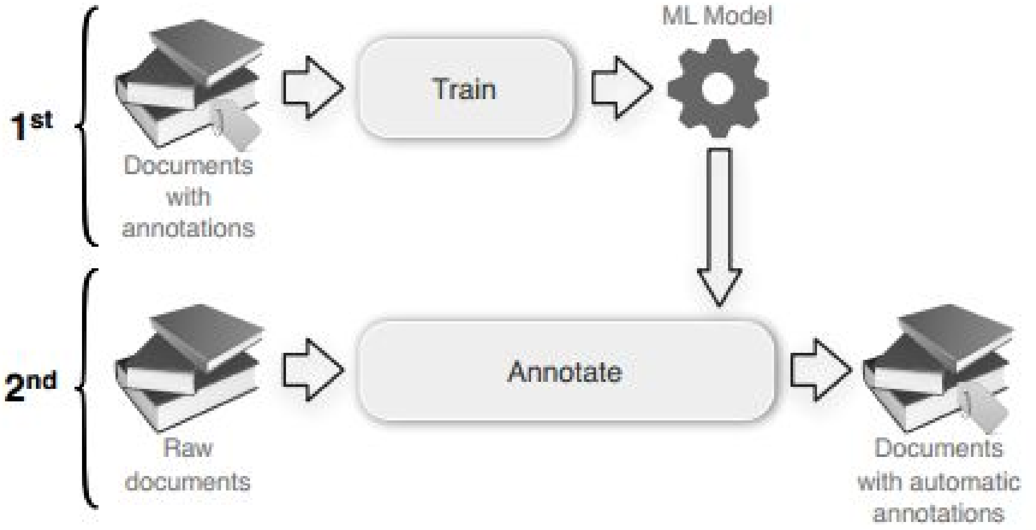
GenBank

“N-acetylcysteine”
“N-acetyl-cysteine”
“NAcetylCysteine”

TCF
T-cell factor
Tissue Culture Fluid



Machine Learning Approach



Corpora

Entity	Corpus	Type	Size (sentences)
Gene and Protein	GENETAG [7]	Sentences	20000
	JNLPBA [6] (from GENIA [8])	Abstracts	22402
	FSUPRGE [9]	Abstracts	≈29447*
	PennBioIE [10]	Abstracts	≈22877*
Species	OrganismTagger Corpus [11]	Full texts	9863
	Linnaeus Corpus [12]	Full texts	19491
Disorders	SCAI Disease [13]	Abstracts	≈3640*
	EBI Disease [14]	Sentences	600
	Arizona Disease (AZDC) [15]	Sentences	2500
	BioText [16]	Abstracts	3655
Chemical	SCAI IUPAC [17]	Sentences	20300
	SCAI General [18]	Sentences	914
Anatomy	AnEM[1]	Sentences	4700
Miscellaneous	CellFinder[2]	Full texts	2100

Corpora Example

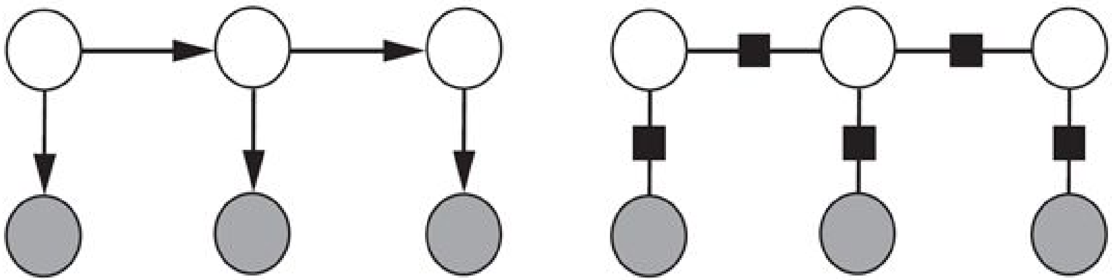
```
### 14008307
### [On trypsin inhibitor activity of amniotic fluid.]
[    9    10   |O
On   10   12   |O
trypsin 13   20   |O
inhibitor    21   30   |O
activity     31   39   |O
of   40   42   |O
amniotic     43   51   |O
fluid    52   57   |O
.    57   58   |O
]    58   59   |O

### 8428048
### Psoriasis and 2,3-biphosphoglycerate blood level.
Psoriasis    8    17   |O
and 18   21   |O
2,3 22   25   |B-IUPAC
-    25   26   |I-IUPAC
biphosphoglycerate  26   44   |I-IUPAC
blood    44   49   |O
level    51   56   |O
.    56   57   |O
```

