WHAT MAKES A GENE NAME?

Gene/Protein Named Entity Recognition

Biomedical Natural Language Processing

Presenter: Chelsea McBride

November 21, 2017

Universität
Stuttgart

Ha Ha Gene Names

- ‘Hamlet’
  - Affects a type of sensory cell (fruit fly)
- ‘Groucho’
  - Excessive bristles on the face (fruit fly)
- ‘Ken and Barbie’
  - Lack of genitalia (fruit fly)
- ‘Superman’
  - Causes extra stamen to grow in flowers (arabidopsis)
- ‘Merlot’, ‘riesling’, ‘cabernet’ etc.
  - A set of genes that inhibit blood cell formation (zebrafish)

Genes aren’t funny.

- Dmel_CG31753,
  - CG10568, CG15906, CG15907, CG31753,
  - Dmel\CG31753, HAM_DROME’

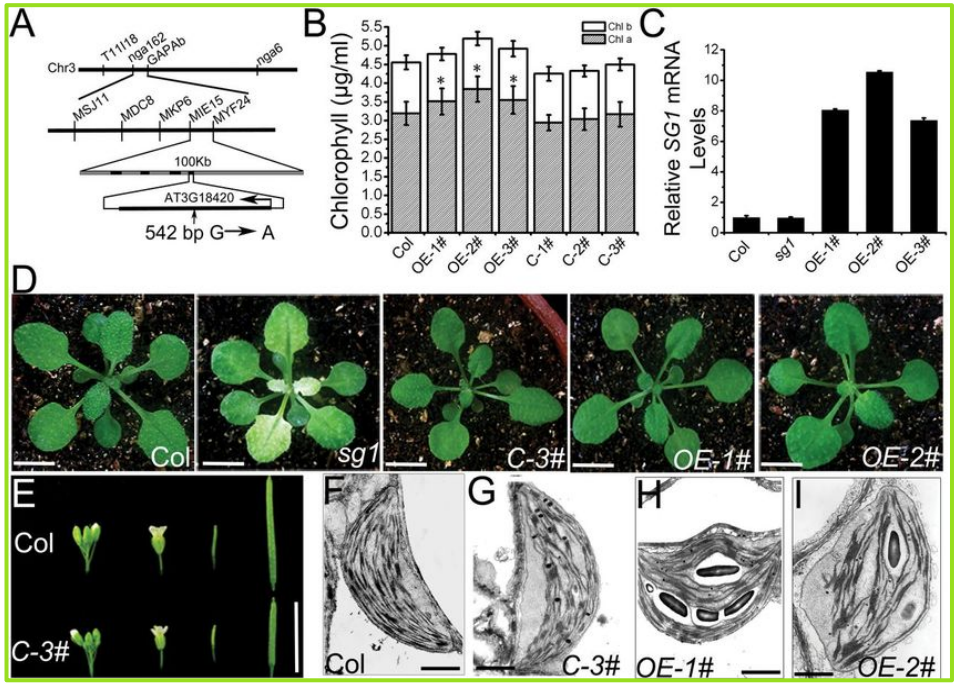
- Dmel_CG5575,
  - 2970, 3907, 5029, 8253, BCL-6/KEN,
  - CG5575, Dmel\CG5575, P420, l(2)00628, l(2)02970, l(2)05029, ok, okina

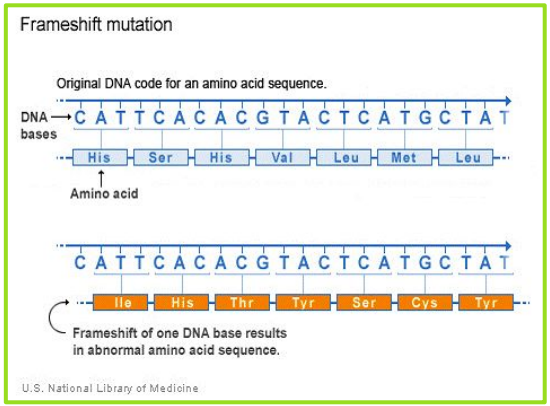
Questions. Questions that need answering.

- What makes a gene name?
  - How do we identify gene/protein mentions?
  - How do we normalize gene/protein names?
  - What is the specific challenge in recognizing gene/protein names and how is this challenge addressed?
  - What is Dictionary-based NER? How does it work?

How do we identify gene/protein mentions?

What is meant by gene **“identification”**? How do we do it?





How do we identify gene/protein mentions?

What we need to do…

- Find and identify individual gene mentioned in a text
  - Ex: Reference to gene families
- Why do we need to do this?
  - Make life easier for poor, overworked geneticists/biololists
    - Use machine learning to pinpoint gene in a text and not just the name
    - Create databases enabling quick, easy information retrieval of scientific publications
    - Devise automated system that extracts relevant information from publications
    - Classify texts for relevance to a certain topic (e.g. biology/genetics)

How do we identify gene/protein mentions?

Named Entity Recognition Approaches

- Dictionary-based approaches
  - Large collections of names, serving as examples for a specific entity class
- Rule-based approaches
  - Definition of rules to separate classes
- Classification-based approaches
  - Consider each word or phrase
  - Most popular technique in BNLP
- Sequence-based approaches
  - Complete ordered sequence of tokens and POS tags in sentence
  - Performs statistical analysis on training corpus
  - Deduce most probable sequence tags for a given word sequence
- Hybrid approaches
  - Multiple-stage processing pipelines using NLP, then machine learning

How do we identify gene/protein mentions?

