Text Mining & Bioinformatics

Patrick Ruch
Bibliomics and Text Mining – BiTeM
HEG-HES-SO & SIB
patrick.ruch@hesge.ch

Swiss Institute of Bioinformatics

Hes·so
Haute Ecole Spécialisée de Suisse occidentale
University of Applied Sciences Western Switzerland

Round table…

- You
  - Bio ?
  - Info ?
  - Other ?
- My team
  - Biol #2 → Bioinformatics
  - Info #8 → 50% Info + 50% Bioinformatics
  - Other #5 [MD, signal processing, statistician, economics, …]
- Me: mathematics, PhD comp sci, HUG, NLM, IBM R, …

Overview

- Introduction and objectives
- Metrics
- Features
- Contents
- Tasks - Methodologies
- Conclusion

Objectives

- Introduce how text mining can support future bioinformaticians
- Explain how text mining operate with biological entities and the « biological » ecosystem
- Stimulate your interest into a satellite - yet very lifeful - bioinformatics field

Text Data Mining

- Text Mining is like Data Mining but works with textual contents
- … So any analysis can be performed with text mining provided the content is available in text ?
- Answer: **Jein !**

Disciplines

- Natural language processing, Natural language understanding, computational linguistics (+)
- Machine learning / data mining (++)
- **Information retrieval – Foundational… and hard to improve (+++)**

Common application fields

- Information retrieval
- Biocuration support tools
- Hypothesis generation – Literature based discovery

User level tasks

- Search – Foundations
- Triage
- Keyword assignment
- (Named-)Entity recognition
- Extract passages or more complex entities (e.g. protein protein interactions)
- Literature based discovery

Metrics

- Precision
- Recall
- Other metrics…

Precision

- Given 5 relevant documents in a collection for a given query, a search engine returns **10** documents, including **3**, which are pertinent

- P = 3/10 = 0.30 or 30%

Recall

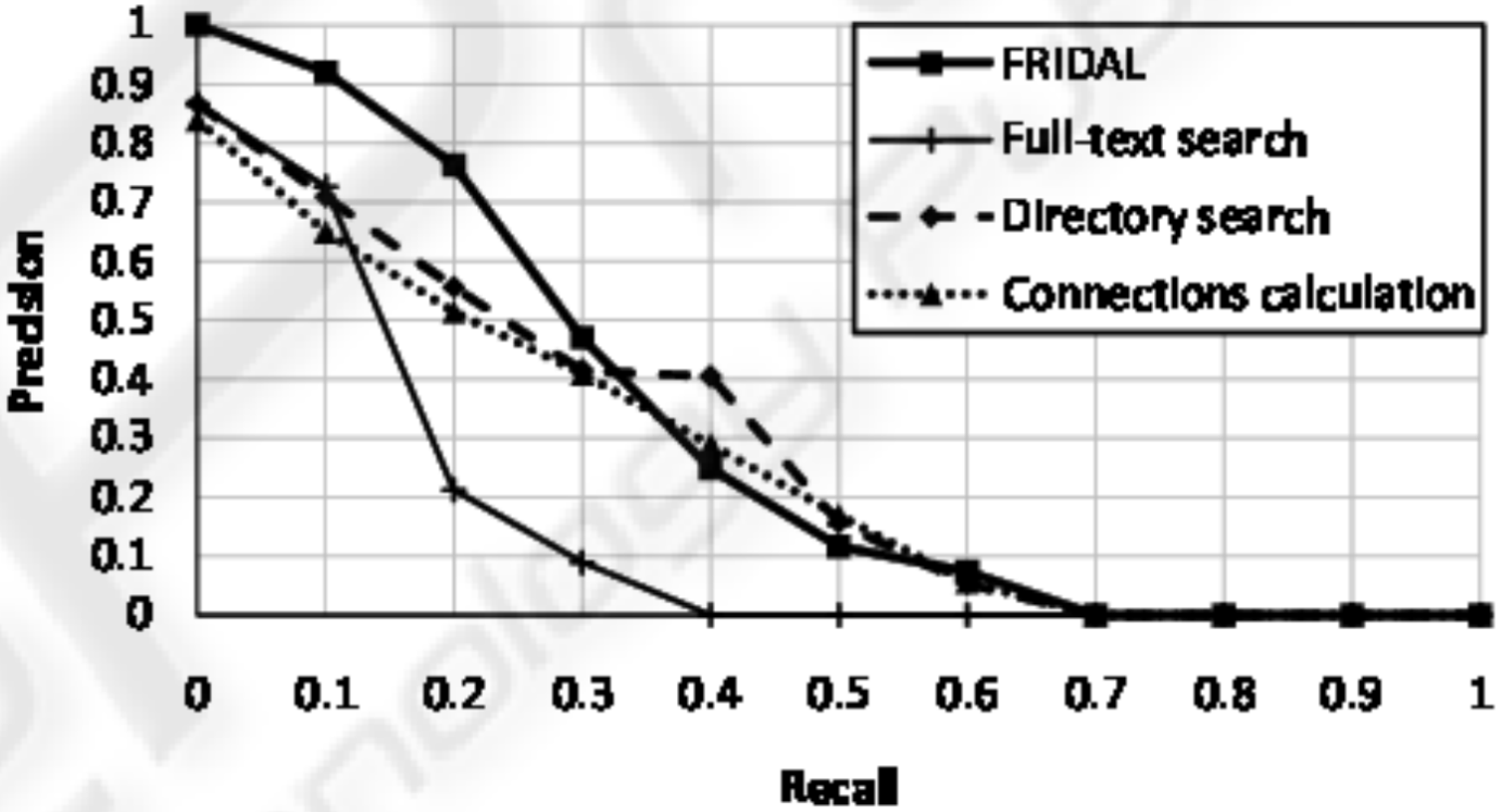
- Given **5** relevant documents in a collection for a given query, a search engine returns 10 documents, including **3**, which are pertinent

- Recall = 3/5 = 0.60 or 60%

Synthetic metrics

- Rank
  - R^{th-1} is more important than R^{th}
  - So, we compute average precision at different rank values (10, 20, … 30%, …)
  - Mean average precision
- F1 and related metrics
  - Harmonic or geometric mean
  - Utility metrics
    - E.g. 0.9 x Recall + 0.1 x Precision

Example


Precision
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Recall
FRIDAL
Full-text search
Directory search
Connections calculation

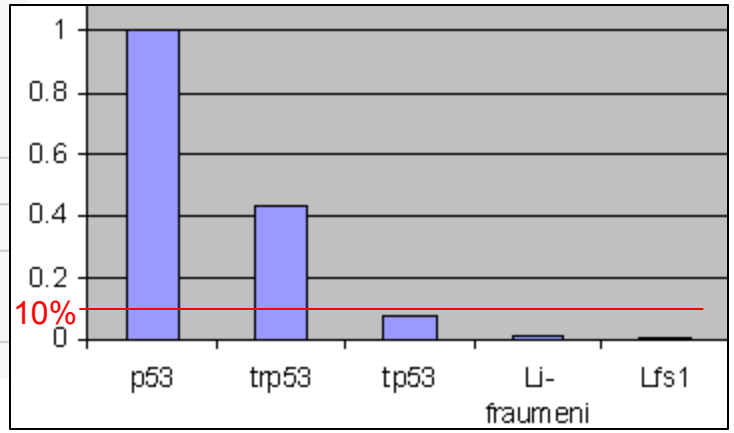
Features

- Words
- Subwords (character N-grams)
- Stems
- Word N-grams
- Syntactic entities (noun phrases, verb phrases, …),
- Semantic entities (gene names, chem. compounds, diseases, …)

Term normalization: database & ontology vs. reality !