Machine Learning Methods

CRF

HMM

MEMM



(a) HMM (b) CRF

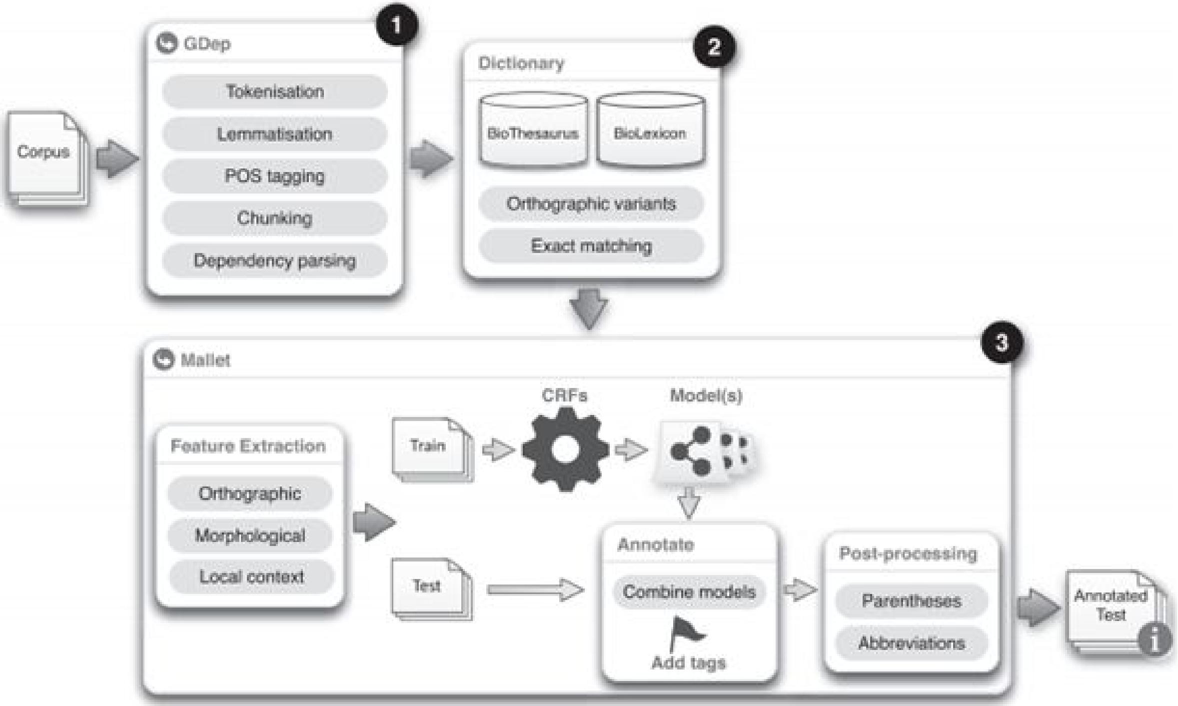
TOOLS

		Open Source								Closed Source				
		2005	2008	2008	2005	2012	2007	2010	2004	2008	2004	2007	2006	2004
		ABNER	BANNER	CBR-Tagger	GENIA Tagger*	Gimli	Lingpipe	NERSuite*	POSBioTM	AIIAGMT	Fin04	IBM Watson	NERBio	Zho04
Reference		[3]	[2]	[32]	[1]	-	[33]	[6]	[4]	[10]	[5]	[7]	[9]	[20]
Programming Language		Java	Java	Java	C++	Java	Java	C++	Java	-	-	-	-	-
Corpora	GENETAG	X	X	X		X	X	X		X		X	X	
	JNLPBA	X			X	X		X	X		X		X	X
Features	Orthographic	X	X		X	X		X	X	X	X	X	X	X
	Morphological	X	X		X	X		X	X	X	X	X	X	X
	Linguistic		X		X	X		X	X	X	X	X	X	X
	Context	X	X		X	X		X		X	X	X	X	
	Lexicons		X			X					X	X		X
Model	CRF	X	X			X		X	X	X			X	
	MEMM				X						X			
	HMM						X							X
	SVM													X
	CBR			X										
	ASO											X		
	Semi-supervised											X		
	Combination					X				X		X		X
Post-Processing	Parentheses		X			X				X		X		
	Abbreviation		X			X								X
	Lexicon									X				
