Maschinelle Sprachverarbeitung

Named Entity Recognition

Ulf Leser

Content of this Lecture

- Named Entity Recognition
  - Dictionary-based approaches
  - Rule-based approaches
  - ML-based approaches
- Named Entity Normalization
- Case studies

Information Extraction: What we need to do

Z-100 is an arabinomannan extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of Z-100 on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived macrophages (MDMs) are investigated in this paper. In MDMs, Z-100 markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. Z-100 was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the env gene is defective and the nef gene is replaced with the firefly luciferase gene) when this vector was transfected directly into MDMs. These findings suggest that Z-100 inhibits virus replication, mainly at HIV-1 transcription. However, Z-100 also downregulated expression of the cell surface receptors CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that Z-100 induced IFN-beta production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (C/EBP) beta transcription factor that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling pathway was involved in Z-100-induced repression of HIV-1 replication in MDMs. These findings suggest that Z-100 might be a useful immunomodulator for control of HIV-1 infection.

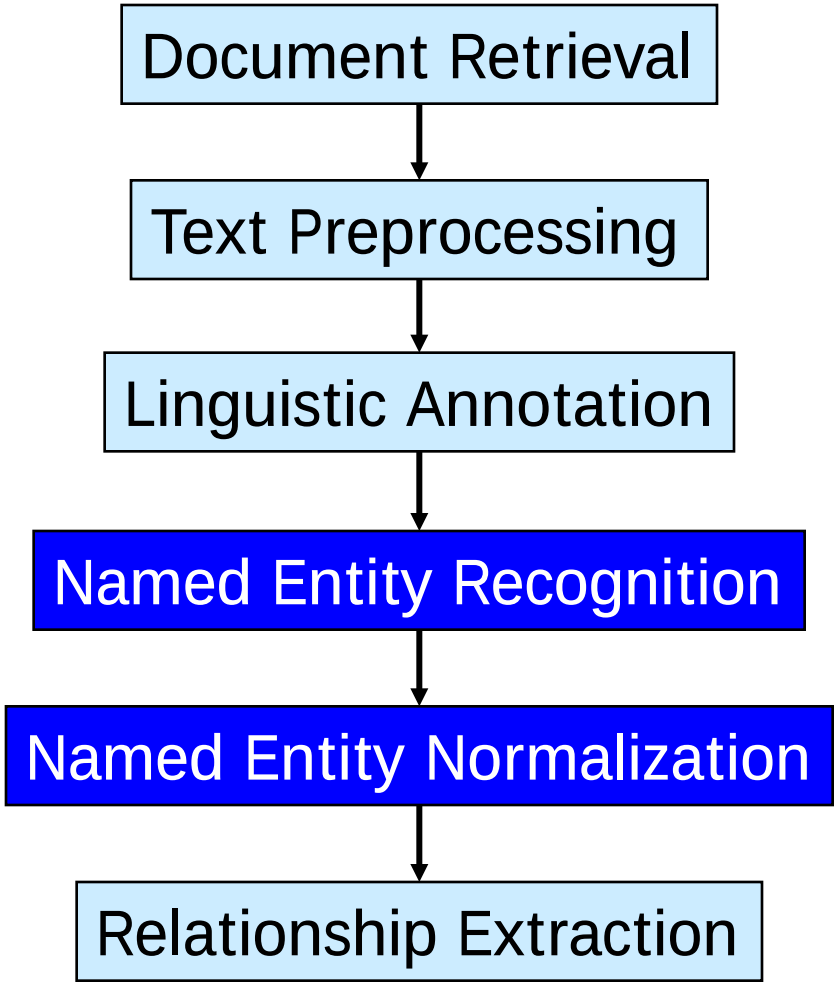
Find Entity Names (Multiple Classes)

Z-100 is an ***arabinomannan*** extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokines. The effects of ***Z-100*** on human immunodeficiency virus type 1 (HIV-1) replication in human **monocyte-derived macrophages** (**MDMs**) are investigated in this paper. In **MDMs**, ***Z-100*** markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. ***Z-100*** was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defective and the ***nef*** gene is replaced with the ***firefly luciferase*** gene) when this vector was transfected directly into **MDMs**. These findings suggest that ***Z-100*** inhibits virus replication, mainly at HIV-1 transcription. However, ***Z-100*** also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that ***Z-100*** induced **IFN-beta** production in these cells, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in ***Z-100***-induced repression of HIV-1 replication in **MDMs**. These findings suggest that ***Z-100*** might be a useful immunomodulator for control of HIV-1 infection.

Relationship Extraction (RE)

Z-100 is an *arabinomannan* extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokines. The effects of *Z-100* on human immunodeficiency virus type 1 (HIV-1) replication in human **monocyte-derived macrophages** (**MDMs**) are investigated in this paper. In **MDMs**, *Z-100* markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. *Z-100* was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defective and the ***nef*** gene is replaced with the ***firefly luciferase*** gene) when this vector was transfected directly into **MDMs**. These findings suggest that *Z-100* inhibits virus replication, mainly at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that *Z-100* induced **IFN-beta** production in these cells, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in *Z-100*-induced repression of HIV-1 replication in **MDMs**. These findings suggest that *Z-100* might be a useful immunomodulator for control of HIV-1 infection.

Information Extraction Workflow



Named Entity Recognition (NER)

- Task: Find all mentions of a given type of entities in a text
  - Genes, diseases, companies, persons, parties, ...
  - Different levels of granularity: Molecular entities, genes, mRNA, exons, human genes, genes implicated in cancer, ...
  - Entities with a fuzzy definition: Earthquakes, symptoms, temporal expressions, relative directions, ...
- Difficulties
  - Set of all entities often not known
  - Spelling variations and spelling errors
  - Entity names may span more than one token (also non-continuous)
- Does usually not include referential mentions
  - Relative pronouns

Examples

- High plasma AVP levels observed in the two cases suggest that SSRIs stimulate AVP secretion, thereby causing SIADH
- A Drosophila shc gene product is implicated in signaling by the DER receptor tyrosine kinase.
- The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.
- The tumor necrosis factor alpha and dkfzp779b086 bind to the human mono-adp-ribosyltransferase.

Examples

- High plasma AVP levels observed in the two cases suggest that SSRIs stimulate AVP secretion, thereby causing SIADH
  - Requires domain knowledge
- A Drosophila shc gene product is implicated in signaling by the DER receptor tyrosine kinase.
  - Has to deal with ambiguities (context is important)
- The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.
  - Sometimes has no clear answer (borders)
- The tumor necrosis factor alpha and dkfzp779b086 bind to the human mono-adp-ribosyltransferase.
  - May use very specific words or consist of rather common words

Some Funny Gene Names

- Dickkopf, zerknüllt, Spätzle
- a (Entrez Gene 43852)
- Lush (40136); (Protein mediates responses to alcohols)
- Van gogh (35922) (Have swirling wing-hair patterns)
- Wish
- Soul
- the
- ...
- Obviously, all of these are homonyms
- Often, a gene, the caused disease, and the mutation share the same name

Abbreviations

- ACE
  - angiotensin converting enzyme
  - affinity capillary electrophoresis
  - Acetylcholinesterase
  - ACE I, a nephrotoxic drug
  - ACE (Anevrysme de l'aorte abdominale: Chirurgie versus Endoprothese)
  - acetosyringone
  - Addenbrooke's cognitive examination
  - Dirección Médica de Fundació ACE
- >60 definitions for ACE in Wikipedia
- Study says: 80% of all acronyms in Medline are not unique

Related Topics

- Single-class (e.g. all genes or all diseases) or multi-class (e.g. all genes and all diseases, ...)
  - Multi-class NER requires disambiguation for mentions which could be both classes
  - E.g. “The company Thomas Cook was named after Thomas Cook”
- Word Sense Disambiguation (WSD)
  - Often, tokens (or sets of tokens) can be of multiple classes
    - Bass can be a fish or an instrument
  - WSD: Assign an entity in a text to its correct semantic class (sense)

Content of this Lecture

- Named Entity Recognition
  - Dictionary-based approaches
  - Rule-based approaches
  - ML-based approaches
- Named Entity Normalization
- Case studies

Dictionary-Based NER

- Gazetteer or dictionary
  - A gazetteer originally is a list of geographic names with locations
  - In TM, a gazetteer is a list of names
- Dictionary-based NER (for single token entities)
  - Build a dictionary of all names of entities you are interested in
    - Dictionaries usually include synonyms
  - Match every token in the text against the dictionary
- Important: Include fuzzy matches

Dynamic Domains

- Can we always build a dictionary of all entities of a class?
  - Finding all street names in Berlin is relatively simple
  - Finding all geographic locations is more difficult
    - Places, buildings, hills, woods, ...
  - Finding all person names in Germany is even more difficult
    - New persons are born all the time
      - Mostly new combinations of known first / last names
    - New names immigrate all the time
    - Other languages are much more innovative with names (initials, J.R: junior, Schewarnadze (son), Saakaschwili (child), Hadschi Halef Omar Ben Hadschi Abul Abbas Ibn Hadschi Dawuhd al Gossarah, ...
  - Finding all company names is even more difficult
    - Companies are created and closed all the time
    - No real naming conventions (Remember the ".com" phase)
    - Often with fixed elements (GmbH, AG, inc., ...)