Antigen NY-CO-13	Protein	SwissProt:P04637
Cellular tumor antigen p53	Protein [preferred]	SwissProt:P04637
FLJ92943	Gene	EntrezGene:7157
LFS1	Gene	EntrezGene:7157
		HGNC:11998
Li-Fraumeni syndrome	Gene	HGNC:11998
p53	Gene	EntrezGene:7157
		HGNC:11998
P53	Gene	OMIM:191170
		SwissProt:P04637
p53 antigen	Gene	EntrezGene:7157
p53 transformation suppressor	Gene	EntrezGene:7157
p53 tumor suppressor	Gene	EntrezGene:7157
phosphoprotein p53	Gene	EntrezGene:7157
Phosphoprotein p53	Protein	SwissProt:P04637
TP53	Gene [preferred]	HGNC:11998
		SwissProt:P04637
	Gene	EntrezGene:7157
		OMIM:191170
transformation-related protein 53	Gene	EntrezGene:7157
TRANSFORMATION-RELATED PROTEIN 53	Gene	OMIM:191170
TRP53	Gene	EntrezGene:7157
		OMIM:191170
tumor protein p53	Gene [preferred]	HGNC:11998

Synonyms	#
p53	53362
trp53	23364
tp53	4156
li-fraumeni	775
lfs1	431



Transformations (ETL, reg. expressions, …)

- i, ii, iii → 1, 2, 3 (e.g. *histone deacetylase iii*)
- Greek letters (e.g *α-tubulin*)
- Hyphenation «-»: {alphatubulin, alpha, tubulin)
- Chemistry
  - Inchi
  - SMILES
  - PubChem, chEBI, DrugBank…

On parvient néanmoins a libérer la vésicule du lit vésiculaire de manière rétrograde et a individualiser le canal cystique . Mise en place d'une ligature au niveau proximal du canal cystique . Section partielle du canal cystique . Il est impossible de cathétériser le canal cystique . On décide de ne pas continuer les manœuvres vu le risque de plaie au niveau de la voie biliaire et , d'autre part , la Cholangiographie rétrograde préopératoire , qui s'est avérée normale . Ligature au niveau du cystique.

Compression power: 80 → 51 → 23 forms !

biliair
canal
cholangiograph
continu
cystiqu
liber
Ligatur
lit
manoeuvr
mis
niveau
normal
part
partiel
plac
plai
preoperatoir
proximal
retrograd
risqu
section
vesicul
voi

Ratio = 1/3 !

Ambiguities and noise often acceptable

0006771|guerison par la **foi**|D029221| **FAITH**
0006771|gycogene phosphorylase, foie|D025001|
0006771|huile foie morue|D003060|
0006771|maladie veino-occlusive foie|D006504|
0006771|maladies alcooliques foie|D008108_1|
0006771|microsome foie|D008862|
0006771|foies, mitochondrie|D008930|
0006771|tachycardie **foyer** auriculaire ectopique|D013612_1|
0006771|tachycardie jonctionnelle **foyer** ectopique|D013613_1| **NODE**
0006771|transplantation foie|D016031|
0006771|tumeur experimentale foie|D008114|
0006771|tumeur foie|D008113|
0006771|veino-occlusive foie, maladie|D006504_1|
0008370|**foie**|D008099|

Principle: «one sense per corpus»

Contents

- Literature +++
- Electronic Health Records ++
- Grey literature, patents ++
- Social media +
- Other contents (news, archives, …)

- Resources: ontologies, terminologies, dictionaries, large corpora, …

Back end tasks – intelligent stuff that computers can do ☺

- Ranker
- Classifier
- … and both are the same: a classifier is a ranker with a threshold !
- Regression (ok for structured data)

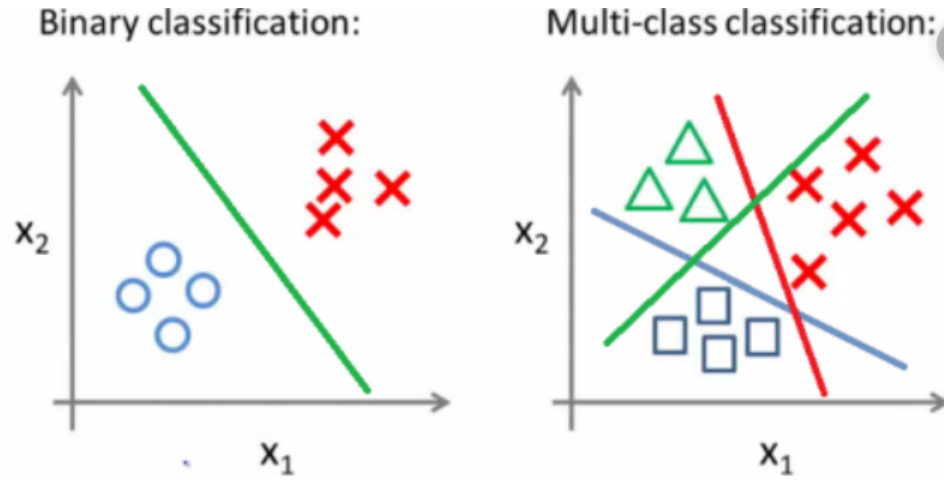
  NB: we can use it but it is not to fit a function !

Ranking and text mining features

- Given an objective function, rank a collection of text !
- Objective function = distance, precision, recall, …

Classification: one-class, binary, N-class

- With N binary classifiers, we can design N-class categorizers



Search 1

- PubMed
  - Boolean & Ante-chronological
  - Vector-space: words are like « dimensions »

  [Text Mining representations are very high dimension sparse matrix models]
- EuropePMC
  - Lucene (vector-space)
- SIBiLS
  - Lucene
  - Combination of weighting schema (Terrier)

Search 2

- Search (Links)

https://pubmed.ncbi.nlm.nih.gov/

https://candy.hesge.ch/SIBiLS/

- CovidTriage

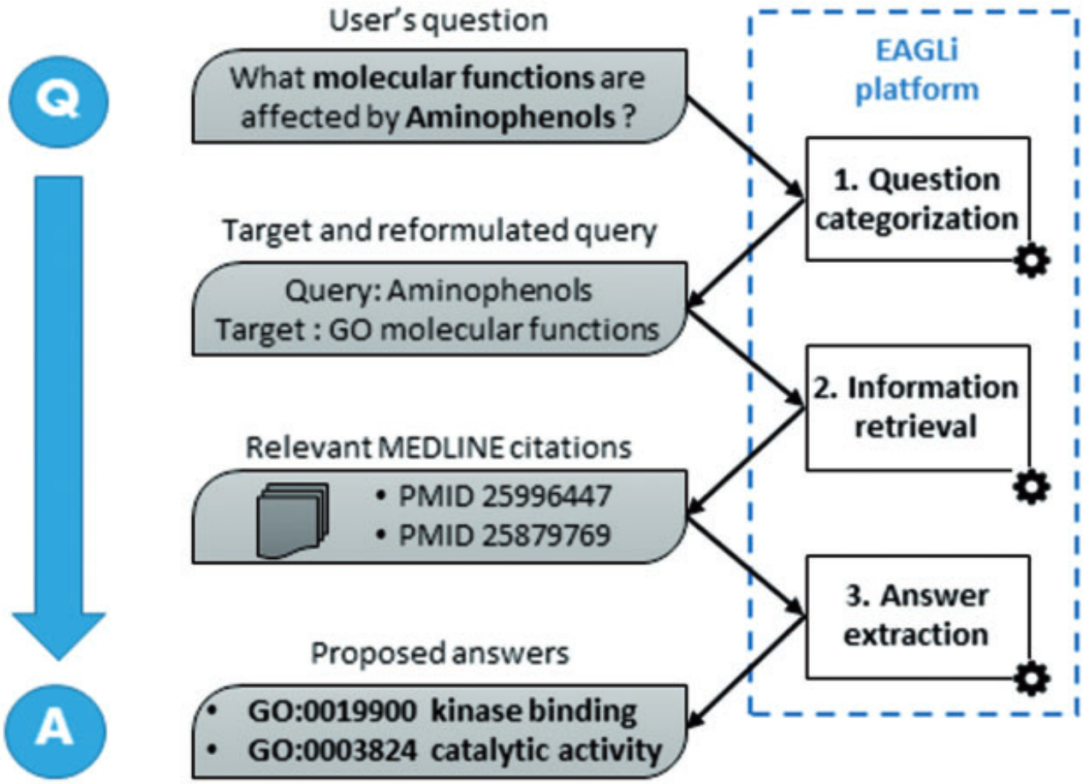
http://candy.hesge.ch/CovidTriage/

- Covidex (Question Answering)

https://covidex.ai/

Search 3 / Question-answering

- Dedicated architecture
  - Information retrieval
  - + NLP
- Data-driven: 100’000 qu
  - MS-MARCO, …



EAGLi: What enzymes can inhibit levamisole ?

what enzyme are inhibited by levamisole?