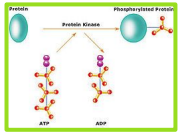
Conditional Random Fields

- Class of statistical modeling method often applied in pattern recognition, machine learning
- Used for structured prediction
- Model for tagging gene and protein mentions from text using probabilistic sequence tagging
- Can model probability *P*(**t**|**o**) of a **tag sequence** given an **observation sequence** directly

How do we identify gene/protein mentions?

Conditional Random Fields

- Use features
  - Reduce each problem to finding a feature set for representation
  - Features based on word, orthographic base
  - Character *n*-gram base $2 \leq n \leq 4$
    - Recognize substrings (e.g. ‘homeo’, ‘ase’)
  - Prefix/suffix predicates take position into account
    - Ex: ‘ase’ at end of word vs. ‘ase’ in middle

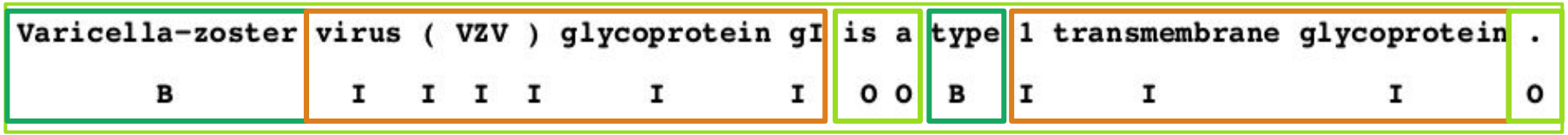


- ‘Kinase’ vs ‘laser’

Table 1: Orthographic features.

Orthographic Feature
Init Caps
Init Caps Alpha
All Caps
Caps Mix
Has Digit
Single Digit
Double Digit
Natural Number
Real Number
Alpha-Num
Roman
Has Dash
Init Dash
End Dash
Punctuation

How do we identify gene/protein mentions?



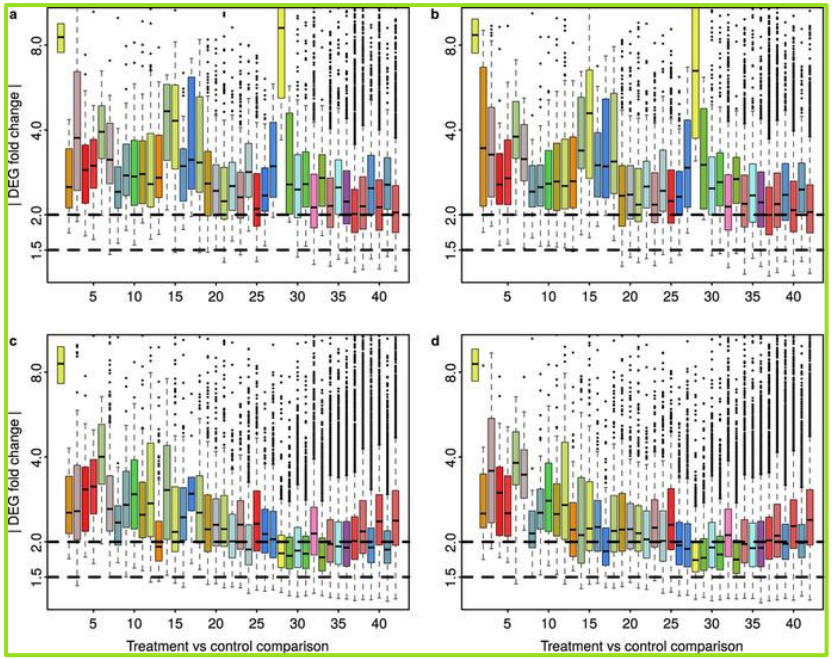
Conditional random fields

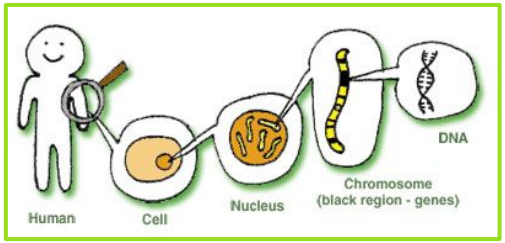
- Identification of gene mentions as tagging task
  - Each token is labeled with a tag
    - Where the token begins **(B)**
    - Where the token is intermediary **(I)**
    - Where the token is outside **(O)**

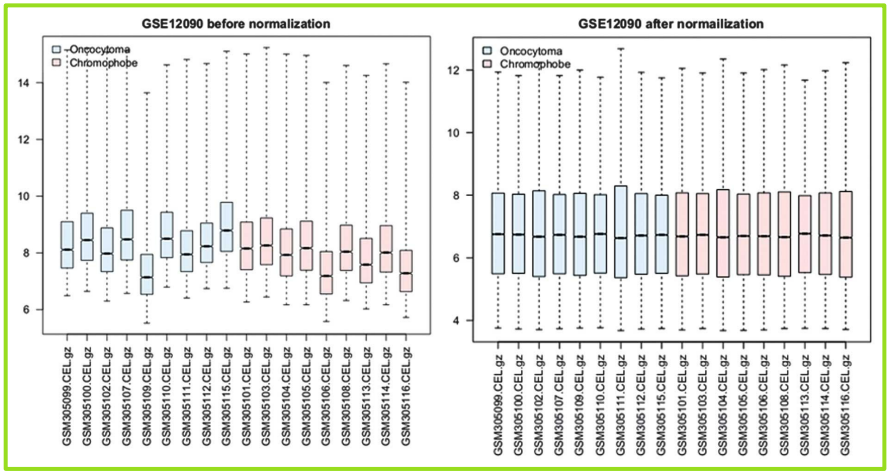
Conditional random fields → the best sequence of IOB labels

How do we normalize gene/protein names?

What is meant by gene **“normalization”**? How do we do it?







How do we normalize gene/protein names?

Gene Normalization...