Funny First Names [Berliner Zeitung, 2008]

- Regulations in Germany: „Die Schreibweise ist den Regeln der Rechtschreibung unterworfen. Biblische Namen mit negativer Assoziation wie Judas oder Kain sind nicht erlaubt, ebenso wenig Markennamen, die nicht mit Vornamen identisch sind, Adelstitel, Orts- und Städtenamen. Also nichts mit Arizona, Sierra Nevada oder Schweinfurt. Ausnahmen wie Mercedes, Paris und San Diego bestätigen allerdings die Regel. Außerdem muss der Vorname das Geschlecht erkennen lassen, weshalb ein Kind namens Kim einen zweiten Vornamen braucht.“
  - Internationale Promis hätten in Deutschland schlechte Karten. Ist der Name von Nicole Kidmans Tochter Sunday Rose weiblich? Nein, der Sonntag ist so männlich wie Freitag aus Robinson Crusoe. Und was ist mit Gwyneth Paltrows Tochter Apple? Im Deutschen wäre es der Apfel ... Da wir schon mal beim Obst sind: Eine Lehrerin in Neuseeland heißt Cherry. Kirsche. Immerhin: die Kirsche. Auch viele Frauen namens Fern gibt es im Land des Silberfarns. Und ganz im Trend der handy- und SMS-süchtigen jungen Generation kamen im vergangenen Jahr reichlich Knaben namens JJ, C, CJ, T, TJ und AJ auf die Welt. Die weibliche Antwort darauf ist Tequila. Zur besseren Verdauung aller schwer verdaulichen Vornamen.
- Genehmigt: Pepsi-Carola, Napoleon, Rasputin, Rapunzel, Sunshine, Sonne
- Abgelehnt: Möwe, Porsche, Pfefferminze, Lenin, Crazy Horse, Störenfried

Example: Gene Names

- Finding all gene names is really hard
  - New genes are found or genes are re-discovered all the time
  - Definition of a gene is not clear at all (splicing, miRNA, ...)
  - Difference between gene, transcripts, encoded proteins not clear
  - No (successful) naming convention
    - Discoverer, disease, location, phenotype, species, cell type, ...
  - Much “legacy” text which is only a couple of years old
  - Frequent use of abbreviations
  - Use of common English words (hedgehog, Dickkopf, soul, ...)
  - Highly distributed creation process, no central repository
    - Contrast: There are regional “repositories” for company names

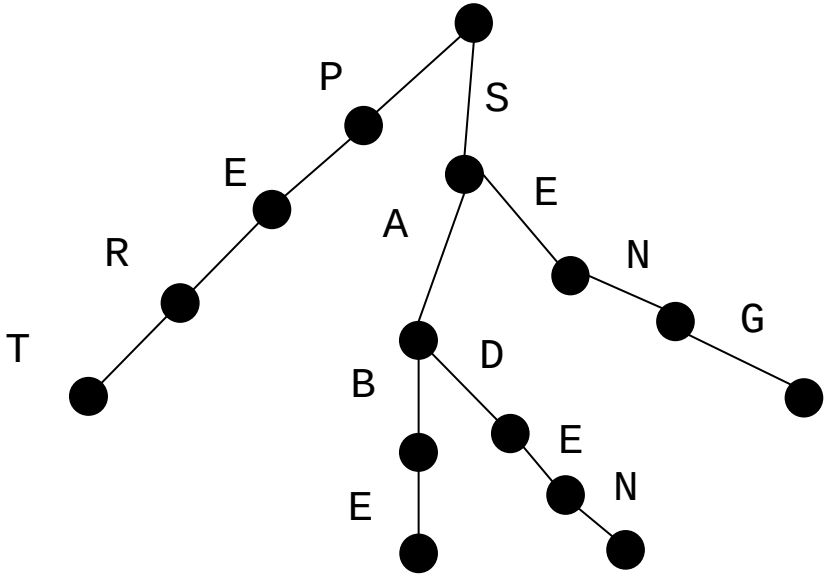
Fuzzy Search

- Even in static areas, names need not appear exactly
  - Yahoo, yahoo, Yahoo!, yahoo.com, yaho (typo), ...
  - Die Geissens, die Geissen's, die Geissen`s, die Geißens, ...
- Solution: Fuzzy or approximate matching
  - Solution 1: Generate a "fuzzified" dictionary
  - Solution 2: Use similarity-based string matching algorithm

Dictionary-Based NER: Exact Matching

- Exact matching: Prefix trees
  - Build a prefix tree of all dict. entries
  - Search each token in the tree
  - For a token of length m, this requires O(m) char comparisons
  - But requires lots of space

P={PERT,
SABE,
SADEN,
SENG}



Dictionary-Based NER: Similarity for Single Token

- Hamming distance: Fast but not appropriate for human language terms
- Edit distance: Slow, should be length-normalized, should have different weights for individual symbol operations
  - Meier – Maier, Tobel – Hobel (distance 1)
  - Tor-Kur, Schifffahrt-Schifffahrten (distance 2)
  - Requires O(m*n) char comparisons (n: length of dict entry)
  - Much research in efficient index structures
- Jaccard-distance over k-grams: Faster, lower bound for edit distance, good for longer token
- Grammar-inspired heuristics: remove “s”, remove “ed”, ...
- Domain-specific; e.g., gene names: Remove ‘ or -

Dictionary-Based NER: Multiple Token

- Prefix-tree: Index all token of all entries
- Move a sliding window over the tokenized text
  - Window length: Difficult! Length of longest dict entry?
  - Match all token of text in prefix tree
  - Compute bipartite matching of matched token of entry with matched token in mention
    - This can be tricky if token have multiple potential matches
    - Bipartite matching: $O(n^3)$ (if n is length of window)
  - Aggregate scores of individual matches to a window-entry score
    - GO terms: "*Negative regulation of anterior neural cell fate commitment of the neural plate by fibroblast growth factor receptor signaling pathway*"
    - "Gesetz zur effektiveren und praxistauglicheren Ausgestaltung des Strafverfahrens"
    - "Gesetz zur effektiveren und praxistauglicheren Ausgestaltung von Strafverfahren"
    - "Gesetz zur effektiven Ausgestaltung von Strafverfahren"
    - „Wir haben ein Gesetz erlassen, dass Strafverfahren beschleunigen soll"
  - Further clues: Distance of token, number of unmatched token, order of token, containment in noun phrases, ...

Properties of Dictionary-Based NER

- Advantages: Simple, fast, can easily include NEN
  - Typical baseline system
  - Easiest solution, lay persons use it as synonym for NER
- Well suited for static (closed) entity types
  - Problems with fuzzy matching and ambiguous names remain
- For dynamic classes
  - Performance depends on dictionary size, level of ambiguity, ...
  - Usually one expects high precision
    - A match should is correct (provided appropriate configuration)
    - But dictionary-based NER usually disregards context
    - Ambiguous names deteriorate precision
  - ... at rather low recall (incomplete dictionaries)

Content of this Lecture

- Named Entity Recognition
  - Dictionary-based approaches
  - Rule-based approaches
  - ML-based approaches
- Named Entity Normalization
- Case studies

Rule-Based Systems

- Define rules that capture indications for of a NE
  - Combine context words, POS tags, surface properties, ...
    - [PERSON] earns [MONEY] USD
    - [PERSON] join* [ORGANIZATION]
    - the [PROTEIN]/NNS receptor
- Potentially very labor-intensive approach
- Typical trade-off
  - Long, precise rules: Very good precision, low recall
  - Short, general rules: Bad precision, good recall
- Often used in combination, e.g., use ML-based NER and rules for post-processing (filtering false positives)
- Somewhat old-fashioned, but ...

Rule-Based or Machine-Learning-Based?

Rule-based Information Extraction is Dead!
Long Live Rule-based Information Extraction Systems!

Laura Chiticariu
IBM Research - Almaden
San Jose, CA
chiti@us.ibm.com

Yunyao Li
IBM Research - Almaden
San Jose, CA
yunyaoli@us.ibm.com

Frederick R. Reiss
IBM Research - Almaden
San Jose, CA
frreiss@us.ibm.com

- 90% of NER papers in top-TM conferences use ML
- 80% of commercial tools and projects are rule-based
- Rule-based: Adaptable, controllable, understandable

Learning Rules

- Rules can be learnt from gold standard corpora
- Learn characteristics of the searched entities
  - Context words, suffixes, position in sentence, ...
  - That appear frequently around positive instances
  - That appear rarely elsewhere
- Rule abstraction is vital
  - Word at this position? Around this position? Word like this?
  - Must be a verb-pasttense-1stperson-sg; verb-1stperson-sg; ...
- Requires very large GS-corpora

Content of this Lecture

- Named Entity Recognition
  - Dictionary-based approaches
  - Rule-based approaches
  - Machine Learning-based approaches
    - NER as classification
    - Sequential tagging: HMMs, MEMMs, CRFs
- Named Entity Normalization
- Case studies

Classification-Based NER

- Classify each token as entity or not
  - Learn model based on manually annotated training text
- Advantages
  - Usually high quality results, but problems with multi-token names
  - Recognizes unseen entities (provided a proper feature set)
  - We “only” need a corpus, learning is automatic
  - Implicitly performs weak form of WSD
    - If context is encoded as features
- Disadvantages
  - Often slow (depends on ML-method)
  - Needs large amount of high-quality training data
  - Requires additional NEN step

Typical Features

- [Biased towards gene / protein name recognition]
- Surface features
  - The word itself – how often at start of / within entity name?
  - Character uni-, bi-, tri-grams
  - POS tag
  - Length
  - Specific properties (to be defined manually)
    - Has capital letters, all capital letters, more capital letters than non-cap
    - Has Greek/Roman letters, special characters, digits, all digits
      - 3'-mRNA, 5-alpha-reductase, EST94F88G, ...
    - Abstraction: is of class DDUU, DDSS, DDCDD, ...
      - Digits, small case letter, upper case letter, special characters, ...
      - Max include contraction: 1.999.000,99 -> D.DDD.DDD,DD -> D.D.D,D
  - ...