EAGL i PubMed

Search

Your question was : *what enzyme are inhibited by levamisole?*, **reformulated as** *enzym inhibit levamisol*

Possible answers are :

- *Alkaline Phosphatase (85 matches in 23 documents)*
- *Isoenzymes (8 matches in 2 documents) ; Cholinesterases (5/2) ; Muramidase (5/2) ; Prostaglandin-Endoperoxide Synthases (4/1) ; Malate Dehydrogenase (4/1) ; Apyrase (6/1) ; 5'-Nucleotidase (7/1) ...* show all

Automatic assignment of Gene Ontology and Swiss-Prot keywords to the selected articles :

Alkaline Phosphatase

Score

A specific alkaline phosphatase of amphibia integument levamisole effect on short circuit current (SCC).
Andreoletti G E , Donna D , Dore B , Lodi G , Savardi L
Boll Soc Ital Biol Sper. 2000 Jul; 76(7-8): 45-50
Pmid : 11449820
PubMed Alkaline Phosphatase is a MeSH term of this PMID

Score

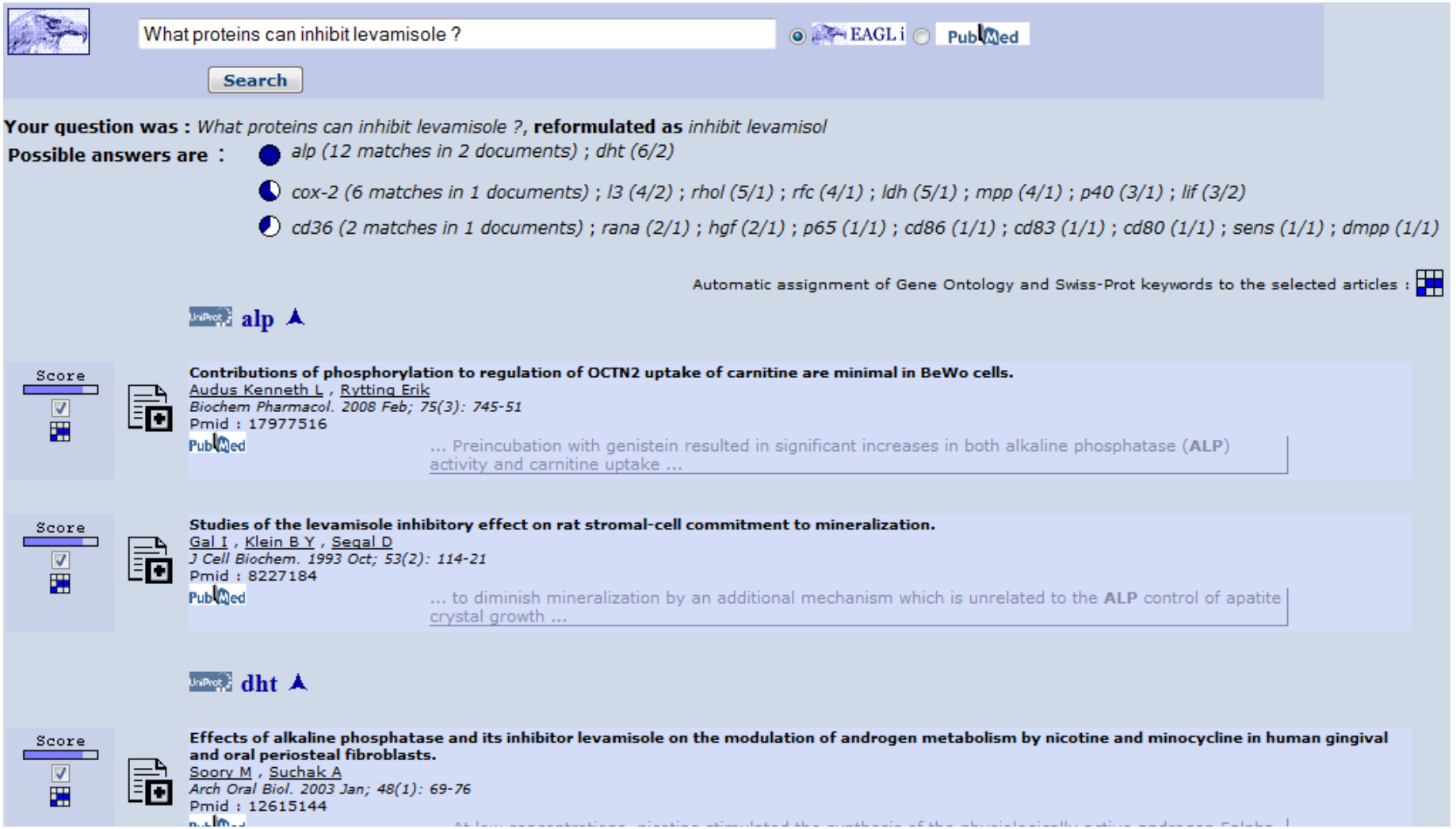
Inhibition of endogenous tissue alkaline phosphatase with the use of alkaline phosphatase conjugates in immunohistochemistry.
Ponder B A , Wilkinson M M
J Histochem Cytochem. 1981 Aug; 29(8): 981-4
Pmid : 7024402
PubMed ... Inhibition of endogenous tissue **alkaline phosphatase** with the use of alkaline phosphatase conjugates in immunohistochemistry ...
Alkaline Phosphatase is a MeSH term of this PMID

Score

Cholinesterase and phosphatase activities in adults and infective-stage larvae of levamisole-resistant and levamisole-susceptible isolates of Haemonchus contortus.
Giménez-Pardo C , Gómez-Barrio A , Martínez-Grueiro M M , Rodríguez-Caabeiro F
Vet Res Commun. 2003 Dec; 27(8): 611-23
Pmid : 14672450
PubMed ... Cholinesterase (ChE) and acid **phosphatase** (AP) activities, but not **alkaline** phosphatase activities, were detected in cytosolic and membrane ...
Alkaline Phosphatase is a MeSH term of this PMID

What enzymes/proteins can inhibit levamisole ?

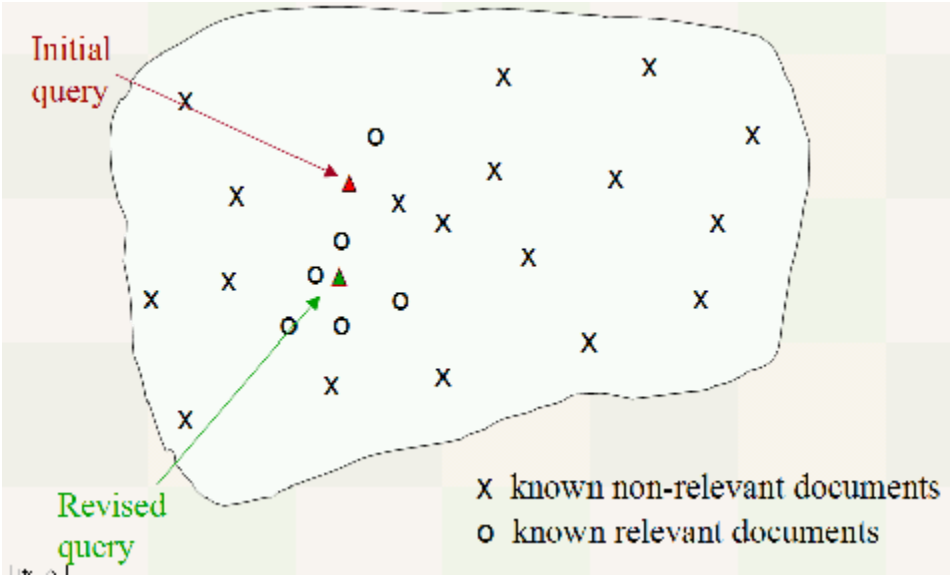
Importance de la question: enzymes → proteins !