- **Links** objects of potential interest...such as genes to detailed information
- Key for integrating different knowledge sources
- Information retrieval facilitates indexing and querying
- Gene mention normalization is challenging
  - Ambiguous gene names: names shared among different genes (e.g. ‘p21’)

    - https://gpsdb.expasy.org/cgi-bin/gpsdb/show

How do we normalize gene/protein names?

GNAT

- First publically available gene mention normalization system
- Handle inter-species GN (gene mention normalization)
- Extensive background knowledge on genes to resolves ambiguous names
  - '*CAT*' represents different genes in cow, chicken, fly, human, mouse, pig, deer and sheep
- Corpus containing genes from 13 species…
  - *F*-measure of 81.4% (90.8% precision, 73.8% recall)
  - Human genes → *F*-measure of 85.4%

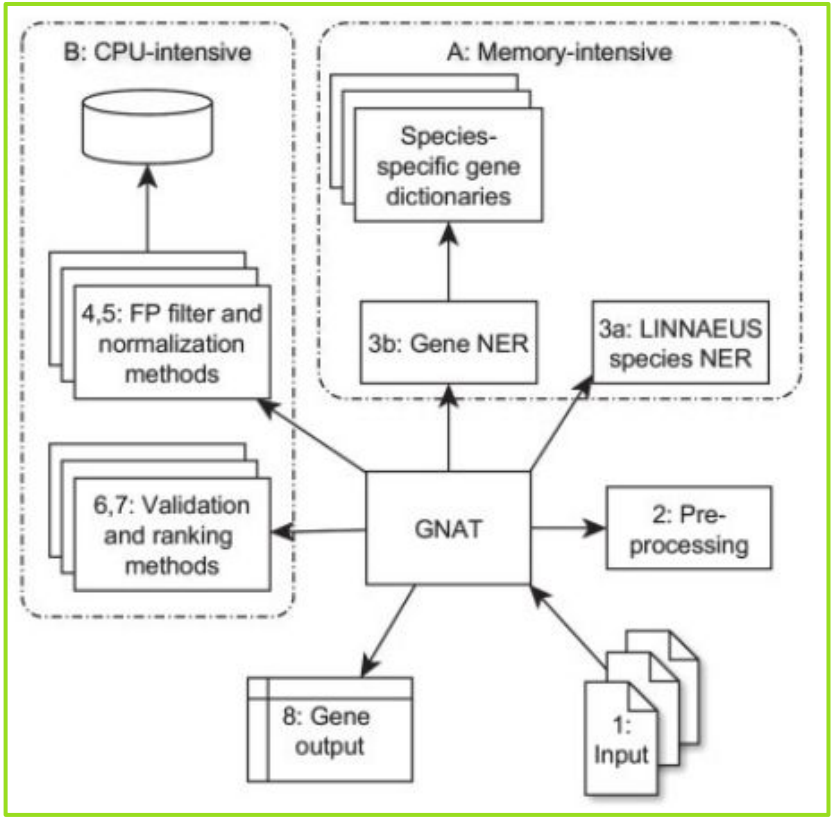
How do we normalize gene/protein names?

Inter-species gene mention normalization is harder…

- *F*-measure
  -  Mouse → 79%
  -  Yeast → 90%
  -  Human → 81%
  -  Fly → 82%
- **GNAT** → Identify every gene in text regardless of species

GNAT Processing Pipeline

1) Retrieve documents
2) Pre-process each text
3) Perform named entity recognition for genes and species
4) Remove likely false positive gene mentions
5) Assign candidate identifiers to genes
6) Validate identifiers, and...
7) Rank candidate gene identifiers.



How do we normalize gene/protein names?

The P54 gene was previously isolated from the chromosome translocation breakpoint region on 11q23 of RC-K8 cells, with t(11;14)(q23;q32). It was found to encode a 472-483-amino-acid (aa) polypeptide belonging to an RNA helicase/translation initiation factor family.

[From PubMed–ID8543178]

Potential candidates for P54, with annotations for each, extracted from EntrezGene and UniProt:

Gene	DDX6	ETS1	FKBP5	NONO	SRFS11
Chromosome	11q23.3	11q23.3	6p21.3-p21.2	Xq13.1	1p31
Length (aa)	483	441	457	471	484
GO (examples)	RNA helicase activity	immune response	FK506 binding	DNA binding	mRNA processing

(Gene Ontology Consortium)

5 human genes that have the synonym ‘P54’

- Retrieve all known information on 5 genes and map to sentence
- Hints: Reference to chromosomal location, length of protein and GO term
  - Human ‘DDX6’ gene as solution for ‘P54’ normalization

What is the specific challenge in recognizing gene/ protein names and how is this challenge addressed?

NER in the life sciences is a rather difficult problem.

- The recognition of biological objects in written language is very difficult due to many factors
  - A general lack of naming conventions
  - Excessive use of abbreviations (‘p54’)
  - Frequent usage of synonyms and homonyms
  - Often have names consisting of many single words,
    - Ex: ‘human T-cell leukaemia lymphotropic virus type 1 Tax protein’

What is the specific challenge in recognizing gene/ protein names and how is this challenge addressed?

What makes BNER complicated? Synonyms, homonyms, abbreviations and ambiguities

- Text represent real-life concepts in our mind, represented in language
  - Ex: “What is a supermarket?”, “How do I name a supermarket?”
- In quickly changing highly specialised domains as molecular biology…
  - Agreements do not have time to build or are subject to frequent modifications
  - Real-life concepts and their textual representations → **not unambiguously defined**
  - No community-wide agreement on how a particular gene should be named
  - Concept denoted by a gene name is usually not clearly defined

What is the specific challenge in recognizing gene/ protein names and how is this challenge addressed?