More Features

- Context features
  - POS tag of surrounding tokens
  - NER tag of preceding tokens
  - Presence of indicator words within a certain distance
    - Protein, human, enzyme, plasma, ...
- External knowledge
  - Token (or closed-by tokens) matches in a dictionary
- Memory
  - Most frequent tag for this token in texts
  - Most frequent tag for surrounding tokens in corpus
- Others (creativity!)
  - E.g. Number of matches in Google versus PubMed

Classifiers and Ensembles

- Popular choice: SVM / Maximum Entropy
- Ensembles: Use different classifiers and vote
- Example results

Classification	LOC	MISC	ORG	PER
MxE24[1]	77.81	57.49	78.83	85.41
TMB24	75.49	53.19	77.44	83.89
MxE25	78.27	58.22	78.64	85.60
TMB25[2]	75.15	52.94	77.79	85.36
HMM[3]	71.15	45.69	72.95	70.20
Voting[1,2,3]	78.46	57.00	78.93	86.52

Source: Kozareva, JRC Workshop, 2005

Content of this Lecture

- Named Entity Recognition
  - Dictionary-based approaches
  - Rule-based approaches
  - Machine Learning-based approached
    - NER as classification
    - Sequential tagging: HMMs, MEMMs, CRFs
- Named Entity Normalization
- Case studies

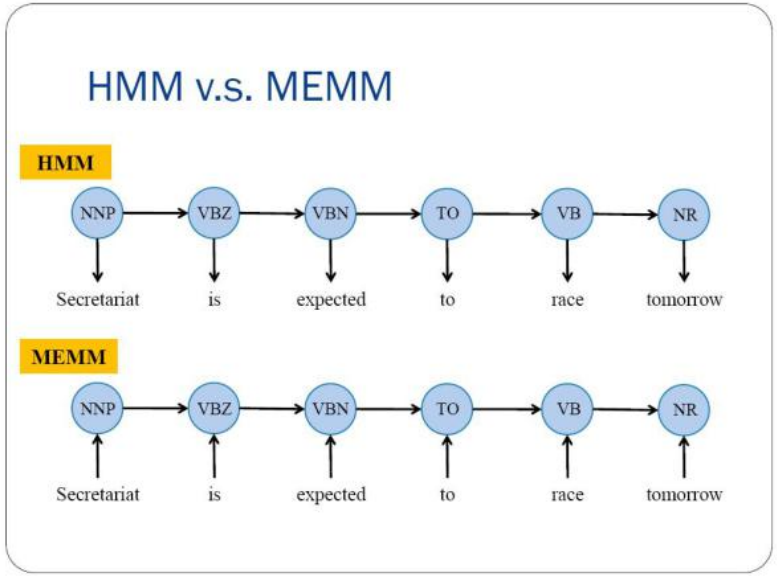
Sequential Tagging using HMMs

- Recall POS tagging with HMMs
  - Fix a set of classes (POS tags)
  - Learn probabilities as state transitions and emissions
  - Encode as Hidden Markov Model
  - Given a new text, find most probable sequence of tags (Viterbi)
- Can readily be applied to NER – with proper tag set
  - Very popular: IOB (Begin of name, In a name, Other)
- But: Using only tag sequence not enough for high quality
  - Too coarse-grained (only three classes)
  - Need to look at the words and their features, not just their tags

MEMM: Maximum Entropy Markov Models (sketch)

- HMMs are generative models (like Naive Bayes)
- MEMM: A discriminative sequential classifier
  - We predict output (e.g. IOB) from sequential observations (token)
  - MEMM model only transition probabilities, but conditional on the observations which are represented using feature functions
    - Feature functions are derived from the observation
    - May take tokens “as is” or use abstractions
      - “is a noun”, “has capital letter”, ...
  - ME principle to learn conditional transition probabilities is applied separately for each transition from a state q to all next states
    - High-order models are possible
  - Training: GIS algorithm for each state as in ME classification
  - Decoding: Variation of Viterbi algorithm

Visual Explanation

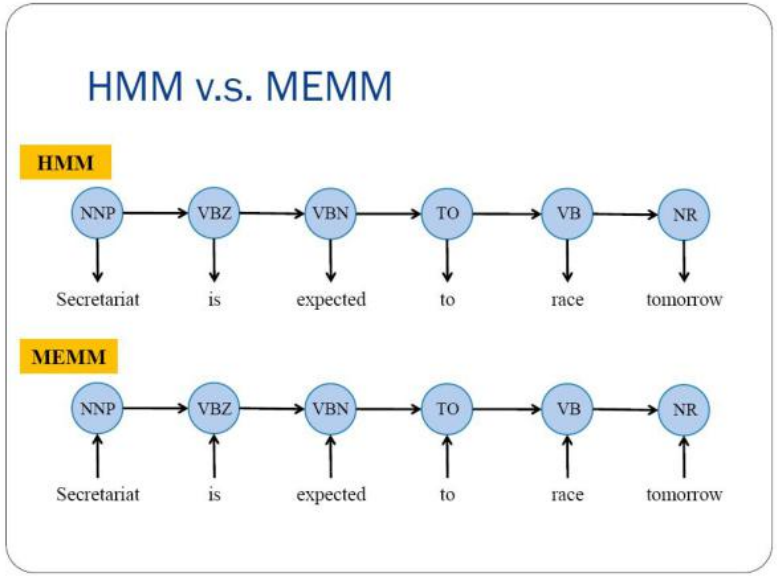


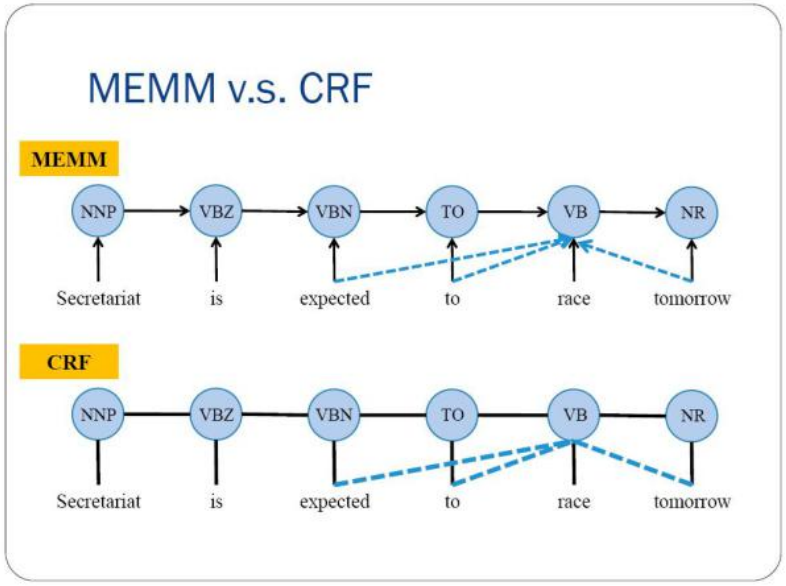
Source: https://liqiangguo.wordpress.com/2011/04/18/

Conditional Random Fields (sketch)

- MEMM suffer from Label Bias Problem: Transitions from labels with few successor states get higher probabilities and thus dominate inference
  - Because outgoing probabilities must sum to 1 in each state
  - MEMM is a local model (in each state)
- CRF are global models and directly estimate p(Y|X) over the entire sequence of labels Y and observations X
  - Lafferty, McCallum, Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data.„. Technical Report, Upenn (2001).
  - Transitions probabilities may depend on future observations and future states – all combinations are considered during inference
- Decoding is simple (Viterbi), learning is complex