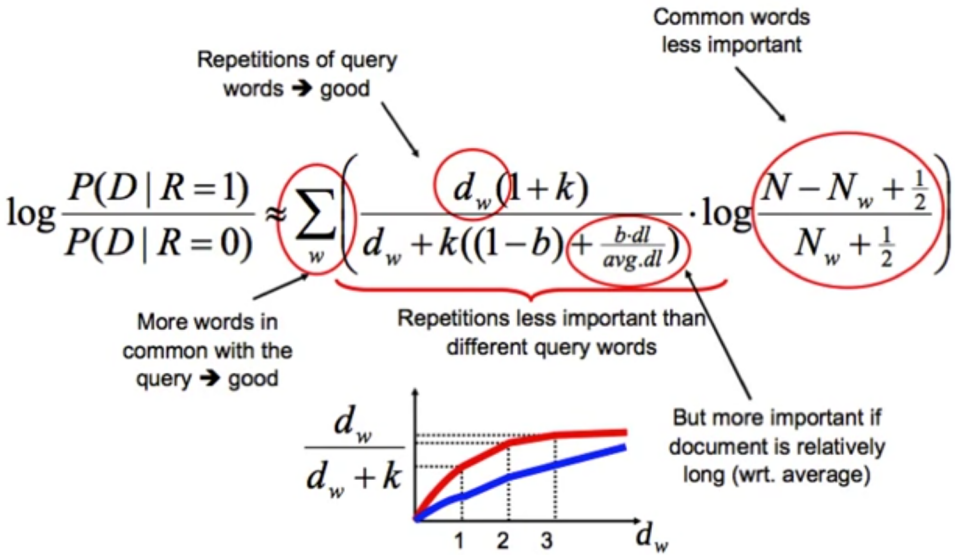


Search 4

- Query normalization / expansion
- Relevance feedback (reranking methods)
  - E.g. Rocchio (1971): automatic vs. interactive
  - Top-N documents and Top-M words
- Cross-references
  - Page rank
  - **Citations**
- Linear combinations of weighting schema (Fox and Shaw 1983)
- Learning, e.g. click-through

Rocchio – Okapi BM25 – Weighting Schema





Model	word	bigram (base)
PB2-nnn	0.2378	0.3729
LM	0.2120*	0.3310*
Okapi-npn	0.2245*	0.3630*
Lnu-ltc	0.2296	**0.3973***
dtu-dtn	**0.2411**	0.3673*
atn-ntc	0.2242*	0.3270*
ltn-ntc	0.2370	0.3708
ltc-ltc	0.1606*	0.2260*
ntc-ntc	0.1548*	0.2506*

Evaluations: relevance judgements

- Queries, min 25-40 (static)
- Large document set (static)
- Relationships

Query ID	0	Document ID	0	Relation
1	0	48679	0	1
2	0	1024195	0	1

“Cranfield paradigm”

Methodology

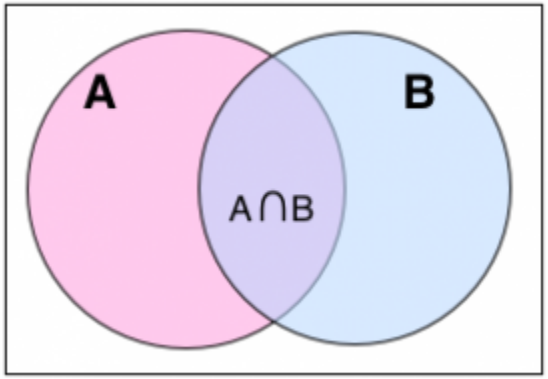
- “Pooling”
- Human assesment
- Inter-rater agreement (rare)

Results

Query ID	Document ID	Score (similarity)
1	56789	0.99
1	1024195	0.5
2	1024195	0.9

Score >> Boolean engine

Query ID	Document ID	Score (similarity)
1	56789	0.99
1	1024195	0.5
2	1024195	0.9



Non textual features: Citations

Impact of the co-citation network	Without re-ranking	5.36%	3.47%
	With re-ranking	**6.76%**	**4.22%**

- +20-25% improvement in prior art search of chemistry patents

https://pubmed.ncbi.nlm.nih.gov/24564220/

Triage

- Triage is a formally a search task, but input are more complex than keywords (e.g. variant sequences) and capturing evidence is key
  - Powered by ML classification
  - Powered by distance computation (~IR), e.g. similarity with a given dictionary
  - Variomes
    - http://candy.hesge.ch/Variomes/
    - http://goldorak.hesge.ch/synvar/generate/litterature/fromMutation?base=BRAF&variant=V600E&type=protein
  - Intrinsically Disordered Proteins
    - http://candy.hesge.ch/disprotGUI/

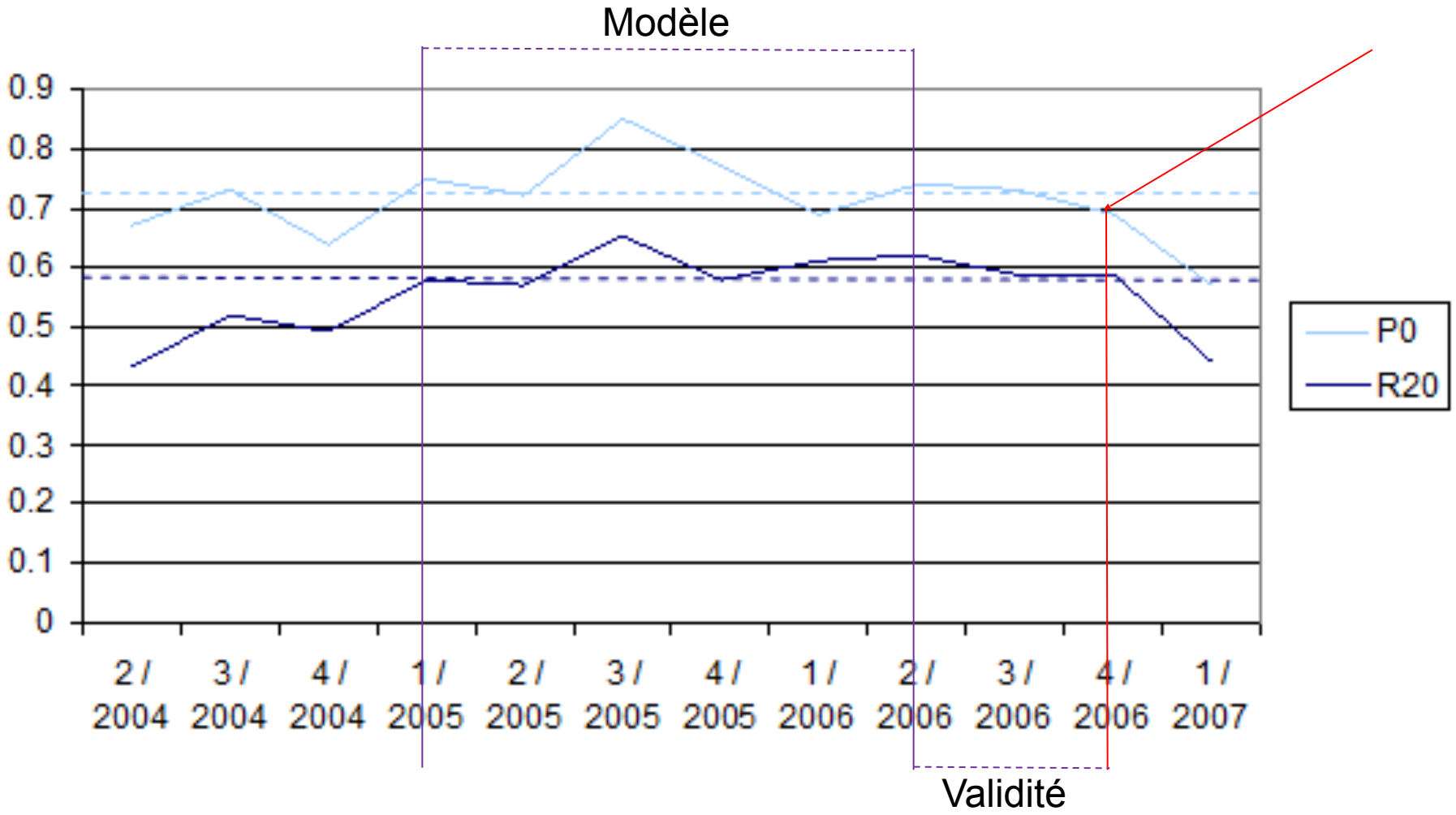
Keyword assignment (multiclass multilabel categorization)