NER tools are important building blocks for text-mining tools supporting biologists

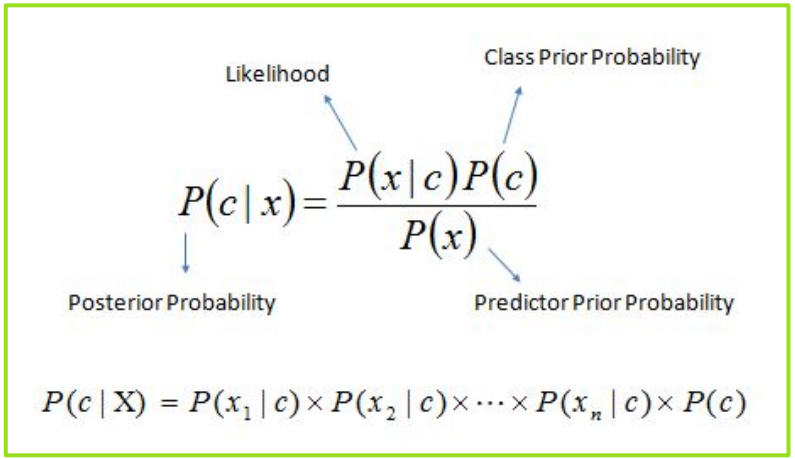
- NER consists of three different problems
  - Recognition of a named entity in text
  - Assignment of a class to this entity (gene, protein, drug, etc)
  - Selection of a preferred term in case that synonyms exist
- Most current systems concentrate on gene/protein names
  - Do not distinguish between these two classes

NER is difficult even for humans...and more difficult for machines

What is the specific challenge in recognizing gene/ protein names and how is this challenge addressed?

So how do we deal with these problems...? Machine learning.

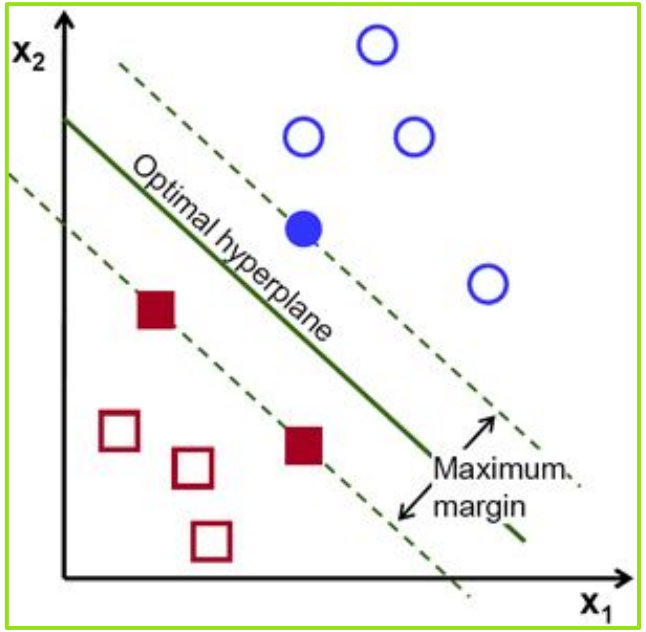
- Naive Bayes
  - Analyse distributions of properties within different classes
  - Calculate probabilities for belonging to either class
  - Used to classify words or phrases as being an entity name or not



What is the specific challenge in recognizing gene/ protein names and how is this challenge addressed?

So how do we deal with these problems...? Machine learning.

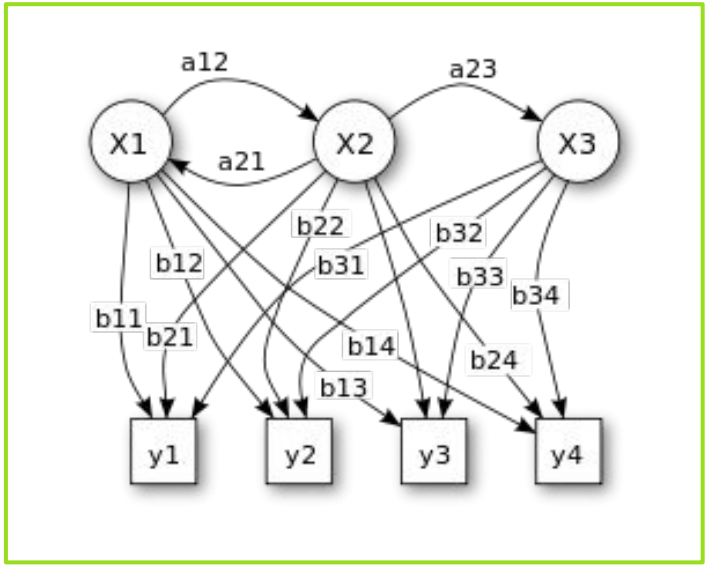
- Support vector machines (SVM)
  - Deduce linear combinations of features from feature vectors
  - Support vectors define a hyperplane in a multidimensional feature space
  - Separate all (ideally) positive examples from all (ideally) negative examples



What is the specific challenge in recognizing gene/ protein names and how is this challenge addressed?

So how do we deal with these problems...? Machine learning.

- Hidden markov models
  - Use order in which features appear in text
  - Aggregate statistical information from labelled examples
  - Predict most probable sequence of events for a given sentence



What is Dictionary-based NER? How does it work?

Dictionaries are large collections of names, serving as examples for a specific entity class

- Dictionary-based NER approaches match text against a fixed name list
- Matching dictionary entries exactly against text is a simple, precise NER method
  - Yields only very low recall
- To compensate, either...
  - Use inexact matching techniques, or...
  - ‘Fuzzify’ the dictionary by automatically generating typical spelling variants for every entry

What is Dictionary-based NER? How does it work?