Visual Explanation

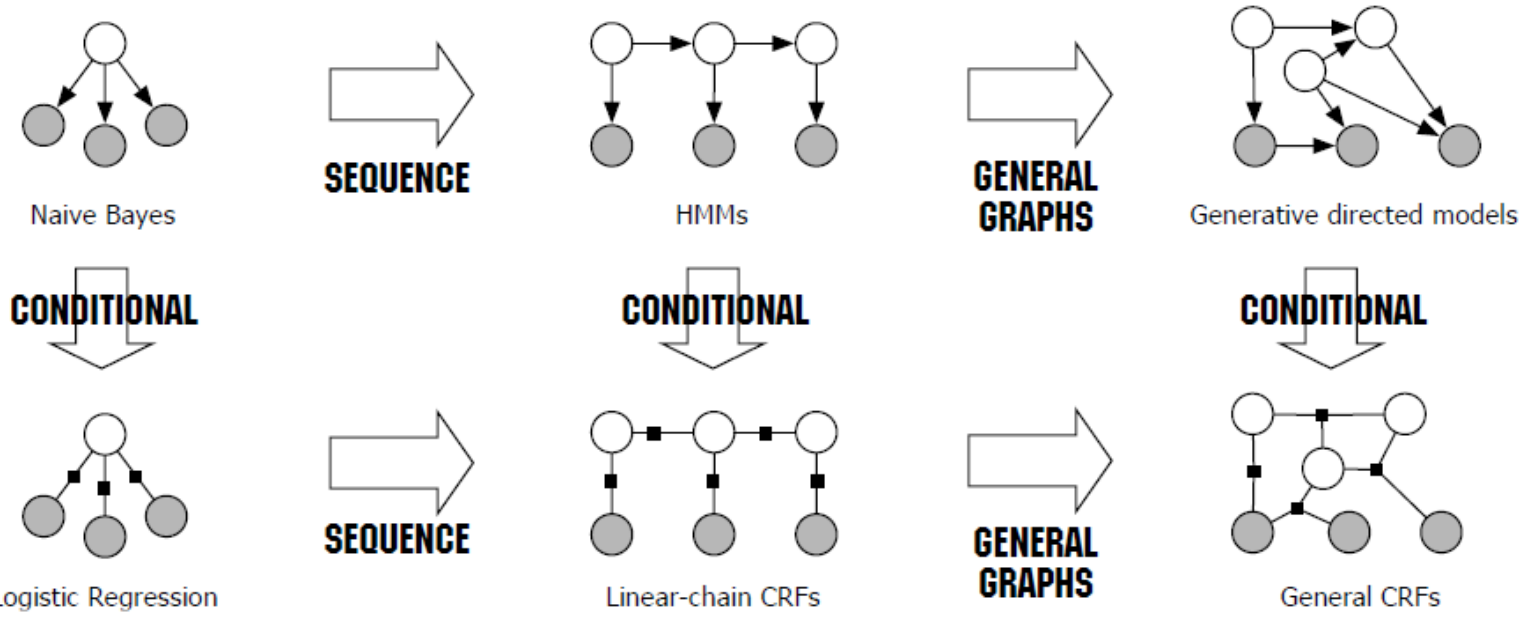




Source: https://liqiangguo.wordpress.com/2011/04/18/

Linear-Chain CRF

- General CRF may condition on every state in the sequence
- Linear-chain CRF restrict the scope of features to those of the surrounding states to make inference more efficient
  - And to make learning less data-demanding



Sutton, McCallum (2011): An Introduction to Conditional Random Fields

Comparison

	HMM	MEMM	CRF
Type	Generative	Discriminative	Discriminative
Model	Local	Local	Global
Decoding method	Viterbi-style	Viterbi-style	Viterbi-style
Independence assumption (token-next state)	Yes	No	No
Arbitrary feature functions	No (difficult)	Yes	Yes
Label bias problem	Yes	Yes	No
Learning	Fast	Fast	Slow
Decoding	Fast	Fast	Fast

Content of this Lecture

- Named Entity Recognition
- Named Entity Normalization
- Case studies

Named Entity Normalization (NEN)

- "It is a gene – but which gene?"
- NEN maps each entity to a canonical ID
  - Highly domain/application specific
  - Coordinates of geo-locations, DB-IDs of genes, passport-numbers of persons, ISBN for books, Orchid-ID for researchers etc.
  - What is "canonical" requires consensus (NCBI gene, ensembl, uniprot, ...)
- Necessary to link recognized entities to further information (data integration)
  - NER without NEN has very few practical applications

NEN Algorithms

- Typical approach: Given a mention, find the most similar term in a dictionary of all names of this entity type
- Same methods as for dictionary-based NER
- But we have to choose a dictionary entry – no thresholds

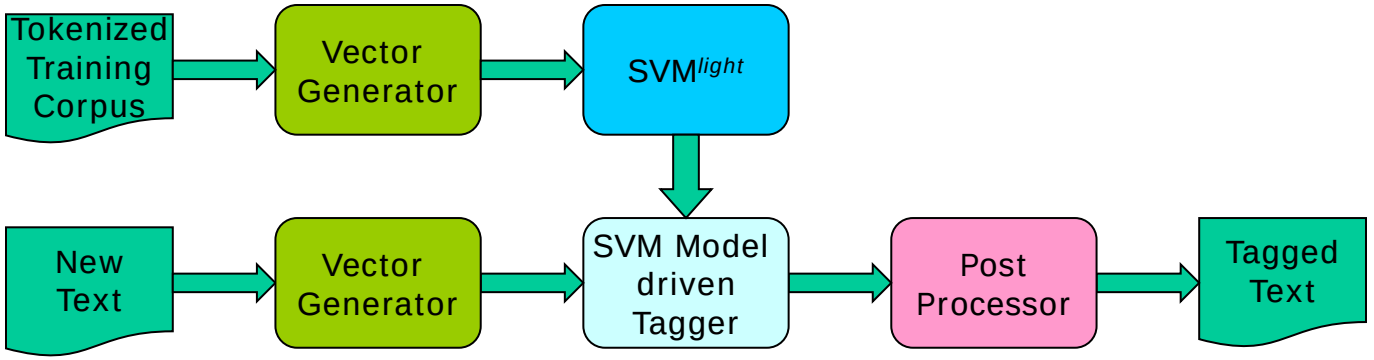
Content of this Lecture

- Named Entity Recognition
- Named Entity Normalization
- Case studies
  - BioCreative
  - MUC conferences
  - Predicting ICD-10 codes

BioCreative Cup 2004

- Critical Assessment of Information Extraction Systems in Biology
- International competition, three tasks
- Training data and evaluation script provided by organizers in cooperation with database curators (Swiss-Prot)
- Test data available for one week
- Evaluation of all submissions by (published) scripts
- Major boost: Top systems reached 84 F1-measure
  - Previous best systems around 60 F1-Measure
  - Possibly not much further improvements since then
  - Fields splits up: Species, NER/NEN, NER/PPI, ...

Example: SVM for NER



- Corpus of 7500 sentences
  - 140.000 non-gene words
- SVM^{light} on different feature sets
- Dictionary compiled from Genbank, HUGO, MGD, YDB
- Post-processing for compound gene names

Features

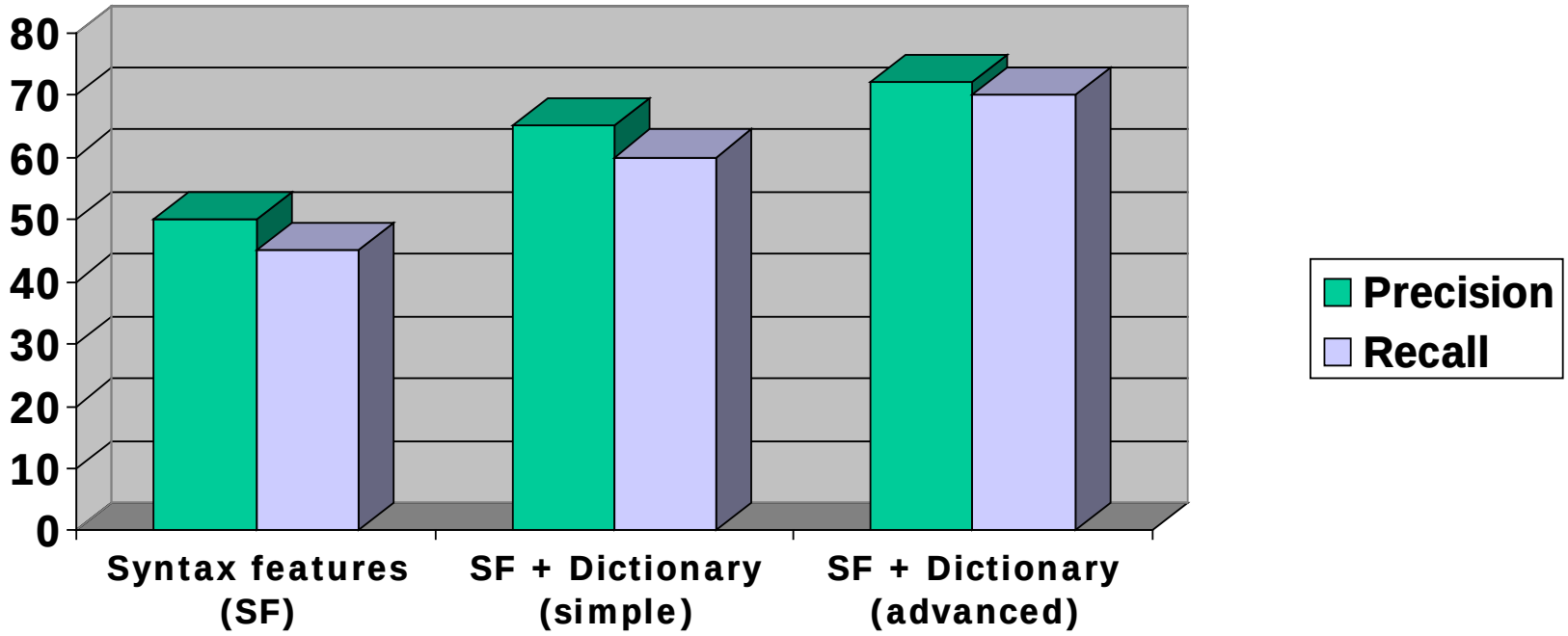
Feature	Weight	Example
Word	tf * idf	kinase
n-grams		
N=1	tf * idf	k, i, n, a, s, e
N=2	tf * idf	ki, in, na, as, se
N=3	tf * idf	kin, ina, nas, ase
Special signs		
HasNumbers	[1|0]	p300
HasCapitals	[1|0]	abLIM
AllCaps	[1|0]	DMD
InitCap	[1|0]	Pax
HasNumbers & Letters	[1|0]	cMOAT2, EST90757
Context		
predecessing word	[1|0]	Gene
succeeding word	[1|0]	Product
distance to keywords	1/(1+dist)	(list of 15)
Dictionary		
Word match	[1|0]	
Phrase match	[1|0]	

Post-processing

- SVM detects only single token candidates
- Most gene names are multi-token names
- Expand detected single-token genes based on set of heuristic rules (found in an unsystematic manner)