- N >> 40000 classes
- Gene Ontology categorization

http://eagl.unige.ch/GOCat/

- Exposed to risky concept drifts…

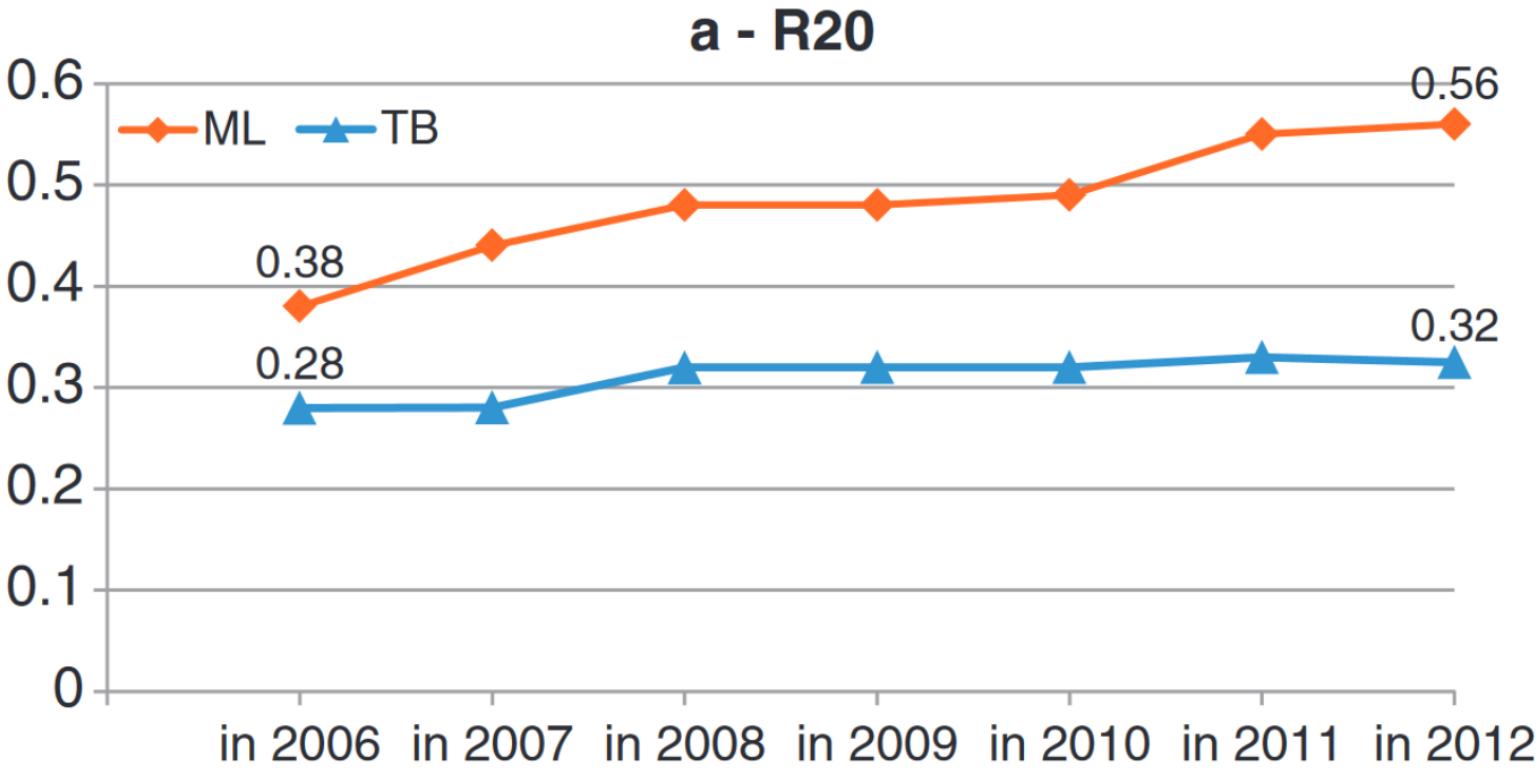
Typical concept drift / Model instability



→ ICD-10 Classification with Electr. Health Records

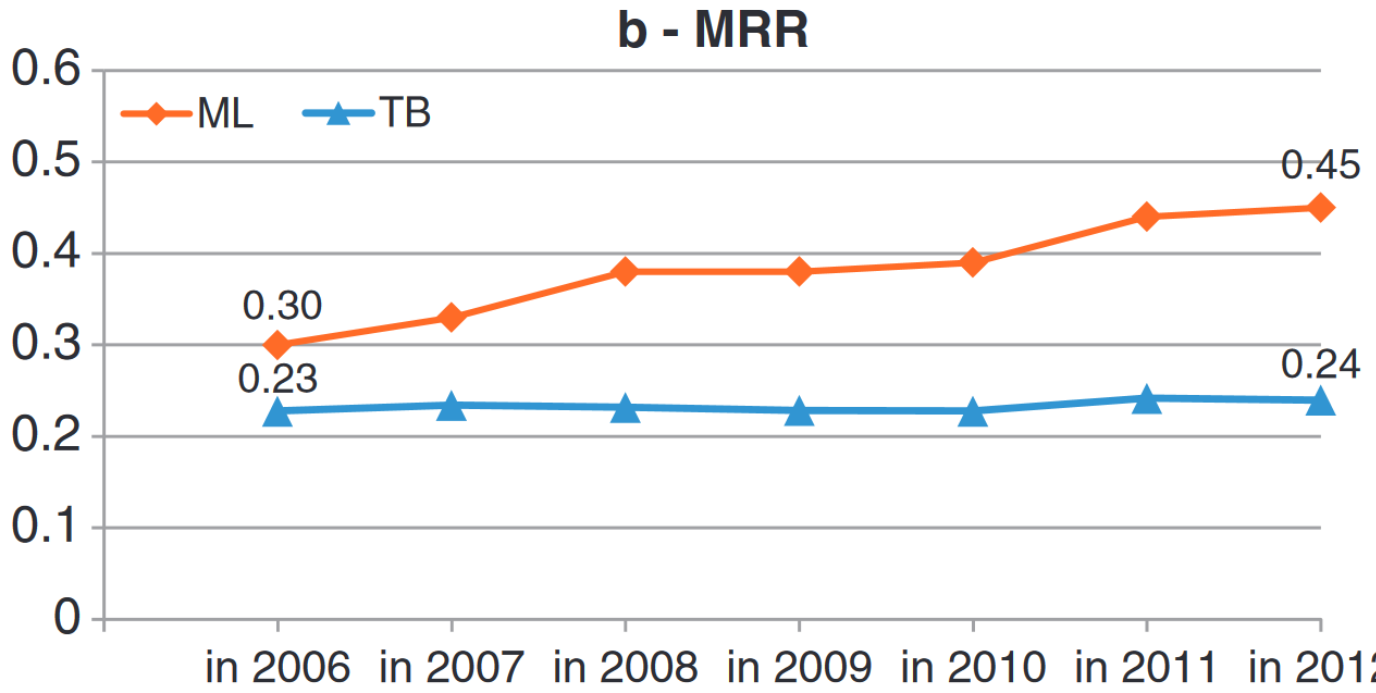
Gene Ontology: Recall

- 2006 to 2013: 19556 terms → 37070



Gobeill et al. 2013 & 2014

Gene Ontology: Precision



- On article from 2006 classified with a model from 2012 (with data from 2006-2012) is better classified than by any other models, incl. 2011, 2010, 09, 08, 07, **and 2006 !**
- **Inter-rater agreement ~40%...** so theoretical upperbound has been reached [Camon et al. 2003]

Challenge in merging text mining pipelines: biocuration !