Why use dictionary-based NER techniques? Why not?

- How much human intervention needed to build your system?
  - Using a dictionary as input to model saves time annotating a corpus…
  - ...but hinders inclusion of contextual informations that machine learning techniques would include
  - Fixed lists of names, collected from expert curated database provide unique identifiers for each instance → not solved by machine learning algorithms
  - Improve recall → refinement of entries or fuzzy matching algorithms
    - Achieve robust spelling variants → leads to severe danger of overfitting

GNAT Error Analysis

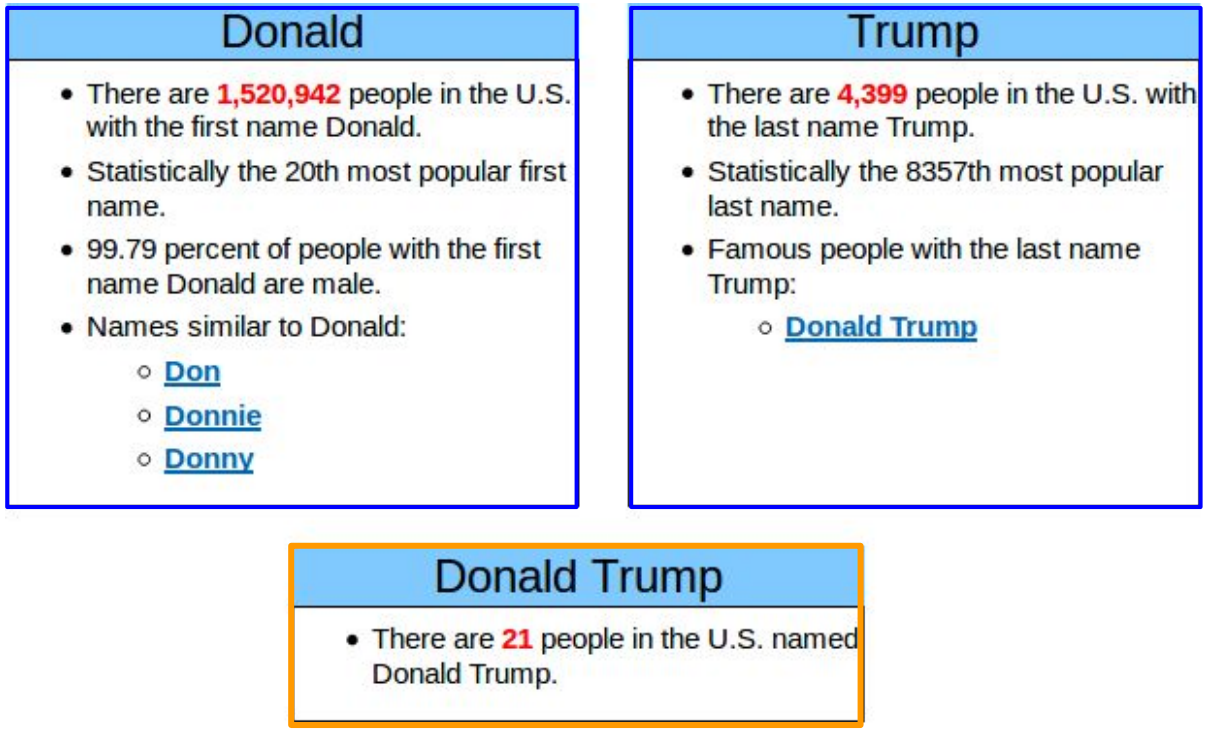
False Negatives

False Negatives

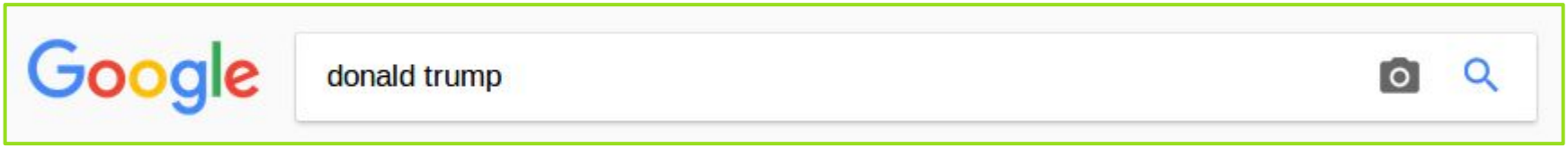
Category	Number
Too restrictive filtering by context	24
No similar synonym known	16
Too dissimilar synonym	10
Too restrictive filtering of stopwords	6
No species/wrong species found in abstract	4
Unspecific species found in abstract	5
No assignment to species	3
Too restrictive filtering of names	2
Too many IDs left after disambiguation	3
Missed conjunction	1
Abbreviation resolution failed	1
Miscellaneous cases	10
Additional/wrong species assigned to gene	14
Disambiguation assigned wrong ID	4
System found a too general mention	1
Closer synonym for wrong gene	3
Name does not refer to a gene	2

What is Dictionary-based NER? How does it work?

Let's normalize Donald Trump: (http://howmanyofme.com/people/Donald_Trump/)



What is Dictionary-based NER? How does it work?



Donald Trump…?

Donald Trump.

TAKE HOME

Personal name recognition is not that difficult.

Donald Trump

TAKE HOME

Business/restaurant name recognition is not that difficult.

The Codfather

TAKE HOME

Place name recognition is not that difficult.

Stuttgart,

Germany

TAKE HOME

Gene name recognition is more difficult.

CAT (catalase)...