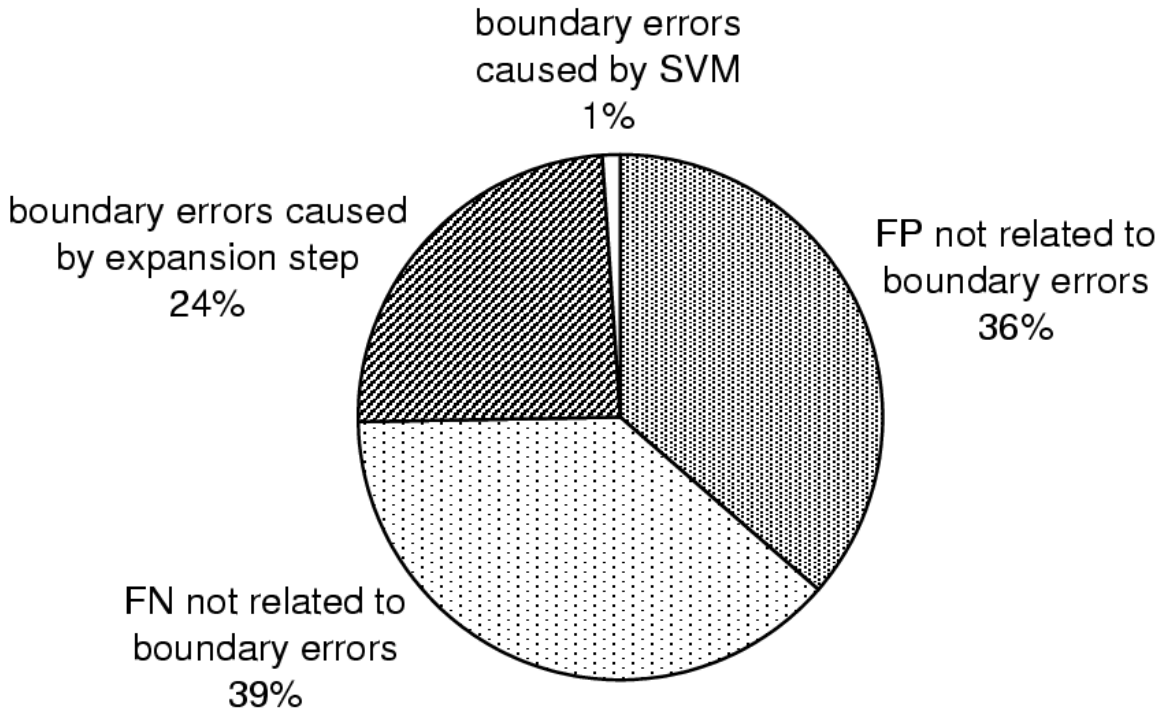
GENE NN*	➔	GENE GENE
NN* GENE	➔	GENE GENE
GENE (NN)	➔	GENE (GENE)
GENE protein	➔	GENE GENE
GENE ADJ GENE	➔	GENE GENE GENE

Performance



- Best result for BioCreative Cup: 73 F-measure
  - 12 percentage point increase by post-processing only
- Raises from 73 to 83 for loose evaluation

Where did we Fail?



- „Boundary error“ – problems with multi-word phrases
- >70% of errors are token classification errors from SVM

Impact of Feature Classes

Feature	Example	Short name	Impact		
Token *	Sro7	Token	=54%	- *baseline* -	
Unseen token *		UToken			
n-grams of token *		1G, 2G,	+15%	1..4grams, P+, R++	
			+14%	1..3-grams	
Previous & next tokens		P/NToke	-5%	[1,1]-window, P+, R-	
			-6%	[-2,2]-window	
n-grams of tokens in window		2PG/2NG..			
Prefixes, suffixes		1P,2P,3P,1..	±0		
Stop word	the, or	Stop	-5%	10.0..0 words, P+, R-	
			-1%	1.0..0 words, P+, R-	
			-.5%	10.. words, P+, R-	
POS ta	NN DT	POS	-50%	P, R-	
Initial					
All cha					
Upper					
Upper					
Single					
Two ca					
Capital					
Lower					
Special					
Charac					
Numbe					
Letters					
Digit,					
Greek					
Roman					
Number followed by '%' ○	75.0 %	percentag	-.1%		P-, R-
DNA, RNA sequences ○	ACCGT	DNA, RN	-.1%		P-, R-
Longest consonant chain *	Sro7→ 2	LCC	-2%		P-, R-
Keyword distance *		keyDist	-20%		P+, R-
Gazetteer *		Gaz	-3%		P-, R-
Prev./next token is NEWGENE		PTG, NT	-18%	pr. only, P+, R-	
Tokens + letter surface clues			+2%		P+, R-
Tokens + 1,2,3-grams + greek + roman + letter surface clues			+14%		P+, R++
Tokens + 1,2,3-grams + keyDist + Gaz + LCC + special + combi + allCaps + initCap *			+ 16%		P+, R++
Tokens + 1,2,3,4-grams + keyDist + Gaz + LCC + special + combi + allCaps + initCap + lowMix ○			+ 18%		P+, R++

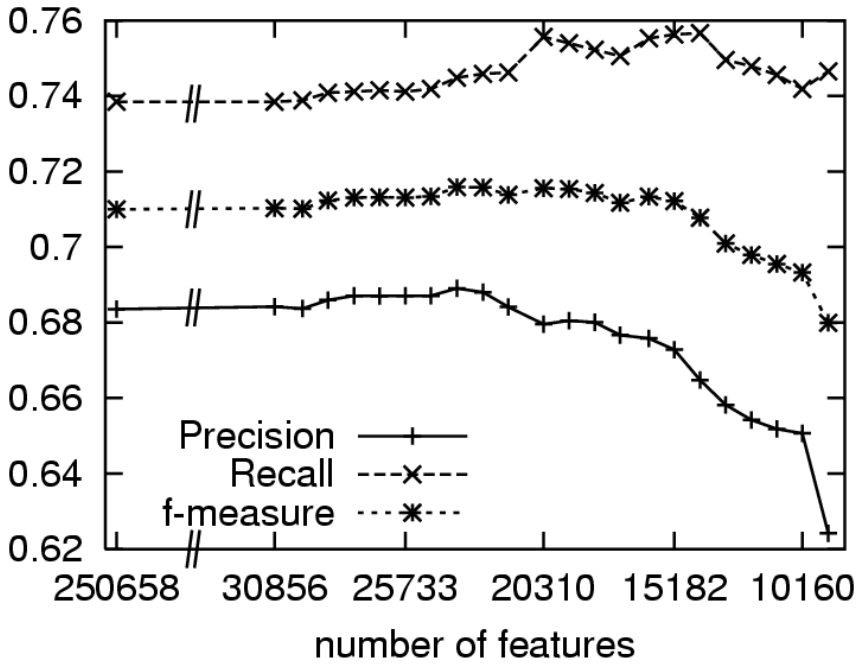
Tokens + letter surface clues +2% P+, R-

Tokens + 1,2,3-grams + greek + roman + letter surface clues +14% P+, R++

Tokens + 1,2,3-grams + keyDist + Gaz + LCC + special + combi + allCaps + initCap * + 16% P+, R++

Tokens + 1,2,3,4-grams + keyDist + Gaz + LCC + special + combi + allCaps + initCap + lowMix ○ + 18% P+, R++

Do we need them all?



- Repeated elimination of 5% least discriminating features
- Eliminating 95% of features costs only 2% F-Measure

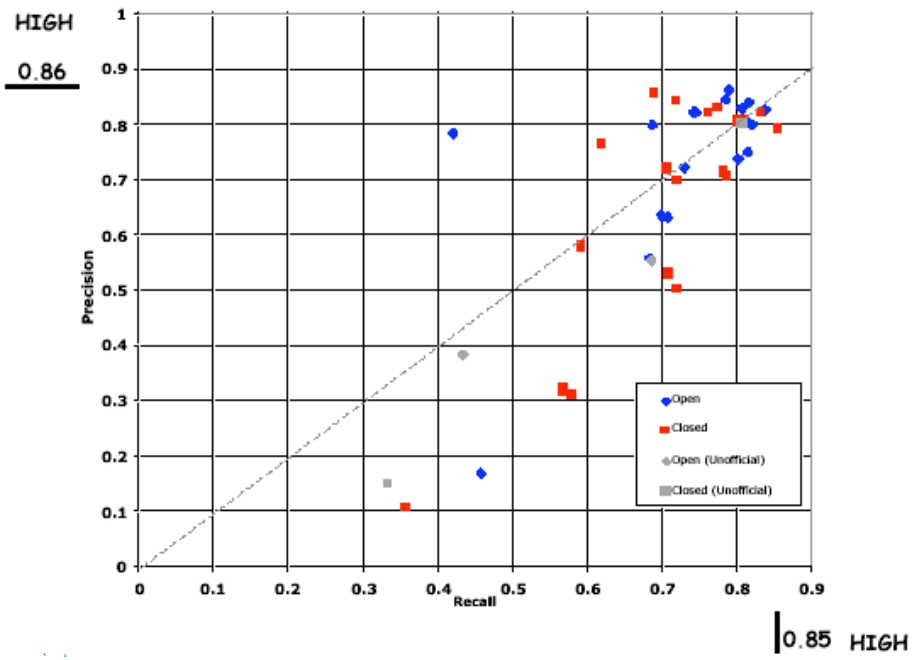
Which Ones?