	Pattern-based												X	

*No complete information is available. Extracted from source code analysis.

GIMLI

- GENETAG
- JNLPBA
- CRF



Evaluation of GIMLI

GENETAG

	Protein
P	90.22%
R	84.32%
F1	**87.17%**

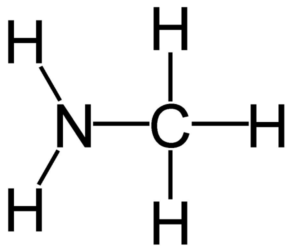
JNLPBA

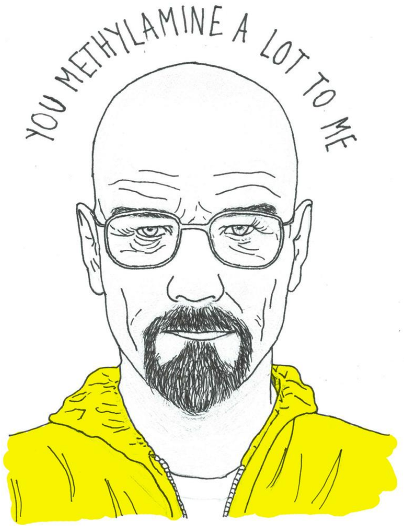
	Protein	DNA	RNA	Cell Type	Cell Line	**Overall**
P	71.53%	74.56%	68.42%	80.44%	61.54%	72.85%
R	78.11%	64.68%	66.10%	62.73%	56.00%	71.62%
F1	74.68%	69.27%	67.24%	70.49%	58.64%	**72.23%**

ChemSpot

- Chemicals can be named in various heterogenous forms.
- Trivial names (e.g. water), brand names (e.g. Voltaren®), (IUPAC) names [e.g. adenosine 3 ,5 -(hydrogen phosphate)], generic or family names (e.g. alcohols), company codes (e.g. ICI204636), molecular formulas (e.g. COOH) and identifiers of various databases.
- Abbreviations introduce a lot of synonyms
- Error prone to; brackets, whitespaces, spelling errors, tokenization errors.
- e.g. methylamine and menthylamine