Aliases for CAT Gene

Aliases for CAT Gene

Catalase [2 3 3 5]

EC 1.11.1.6 [4 61]

External Ids for CAT Gene

HGNC: 1516 Entrez Gene: 847 Ensembl: ENSG00000121691 OMIM: 115500 UniProtKB: P04040

Previous GeneCards Identifiers for CAT Gene

GC11P036013, GC11P035138, GC11P034499, GC11P034424, GC11P034417, GC11P034157

Search aliases for CAT gene in PubMed and other databases

Or cat?

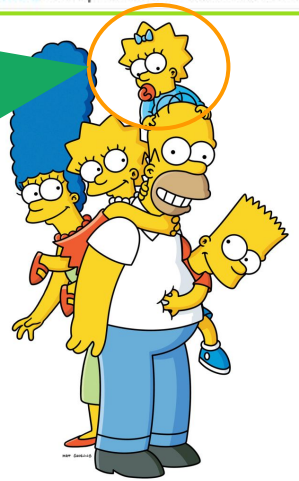
TAKE HOME

Gene name recognition is more difficult.

MAGGIE...

or Maggie?

Protein names	*Submitted name:* Maggie, isoform C Imported
	*Name:*mge Imported
	Synonyms:BcDNA:LP07226 Imported, CG14981 Imported
	ORF Names:Dmel_CG14981 Imported
Organism	Drosophila melanogaster (Fruit fly) Imported
Taxonomic identifier	7227 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Ecdysozoa › Arthropoda › Hexapoda › Insecta ›
Proteomes	UP000000803 Component: Chromosome 3L



Answers. Answers to questions.

- Lots of things including but not limited to...
  - Find and identify individual genes mentioned in a text using NLP and machine learning approaches
  - Normalization links objects of potential interest such as genes to detailed information
  - Synonyms, homonyms, abbreviations and ambiguities (to name a few)
  - Dictionary NER matches dictionary entries exactly against text simply and precisely

REFERENCES I

Hakenberg, Jörg, et al.
Inter-species normalization of gene mentions with GNAT.
Bioinformatics, Volume 24, Issue 16, Pages i126–i132, 15 August, 2008.

Leser, Ulf, and Jörg Hakenberg.
What makes a gene name? Named entity recognition in the biomedical literature.
Briefings in bioinformatics 6.4, 2005.

McDonald, Ryan, and Fernando Pereira.
Identifying gene and protein mentions in text using conditional random fields.
BMC bioinformatics 6.1, 2005.

REFERENCES II


Wei, Chih-Hsuan, Hung-Yu Kao, and Zhiyong Lu.
GNormPlus: an integrative approach for tagging genes, gene families, and protein domains.
BioMed research international, 2015


Hakenberg, Jörg, et al.
The GNAT library for local and remote gene mention normalization.
Bioinformatics 27.19, 2011

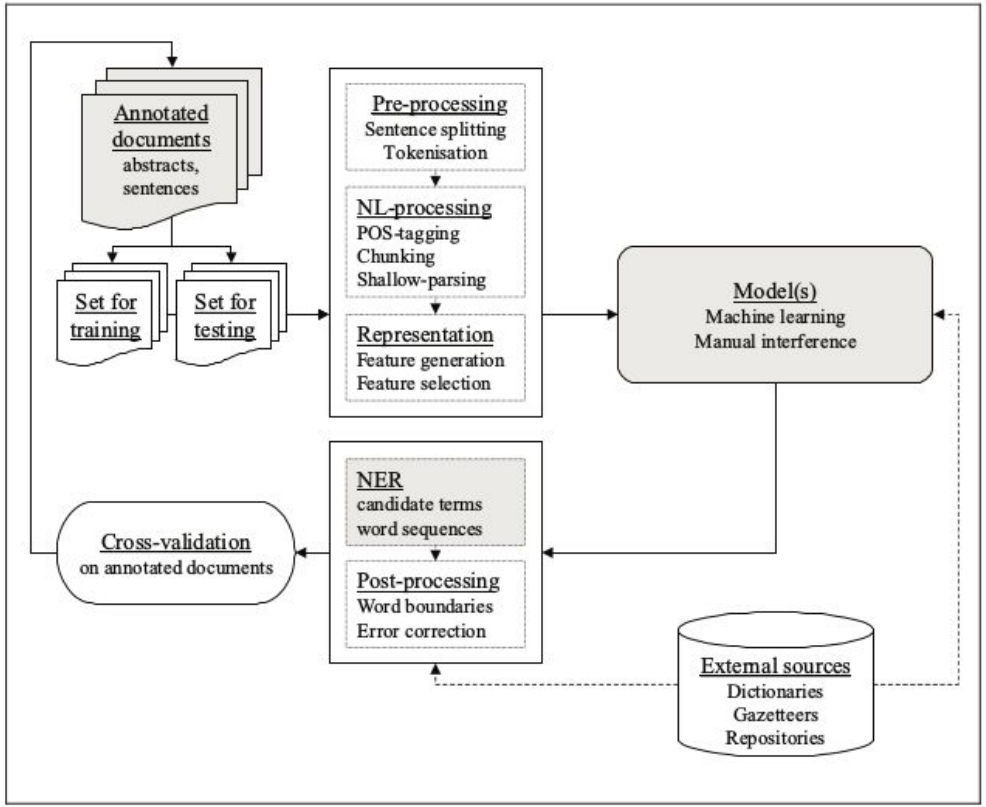

Lafferty, John, Andrew McCallum, and Fernando CN Pereira
Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
Conference Paper, University of Pennsylvania, 2001