- Single features from different classes are among the most important ones
- Difficult to remove entire classes of features

Feature	Class	Weight
	Gaz	1.497386
insulin	Token	0.632708
protein	Token	0.628168
kinase	Token	0.608392
human	Token	0.536695
proteins	Token	0.535368
	greek	0.498111
	combi	0.489201
serum	Token	0.480326
	lowerUpper	0.457806
	singleCap	0.438028
factor	Token	0.438028
wild-type	Token	0.389359
	initCaps	0.366269
mutants	Token	0.340689
genes	Token	0.340352
promoter	Token	0.327395
receptor	Token	0.323412
polymerase	Token	0.305972
complex	Token	0.292019
receptors	Token	0.292019
c-myc	Token	0.292019
sites	Token	0.243349
mutant	Token	0.243349
domain	Token	0.231541
sequence	Token	0.216691
sequences	Token	0.216683
domains	Token	0.215116
	specialnumber	0.205077
isoforms	Token	0.194679
	specialupperCase	0.179926
	capMixLetters	0.179394

Other Systems

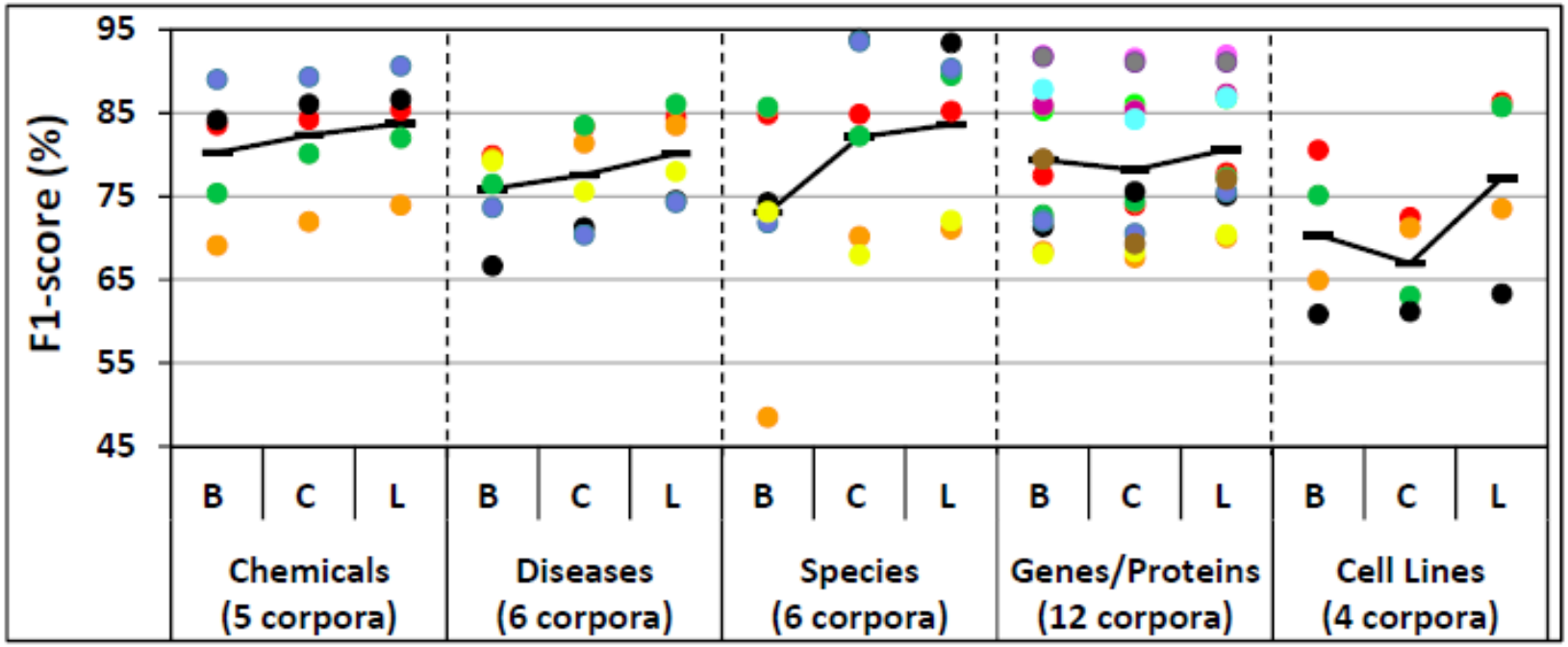
- Best: MMEM or CRF
- Much larger feature sets
- Use of ensembles trained on different corpora
- Current state-of-the-art
  - F-measure ~85%
  - Strongly dependent on eval corpus
  - Inter-annotator agreement assumed at ~90%
  - Loose evaluation reaches >90%
  - Still less than MUC results



Gene-NER: Why is it hard?

- „Scientists would rather share each other‘s underwear than use each other‘s nomenclature“ [Keith Yamamoto]
- Ambiguous gene names and high number of acronyms
  - The, white, ACL, ...
- Small training and eval corpora, mostly only abstracts
- Strict vs. loose matching (up to 20% in F1 difference)
- Generally little agreement on gene names (low IAA)
- Cross-corpus performance
  - All corpora differ in scope
  - Method trained on corpus A performs bad on corpus B
  - Domain Adaptation Problem

State-of-the-Art: LSTM-CRF with Word Embeddings



Content of this Lecture

- Named Entity Recognition
- Dictionary-based approaches
- Rule-based approaches
- ML-based approached
- Case studies
  - BioCreative
  - MUC conferences
  - Predicting ICD-10 codes

Message Understanding Conferences (MUC)

- Large conferences and competitions (1987 – 1998)
- Initiated and funded by DARPA (among other)
- Similar to TREC, but focusing on information extraction / named entity recognition
- Tasks including co-reference resolution
- Template filling / “model-based” IE

Mr. John Smith was appointed CEO of ACME last December 31.

Name:	John Smith
Post:	CEO
Company:	ACME
Date:	December 31

Corpora

Year	Conference	Domain
1987	MUC-I	Navy messages
1989	MUC-II	Navy messages
1991	MUC-3	News about terrorist attacks
1992	MUC-4	News about terrorist attacks
1993	MUC-5	Company news (joint-ventures, micro-electronics production)
1995	MUC-6	Company news (management succession)
1998	MUC-7	Airline company orders

Source: Boullosa, NER

Results (MUC-7, 1998)

Task	Recall (%)	Precision (%)
Named Entity (NE)	92	95
Coreference	63	72
Scenario Template (complete events)	47	70

Systems (MUC-7, 1998)

F-Measure	Error	Recall	Precision
93.39	11	92	95
91.60	14	90	93
90.44	15	89	92
88.80	18	85	93
86.37	22	85	87
85.83	22	83	89
85.31	23	85	86
84.05	26	77	92
83.70	26	79	89
82.61	29	74	93
81.91	28	78	87
77.74	33	76	80
76.43	34	75	78
69.67	44	66	73

Annotators:

97.60	4	98	98
96.95	5	96	98

- Best system is a hybrid between an extensive set of rules and a ME classifier

Context Rule	Assign	Example
Xxxx+ is a? JJ* PROF	PERS	Yuri Gromov is a former director
PERSON-NAME is a? JJ* REL	PERS	John White is beloved brother
Xxxx+, a JJ* PROF,	PERS	White, a retired director,
Xxxx+ ,? whose REL	PERS	Nunberg, whose stepfather
Xxxx+ himself	PERS	White himself
Xxxx+, DD+,	PERS	White, 33,
shares of Xxxx+	ORG	shares of Eagle
PROF of/at/with Xxxx+	ORG	director of Trinity Motors
in/at LOC	LOC	in Washington
Xxxx+ area	LOC	Beribidjan area

Source: Mikheev, Grover, Moens, „DESCRIPTION OF THE LTG SYSTEM USED FOR MUC-7“

Content of this Lecture

- Named Entity Recognition
- Dictionary-based approaches
- Rule-based approaches
- ML-based approached
- Case studies
  - BioCreative
  - MUC conferences
  - Predicting ICD-10 codes (recall from intro)