1. Search

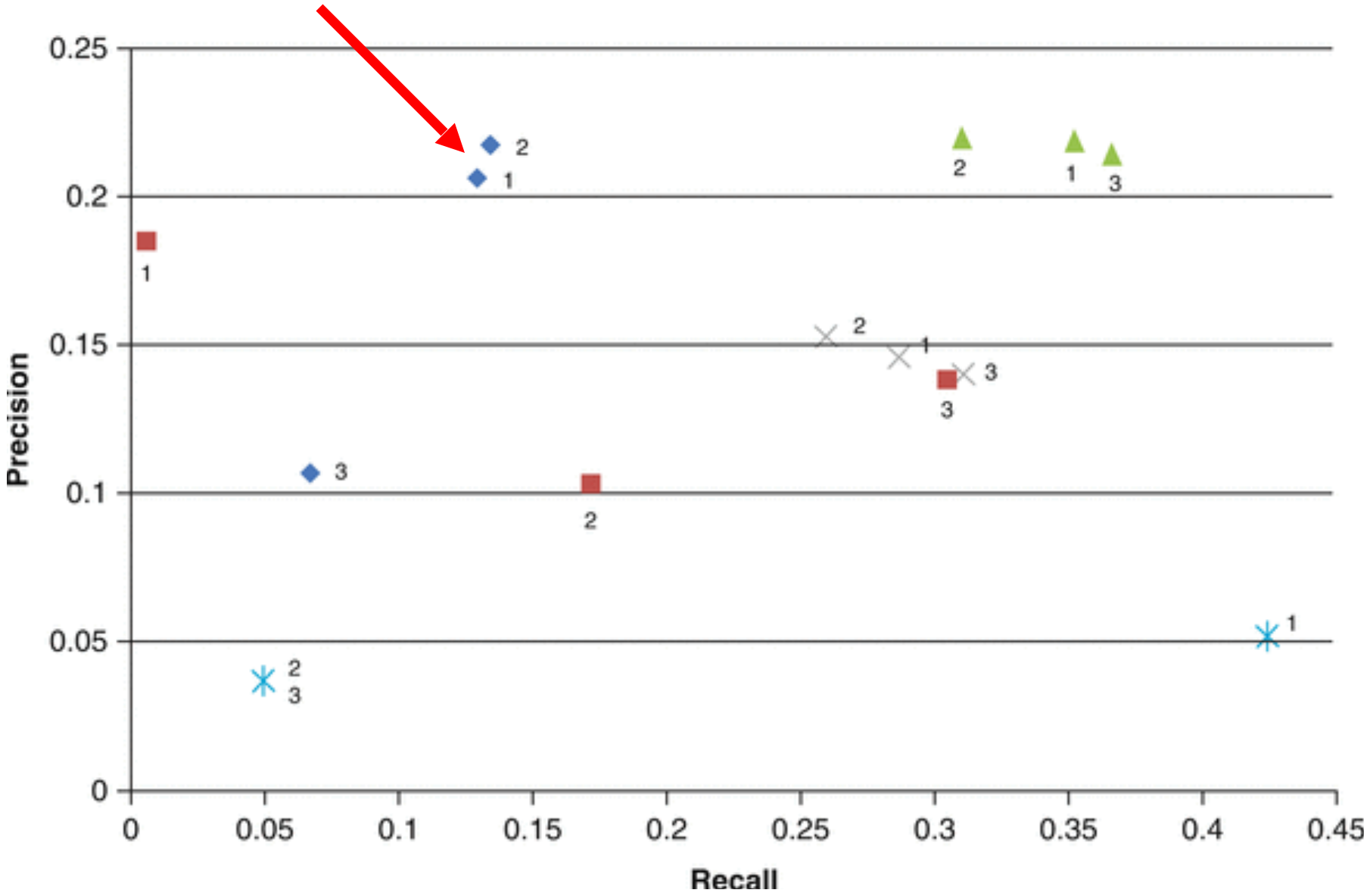
2. Passage extraction

3. Categorization

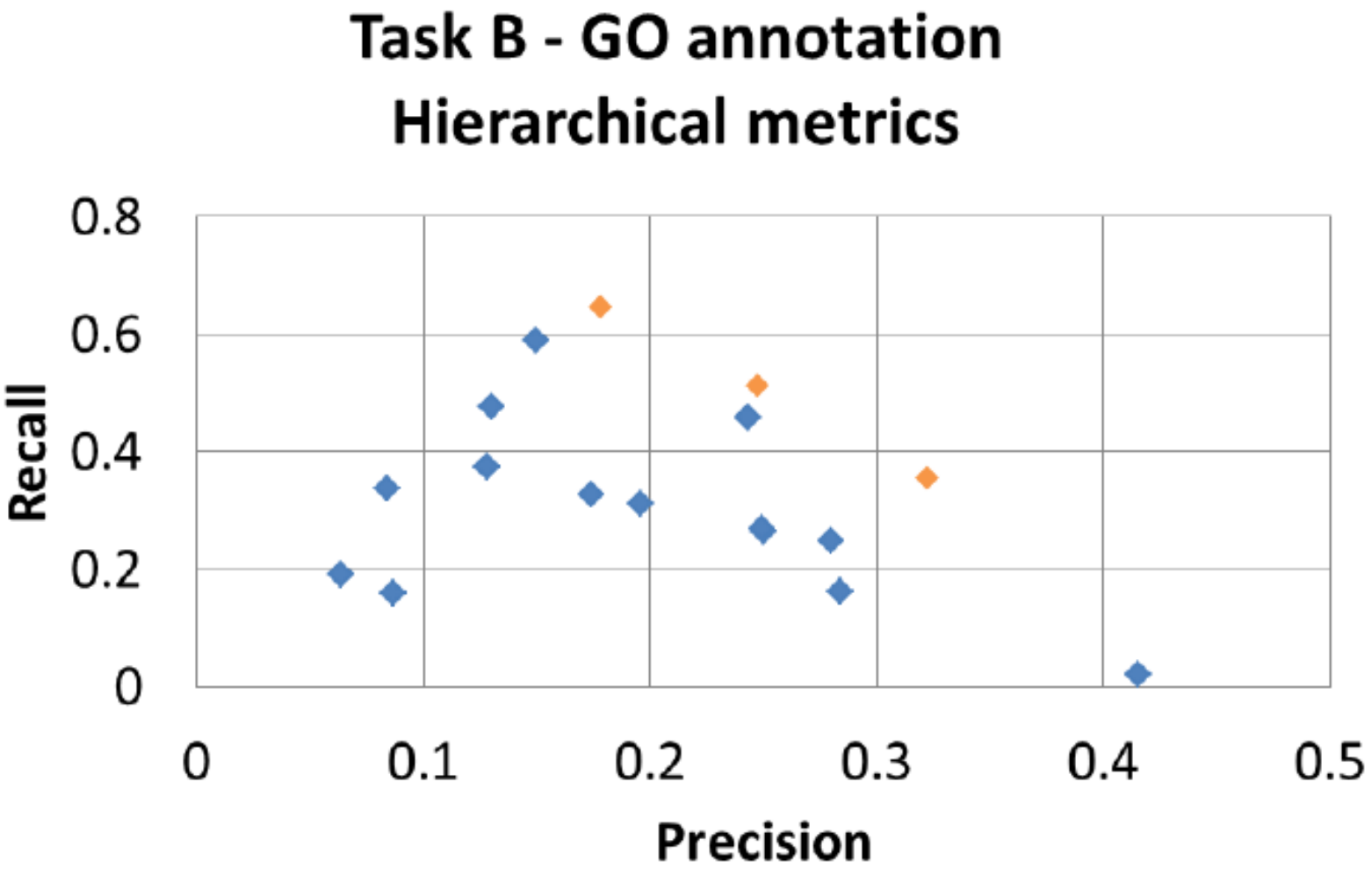
- https://pubmed.ncbi.nlm.nih.gov/27812936/
  Example of a good effort allocation between 2 & 3

NB: although GO has increased, no degradation in time of the model is observable ☺