IUPAC=International Union of Pure and Applied Chemistry





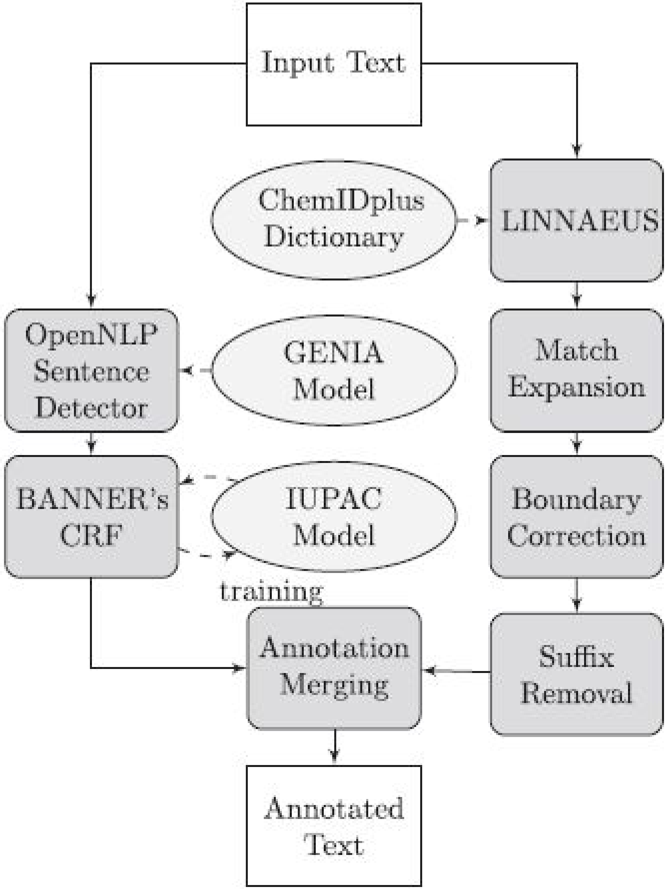
ChemSpot

Hybrid system that uses both CRF and Dictionary

- Cover the different naming conventions for entities commonly subsumed under the term ‘chemical’.
- CRF for IUPAC entities since morphologically more complex than other chemical entities
- Dictionary for brand names, drugs and small molecules since these hardly follow any rule and are best captured by an exhaustive dictionary

CRF

- Sentence boundaries not defined in the corpus
- Better than HMM
- Better than MEMM
  - Label bias problem
- Tagging uses Viterbi Algo.

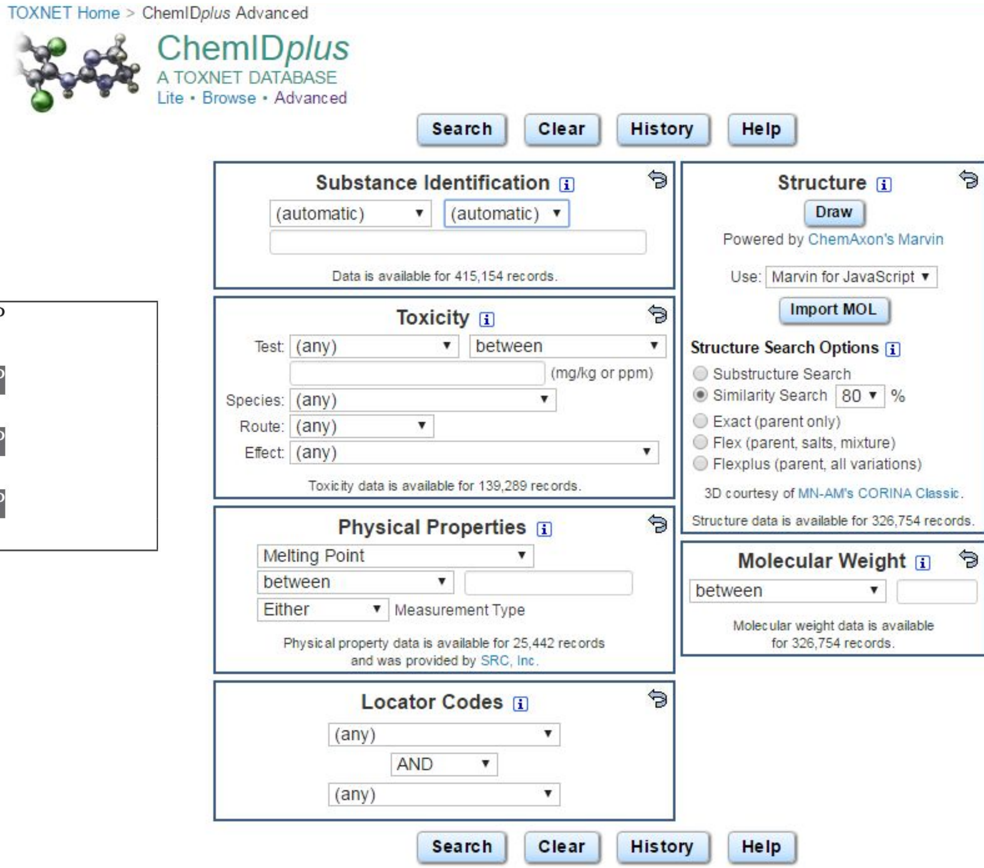


Dictionary

- Search is slow
  - 260 393 concepts
  - 1 378 808 terms
- LINNAEUS
  - Deterministic finite state automata
  - Linear time