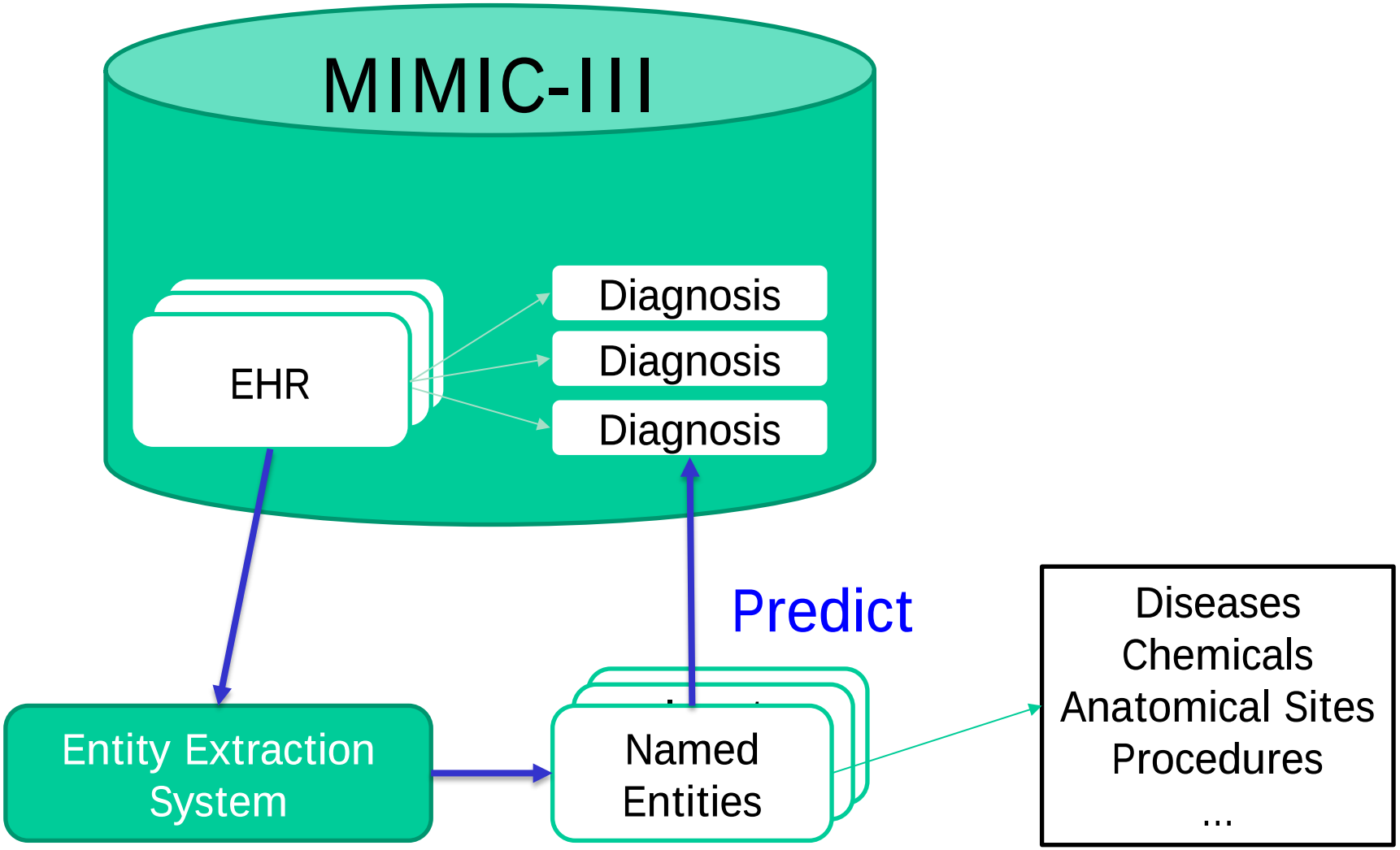
Predicting Disease Codes based on Patient Records

- Medical diagnosis are encoded in fixed vocabularies
  - For accounting, for statistics, for integration, for data mining
- Most important taxonomy: ICD-9/10
  - International Classification of Diseases
  - “codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases”
  - Roughly 15.000 codes in hierarchical organization
  - DRG: German “disease related groups”, derived from ICD-9, used for accounting of medical treatments

Problem

- Proper ICD-10 annotation is vital for any hospital
- Physicians do not use ICD codes for documentation
  - Too clumsy, too many, not precise enough, much relevant information not expressible (temporal development, dosage, ...)
- Currently, a “Medizinischer Dokumentarist” reads EHR’s and adds ICD codes
- Task: Can we automatically predict ICD codes based on medical records?
  - Results here: J. Bräuer, Clinical Entity Recognition for ICD-9 Code Prediction in Clinical Discharge Summaries, Diplomarbeit, 2017

Architecture

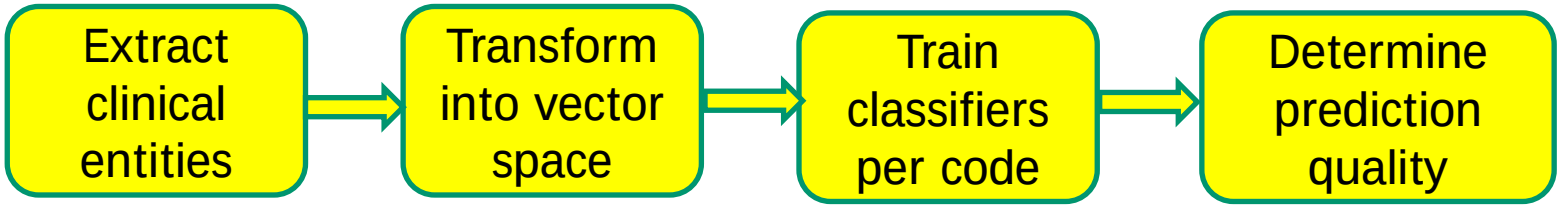


- **DATE OF ADMISSION:** MM/DD/YYYY
- **DATE OF DISCHARGE:** MM/DD/YYYY
- **DISCHARGE DIAGNOSES:**
- 1. Vasovagal syncope, status post fall.
  2. Traumatic arthritis, right knee.
  3. Hypertension.
  6. History of chronic obstructive pulmonary disease.
- **BRIEF HISTORY:** The patient is an (XX)-year-old female with history of previous stroke; hypertension; COPD, stable; renal carcinoma; presenting after a fall and possible syncope. While walking, she accidentally fell to her knees and did hit her head on the ground, near her left eye. Her fall was not observed, but the patient does not profess any loss of consciousness, recalling the entire event. The patient does have a history of previous falls, one of which resulted in a hip fracture. She has had physical therapy and recovered completely from that...
- **DIAGNOSTIC STUDIES:** All x-rays including left foot, right knee, left shoulder and cervical spine showed no acute fractures. The left shoulder did show old healed left humeral head and neck fracture with baseline anterior dislocation. ...
- **HOSPITAL COURSE:**
- 1. Fall: The patient was admitted and ruled out for syncopal episode. Echocardiogram was normal, and when the patient was able, ...
- 2. Status post fall with trauma: The patient was unable to walk normally secondary to traumatic injury of her knee, causing significant pain and swelling. Although a scan showed no acute fractures, ...

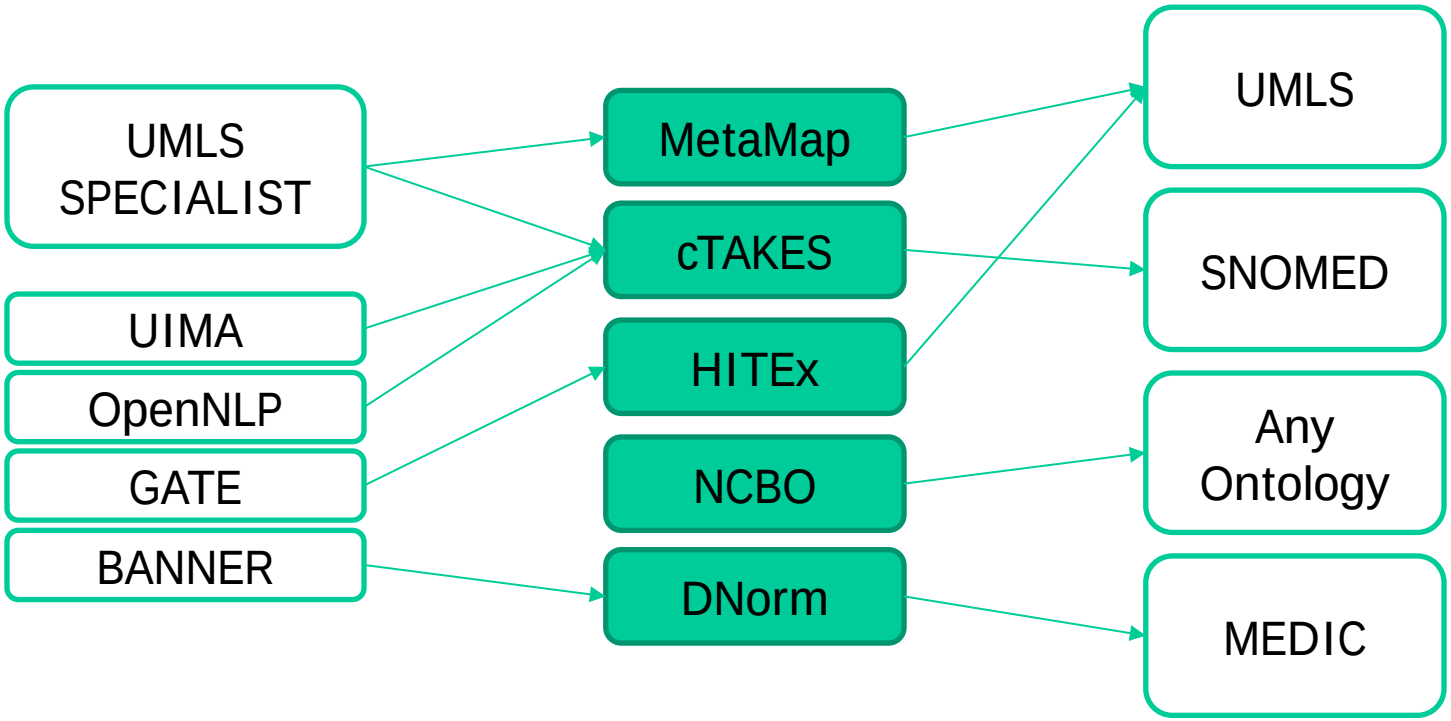
Goals and Methods

- Predict discharge diagnosis based on clinical texts
- Approach 1: Recognize diseases in text (NER-based approach)