Performance comparée @BioCreative IV



Performance comparée @BioCreative IV



Max 65% vs. Kappa ~40% (+225% BC#1)

http://eagl.unige.ch/GOCat/

Named entity recognition

- PubTator

https://www.ncbi.nlm.nih.gov/research/pubtator/?view=publication&pmid=33009979&query=v600e&page=1

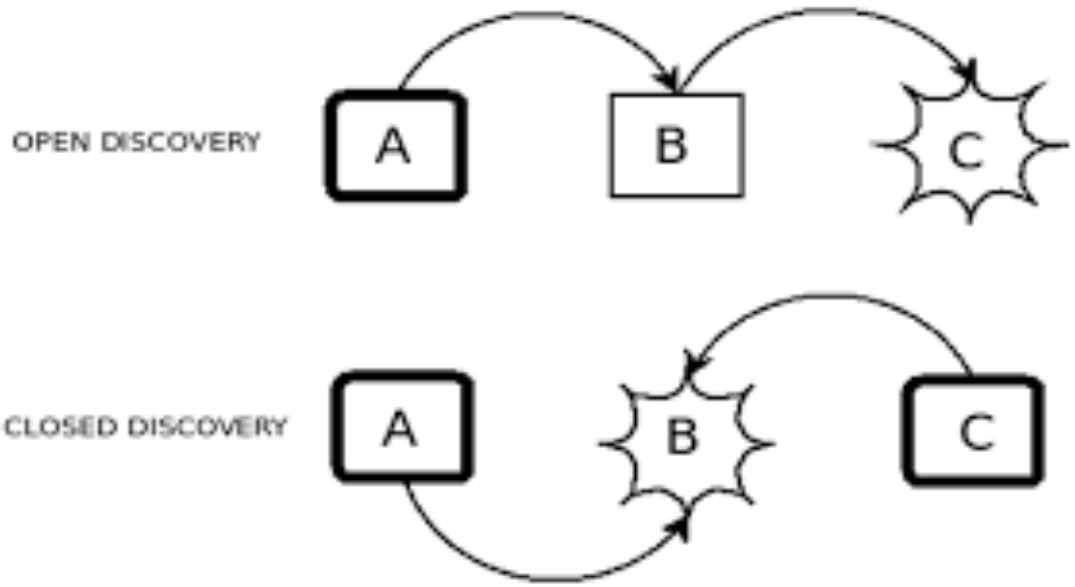
Extraction of more complex entities

- Protein protein interactions
- Mutation causing abnormalities
- […]
- GeneRiFs in EuropePMC

https://europepmc.org/article/MED/31024001

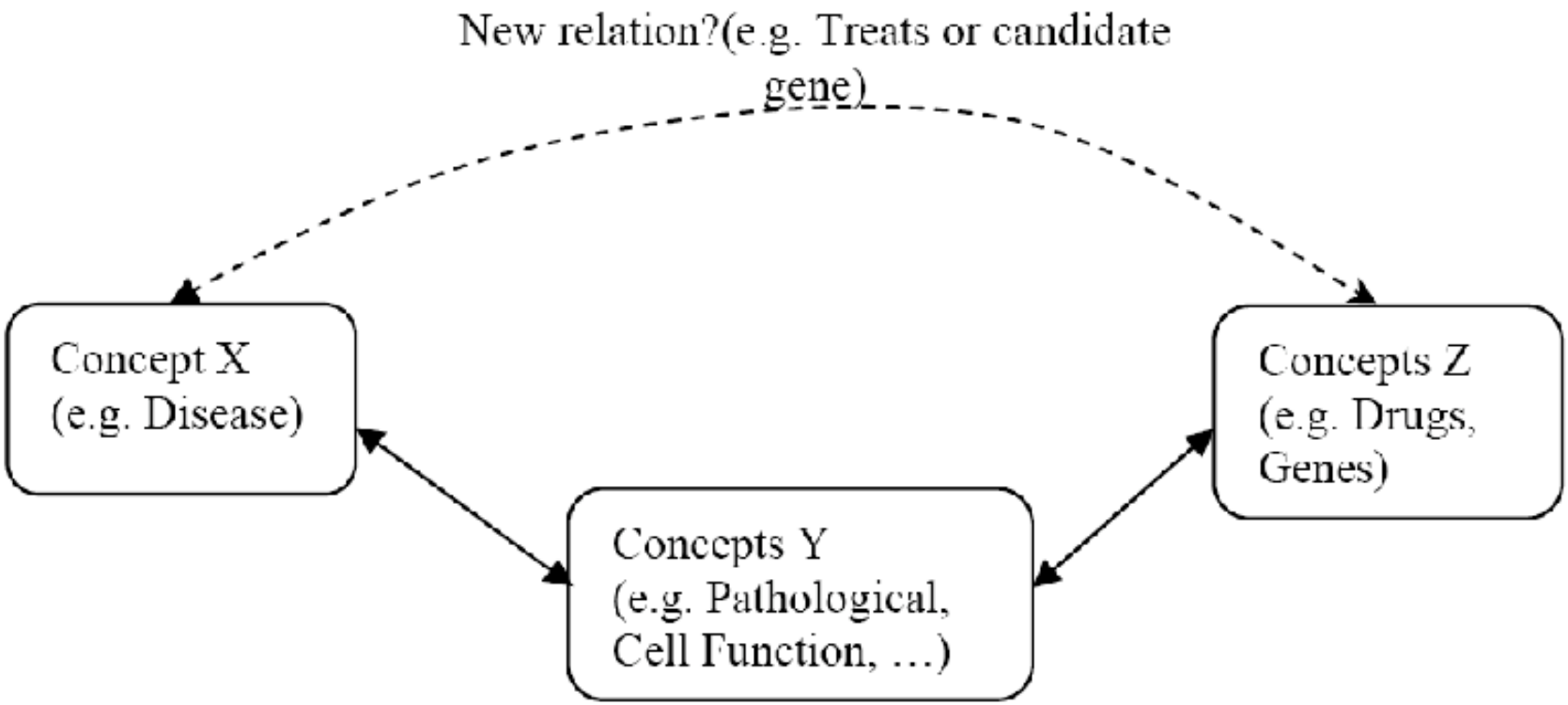
Literature based discovery → Hypothesis generation

- ABC model



*Swanson, Don (1988). "Migraine and Magnesium: Eleven Neglected Connections". Perspectives in Biology and Medicine. **31** (4): 526–557*

Typical LBD problem



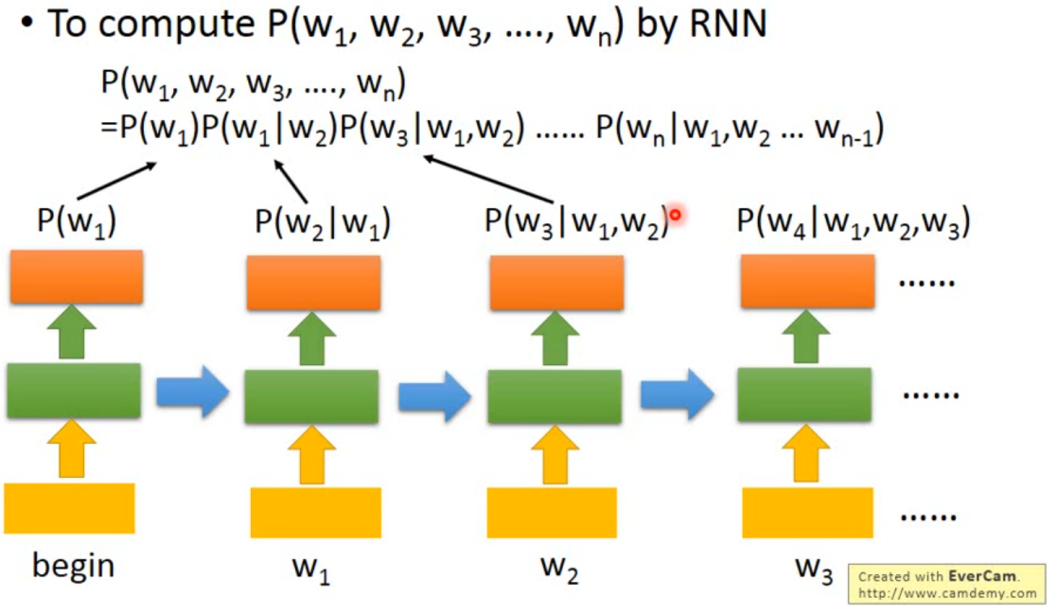
Technologies

- Rule-base systems
- Markovian models (HMM, CRF, LSTM, …)
- Classifiers (kNN, BN, SVM, …)
- **Deep neural networks → word embeddings, transformers, …**

Language models

- A statistical **language model** is a probability distribution over sequences of words. Given such a sequence, say of length m, it assigns a probability. to the whole sequence. The **language model** provides context to distinguish between words and phrases that sound similar.