Dictionary

Input Text	"…inactivation was slowed by MgATP in the case of N6-CH3-N6-R-ATP [R = (CH2)4N(CH3)CO(CH2)5NHCOCH2I]."
LINNAEUS	"…inactivation was slowed by MgATP in the case of N6-CH3-N6-R-ATP [R = (CH2)4N(CH3)CO(CH2)5NHCOCH2I]."
Match Expansion	"…inactivation was slowed by MgATP in the case of N6-CH3-N6-R-ATP [R = (CH2)4N(CH3)CO(CH2)5NHCOCH2I]."
Boundary Correction	"…inactivation was slowed by MgATP in the case of N6-CH3-N6-R-ATP [R = (CH2)4N(CH3)CO(CH2)5NHCOCH2I]."



Comparison with Other Tools

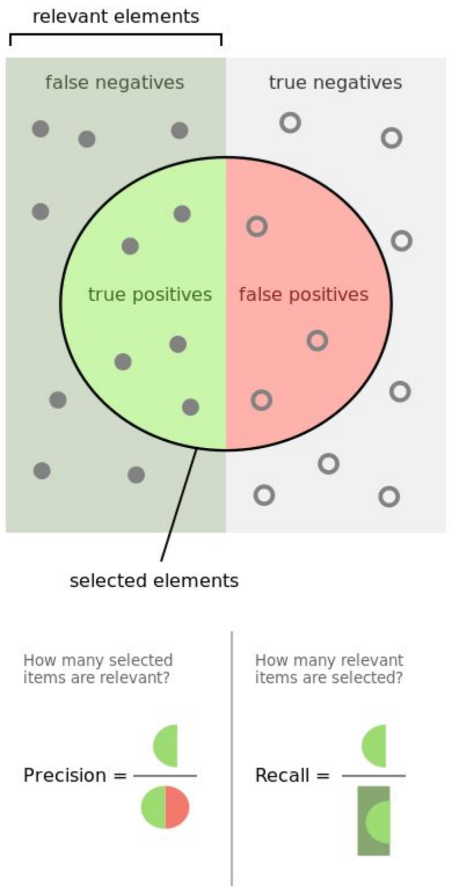
	IUPAC training corpus	IUPAC test corpus			SCAI corpus		
		P	R	F	P	R	F
OSCAR3 (Kolářik *et al.*)					52	72	60
OSCAR3 (Hettne *et al.*)					45	**82**	58
OSCAR3					41.4	81.6	54.9
OSCAR4		2.3	81.5	4.4	45.7	76.5	57.3
CRF (Klinger *et al.*)	X	**86.5**	**84.8**	**85.6**			
CRF (our impl.)	X	61.7	80.1	69.7	**88.3**	28.1	42.6
Dictionary (Hettne *et al.*)					71	37	49
Dictionary (our impl.)					60.8	56	58.3
ChemSpot	X				67.3	68.9	**68.1**

State of the Art

- Gimli => Gene and protein NER
- Chemspot => Chemical, protein and other IUPAC NER
- For statistical approaches
  - CRF > MEMM > HMM
  - HMM - Limited features
  - MEMM - Label bias problem
  - CRF overcomes the problem by a global normalizer
- Deep learning emerged in many fields
- No tools in Biomedical NER yet!

Evaluation

- Data is trained over %80 and tested over %20
- In some cases K-fold cross validation
- Metrics used are;
    - Precision:
    - Recall:
    - F-measure: Harmonic mean of precision & recall



Usecases

- Relation extraction
  - Protein to protein (PPI)
    - “The distribution of actin filaments is altered by profilin overexpression,” the interaction between protein entities “actin” and “profilin” would be extracted
  - Some other interactions gene/disease, protein/chemical
  - Helps scientist in drug development
- Classification
- Topic modeling

Conclusion

- Important part of NLP
- Essential for real world tasks and medicine development
- CRF is mostly used
- Room for improvement - deep learning ?

Thank you for listening…