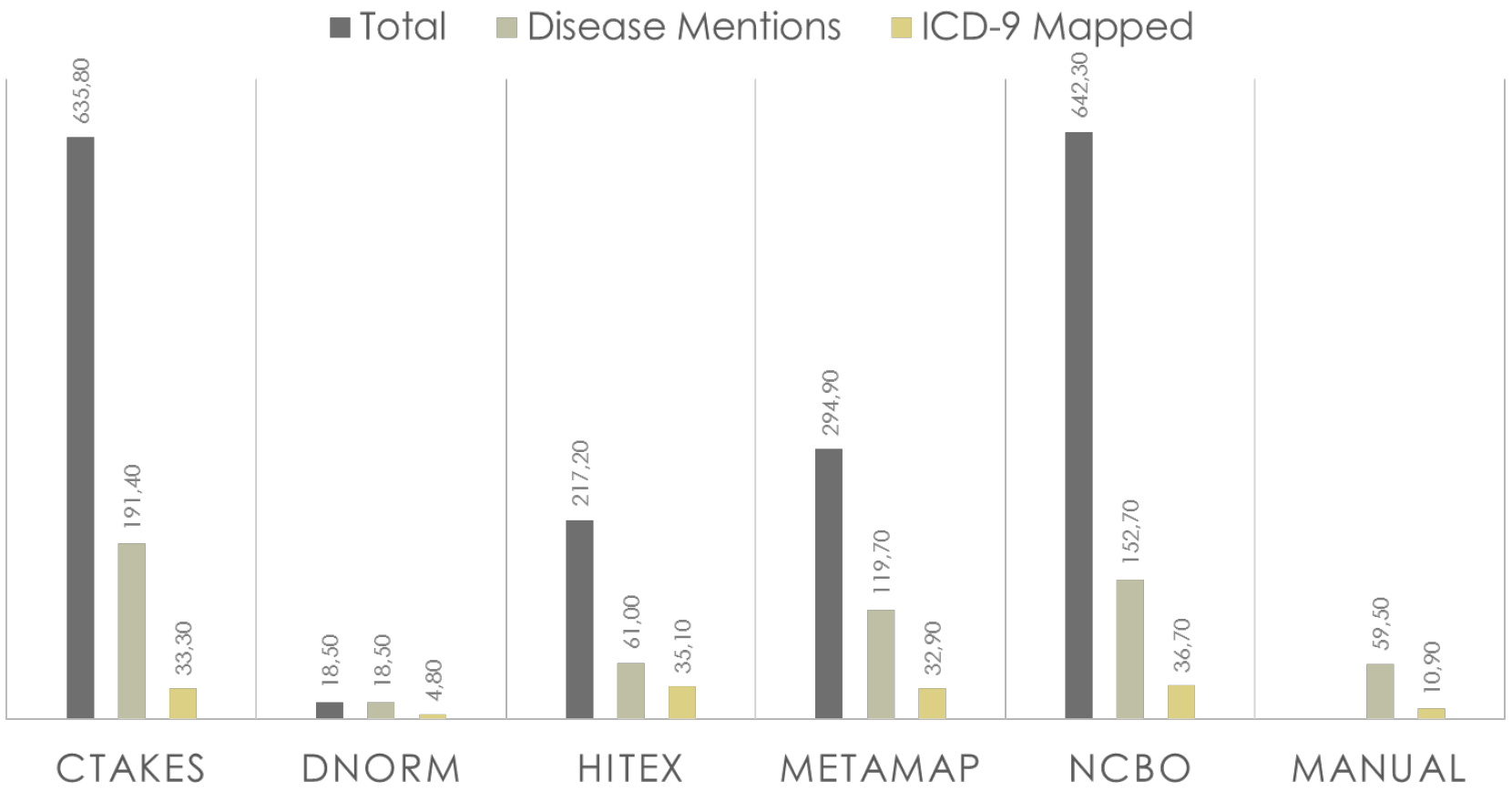
- Approach 2: Predict disease based on (entire, partial) text (classification-based approach)



Medical NER Tools Evaluated

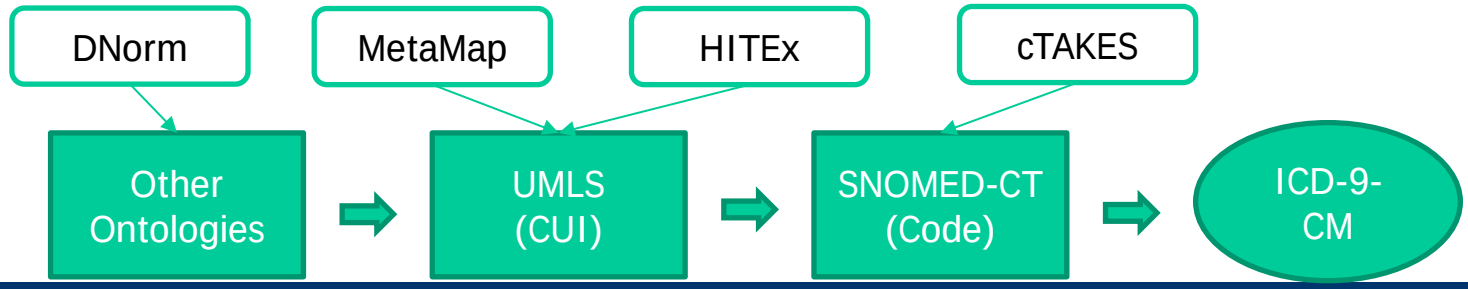


Number of Extracted Concepts (Per Document)



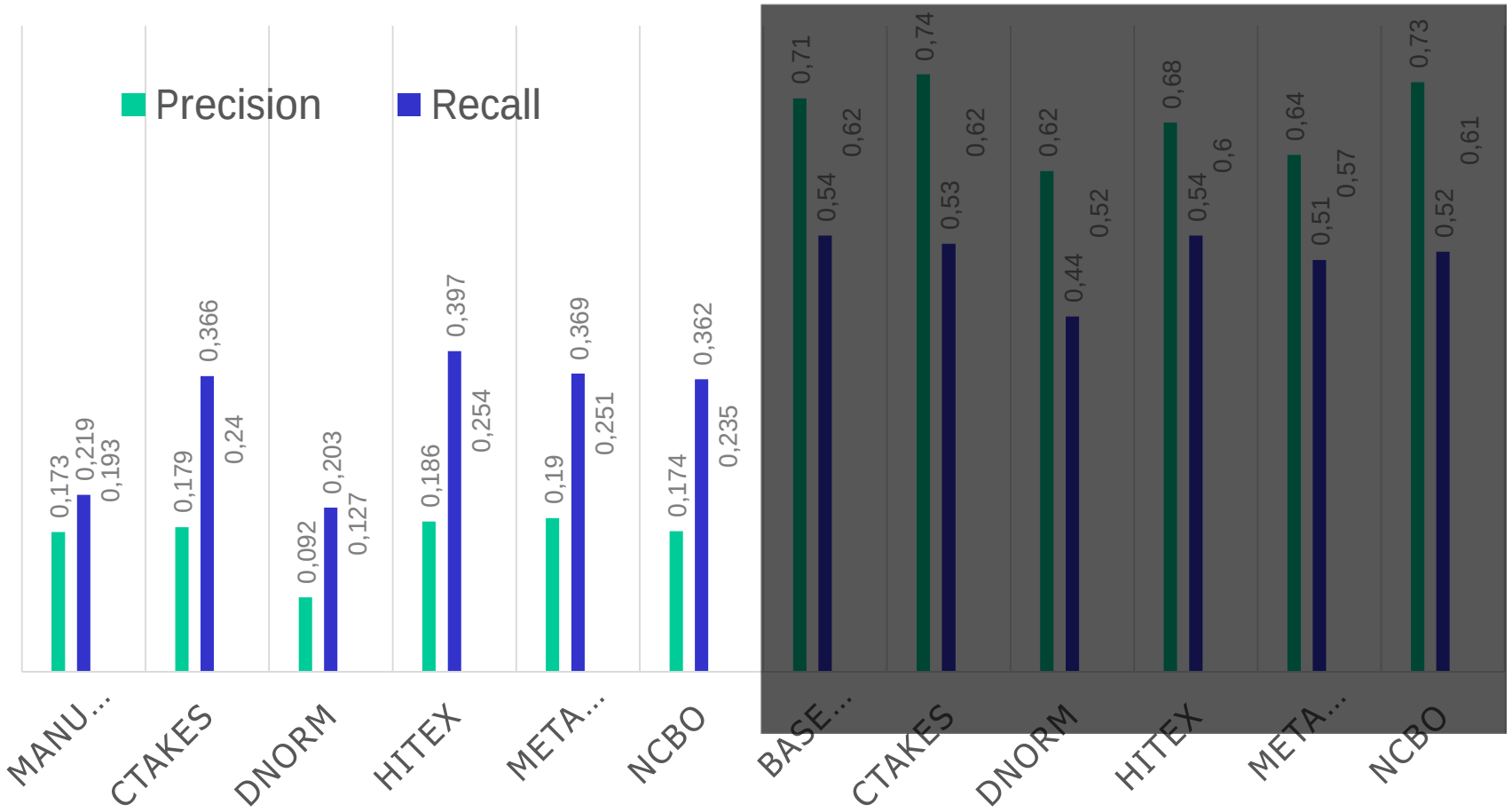
Issues (Typical)

- Hierarchical classification – which level of ICD-9?
  - Higher levels: More training data, few classes, high accuracy
    But: Little value
  - Lower levels: Little training data, many classes, low accuracy
    But: High value
- Mapping between ontologies
  - Concepts with different syntax & synonyms
  - Concepts at different granularities
  - Conflicting subsumption relationships
  - Diverging coverage
  - …



Results / Evaluation

- 50 k discharge summaries
- 7 k classes (diagnosis codes)



Results / Evaluation

- Baseline: 10 k top concepts 7 k
- Train/test split 90% / 10%