Wpedia, Oct 5, 2020



Sometimes it does work… sometimes…

- **Experimentation of language models**

  CamemBERT pre-trained on French documents

- **Preliminary results are pretty good !**

  Accuracy : 0.86% (RECIST) and 0.91% (binary)

	Class	F1-score (CamemBERT)	F1-score (GradientBoosting)
RECIST	Complete Response	0.83	0.75
	Partial Response	0.79	0.88
	Stable Disease	0.90	0.89
	Progressive Disease	0.85	0.86
Binary	Non Progressive	0.91	0.89
	Progressive	0.90	0.87

What you need ?

- Large corpora
- Transformer architecture (word2vec, GloVe, BERT, …)
  - Tensor, PyTorch, Keras, …
- GPU
- Cash (50-100 000 $)

→ Model

Linguistic features

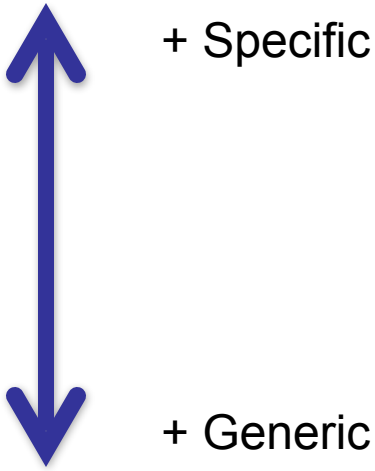
- All models assume a «bag od words» approach

→ Independence between words ☺

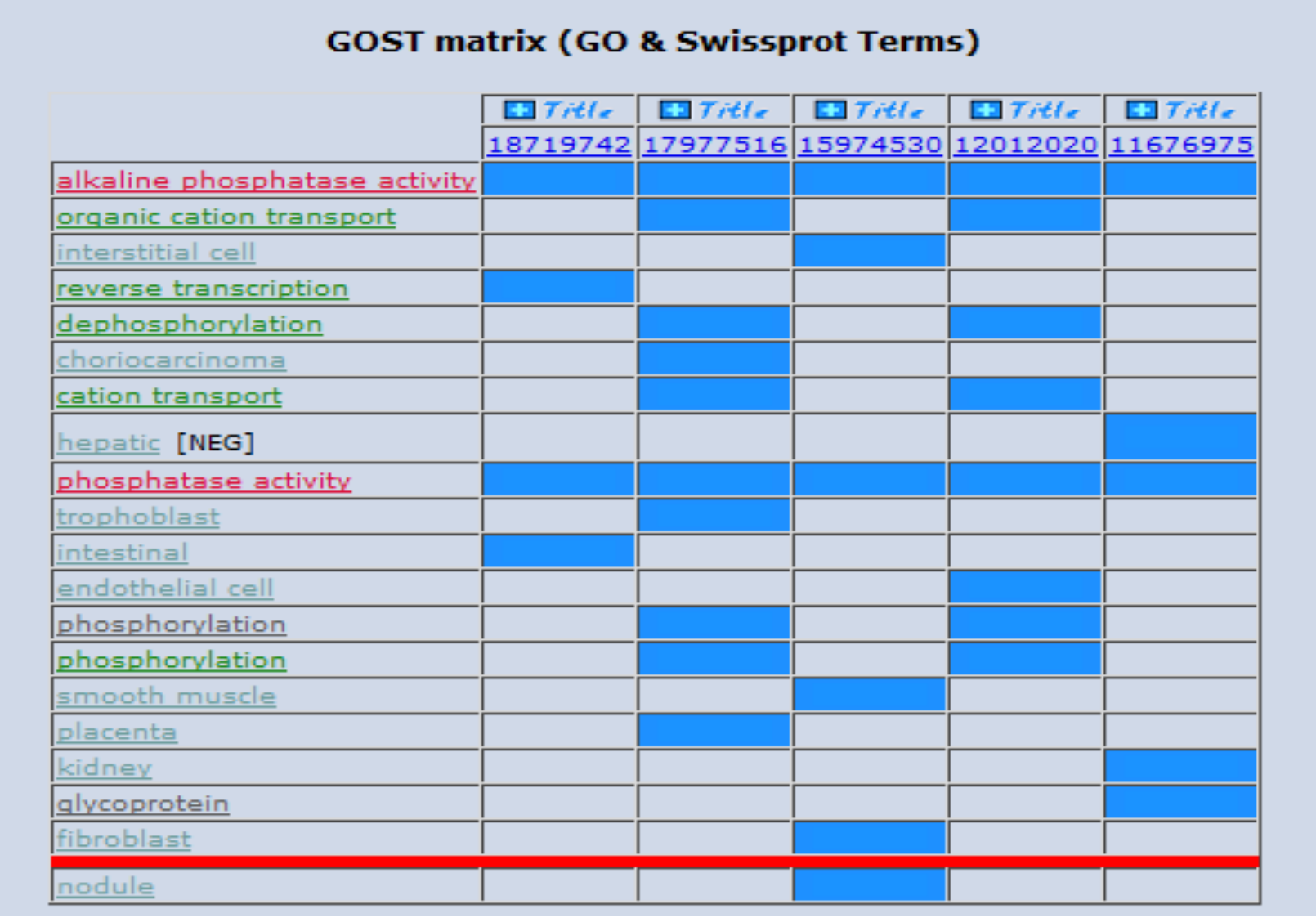
→ Models, which tries to compensate

→ **Linguistif features: negation, level of evidences, …**

- Alternatives
  - Terminologies
  - N-words
  - Words
  - Stems (subwords)



Natural language processing: negations and doxic features



Resources

- Information retrieval: TREC.nist.gov
- Named-entity recogniton: Biocreative
- Categorization MeSH: TREC/PubMed
- Categorization GO: BioCreative/GOA

Remerciements

Thank you very much !

Jung Yi Li [Language model schema]
Michael Baudis, organisation

Some more…

Methods

Singhal A (2001) Modern Information Retrieval: A Brief Overview. IEEE Data Eng. Bull. 24. p 35-43

Ruch P (2006) Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 22(6): 658-664 (2006)

Gene Ontology

Ruch P and Grabar N. (2007), Gene Ontologie : Un outil pour l'annotation fonctionnelle des gènes et de leurs produits, Omnisciences, PUF, Paris.

Camon E, Barrell D, Dimmer E, Lee V, Magrane M, Maslen J, Binns J, Apweiler R. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics, Vol. 6 Suppl 1 (2005)

IT Tools

SOLR (distribution de Lucene/moteur de recherche), Lemur, Terrier…
Protégé, OBOEdit (editeur d'ontologie)

Categorisers

- Precision

  Ratio between positive & returned categories and all returned categories
- Rappel

  Ratio between positive & returned categories and expected positive categories
- Law: F(Precision) = 1/F(Rappel)

Exercice

Expected categories	Returned categories
Diabetes	1. Human
Female	2. Female
	3. Type II Diabetes
	4. New born
babies	

Solution

- Precision = 0.25 (1/4) ; Recall = 0.5 (1/2)

Expected	Returned
Diabetes	1. Primipare
Female	2. Female
	3. Diabète insulino-dépendant
	4. Enfant normal unique

Schémas de pondération… [Okapi BM25: tf, df, dl]

$$Score(D_i, Q) = \sum_{t_j \in q} qtf \cdot \log\left(\frac{n - df_j}{df_j}\right) \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}}$$

$$\text{where } K = k_1 \cdot \left[(1 - b) + b \cdot (l_i / avdl)\right]$$