Thank You for Your Attention

COMMENCE QUESTIONS

Universität Stuttgart

Typical Process Flow in Named Entity Recognition



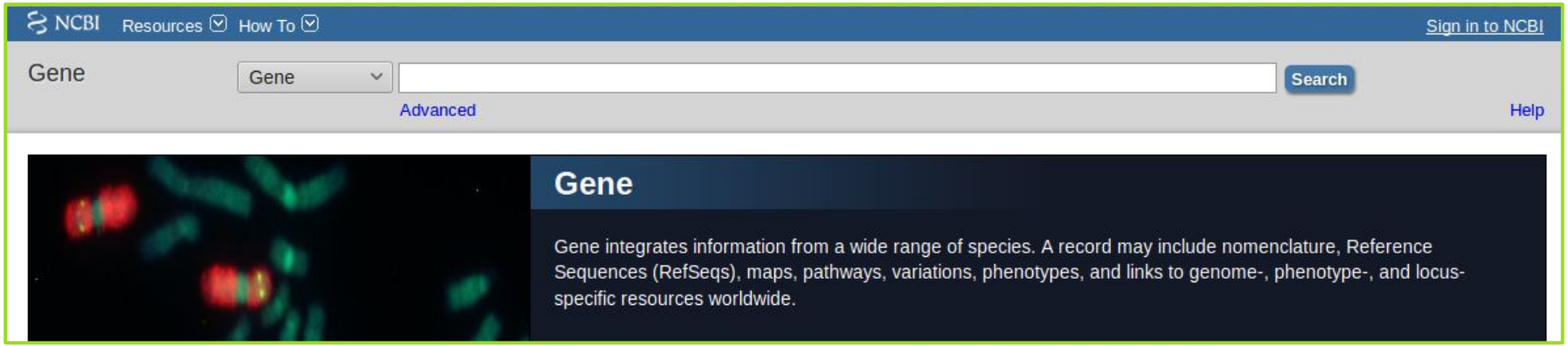
Effects of System Components

Table 2: Effect of system components on development data

System	Precision	Recall	F-Measure
A. No Lex, No Feat. Ind.	0.793	0.731	0.761
B. No Lexicons	0.807	0.744	0.774
C. Trigrams	0.811	0.759	0.784
D. Non-gene Lexicons	0.818	0.743	0.778
E. Gene Lexicons	0.812	0.775	0.793
F. All Lexicons	0.817	0.782	0.799

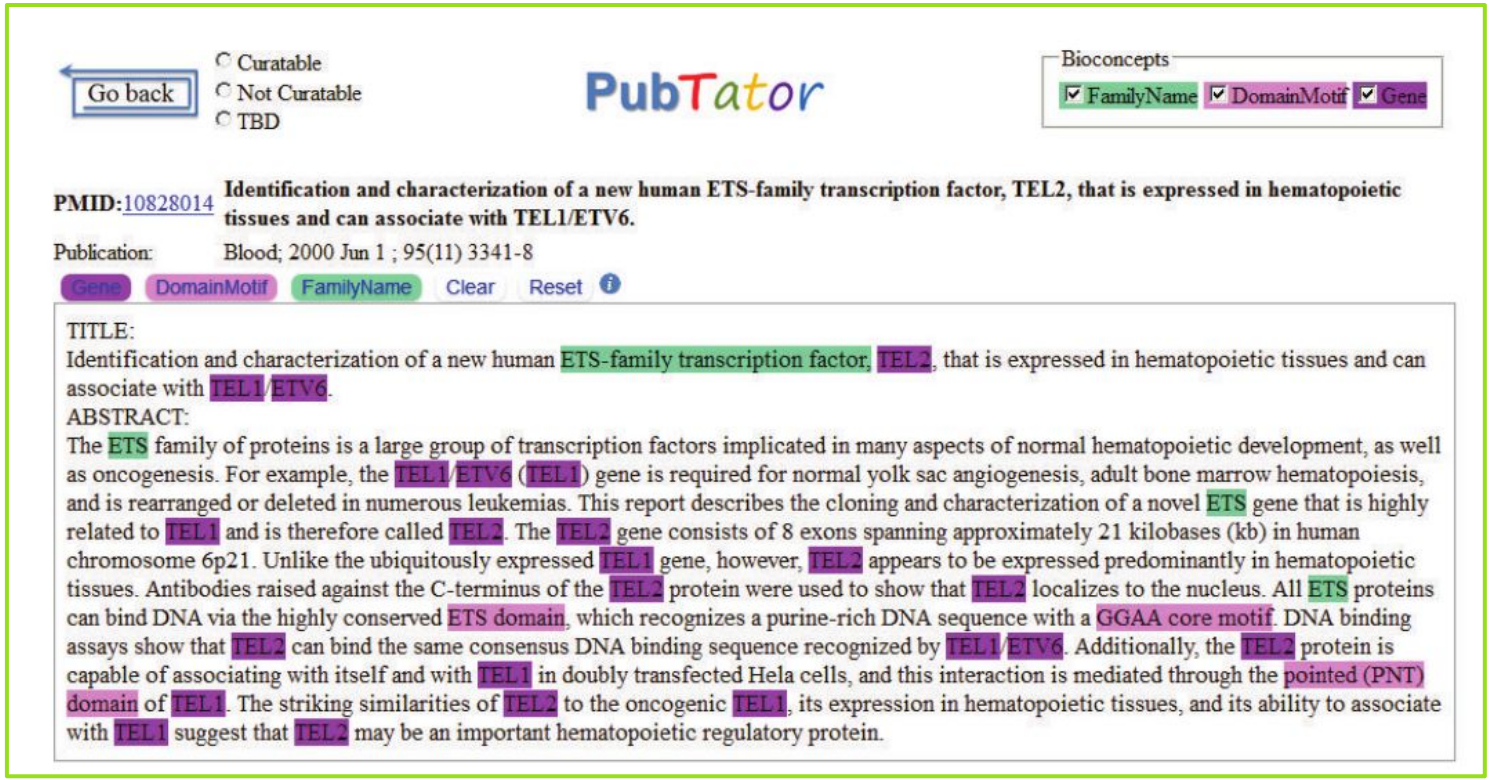
A) System containing no lexicon features and does not use feature induction. **B)** Same as A, except feature induction is used. **C)** Same as B, except features using the infrequent trigram lexicon are used. **D)** Same as B, except features using the non-gene lexicons are used. **E)** Same as B, except features using the gene lexicon are used. **F)** Same as B, except features using all lexicons are used.

What is Dictionary-based NER? How does it work?



[Let's look up something else]

PubTator



Pubtator Entity Mention for TEL2

Named Entity Recognition Corpora

genia

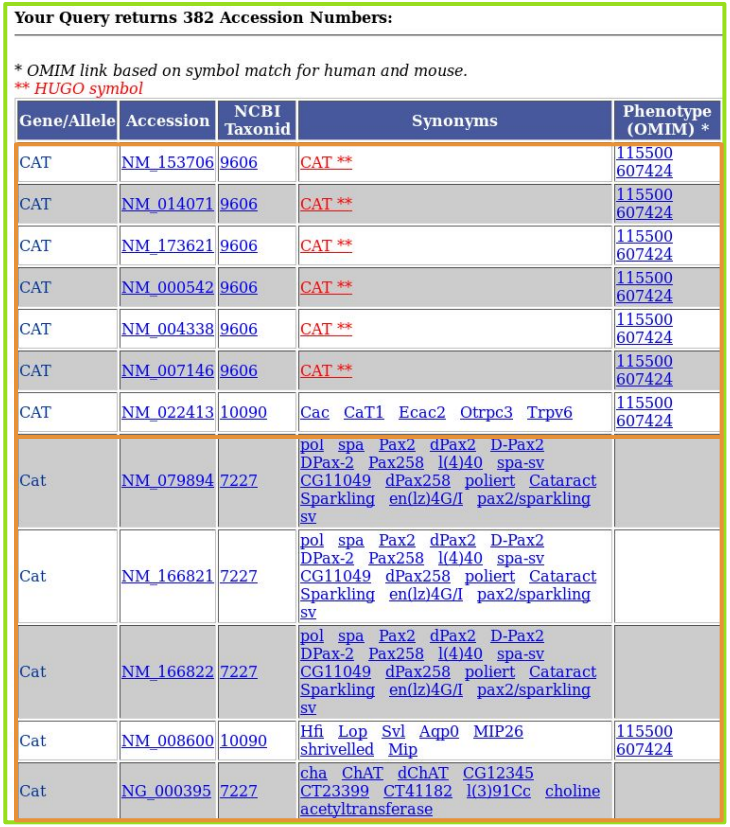
	Size	Entity types	Number of tagged entities	Evaluation scheme	Text selection	Metadata shipped with corpus
BioCreative	15,000 sentences	Gene/protein	17,800	Strict, with some alternative word boundaries	Random selection by Genbank curators	PubMed-ID, POS, tokenisation, sentence splitting
GENIA	2,000 abstracts, approx. 19,000 sentences	Various	eg 21,800 protein_molecule, 8,353 DNA_domain_or_region	Strict	PubMed query	PubMed-ID, POS, tokenisation, sentence splitting
Yapex	201 abstracts	Protein	3,711	Strict	PubMed query + at random from GENIA	PubMed-ID

What is the specific challenge in recognizing gene/ protein names and how is this challenge addressed?

What makes NER complicated? Synonyms, homonyms, abbreviations and ambiguities

- Clones in the Human Genome Project? → up to 15 different names
- Gene names not distinguished from normal language
  - Ex: “White” (symbol *w*), “shaggy” (symbol *ssg*), “mind the gap” (symbol *mtg*)
- Prefixes or suffixes with digits or letters...or Greek
  - ‘MRP2’, ‘MRP3’, ‘Dbf4p’, ‘CCAAT/enhancer binding protein (C/EBP)’, ‘alpha’

What is Dictionary-based NER? How does it work?



Your Query returns 382 Accession Numbers:

** OMIM link based on symbol match for human and mouse.*
*** HUGO symbol*

Gene/Allele	Accession	NCBI Taxonid	Synonyms	Phenotype (OMIM) *
CAT	NM_153706	9606	CAT **	115500 607424
CAT	NM_014071	9606	CAT **	115500 607424
CAT	NM_173621	9606	CAT **	115500 607424
CAT	NM_000542	9606	CAT **	115500 607424
CAT	NM_004338	9606	CAT **	115500 607424
CAT	NM_007146	9606	CAT **	115500 607424
CAT	NM_022413	10090	Cac CaT1 Ecac2 Otrpc3 Trpv6	115500 607424
Cat	NM_079894	7227	pol spa Pax2 dPax2 D-Pax2 DPax-2 Pax258 l(4)40 spa-sv CG11049 dPax258 poliert Cataract Sparkling en(lz)4G/I pax2/sparkling sv	
Cat	NM_166821	7227	pol spa Pax2 dPax2 D-Pax2 DPax-2 Pax258 l(4)40 spa-sv CG11049 dPax258 poliert Cataract Sparkling en(lz)4G/I pax2/sparkling sv	
Cat	NM_166822	7227	pol spa Pax2 dPax2 D-Pax2 DPax-2 Pax258 l(4)40 spa-sv CG11049 dPax258 poliert Cataract Sparkling en(lz)4G/I pax2/sparkling sv	
Cat	NM_008600	10090	Hfi Lop Svl Aqp0 MIP26 shrivelled Mip	115500 607424
Cat	NG_000395	7227	cha ChAT dChAT CG12345 CT23399 CT41182 l(3)91Cc choline acetyltransferase	