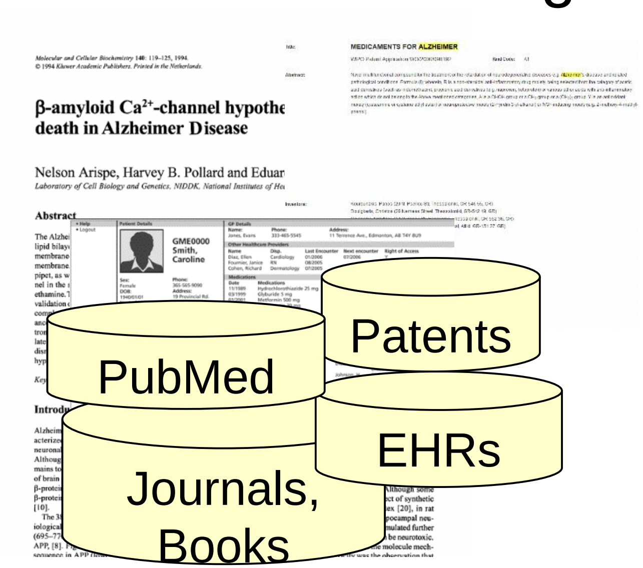
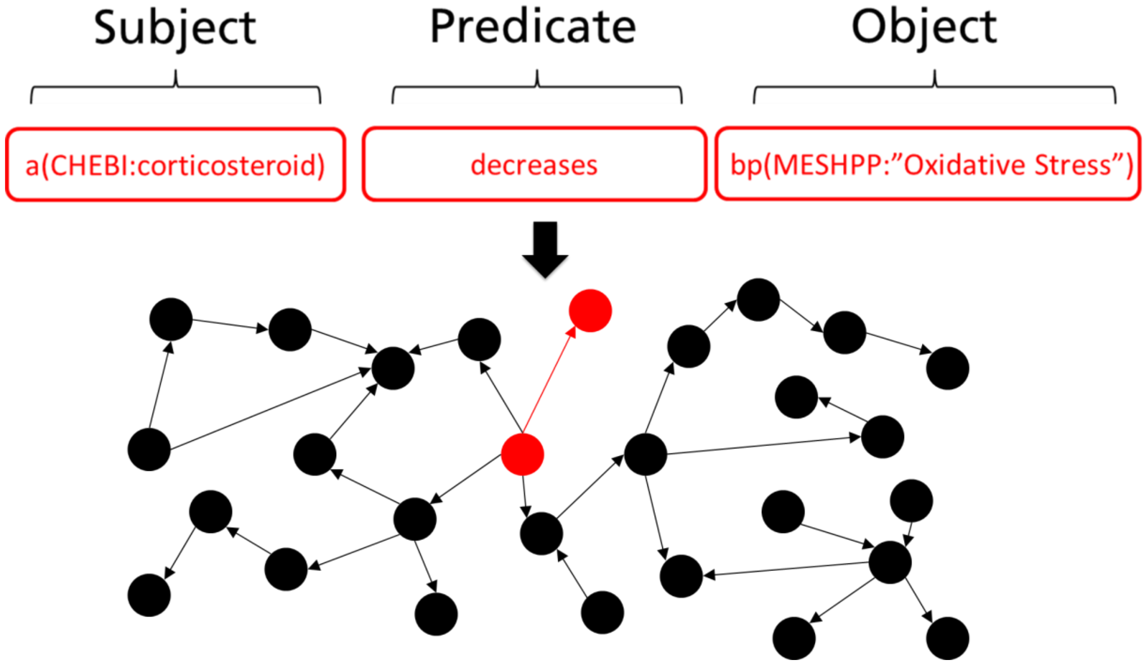
Bio-Medical Text Mining with Machine Learning

Sumit Madan

Department of Bioinformatics - Fraunhofer SCAI

Textual Knowledge


Molecular and Cellular Biochemistry 140: 119–125, 1994.
© 1994 Kluwer Academic Publishers. Printed in the Netherlands.
β-amyloid Ca2+-channel hypothe
death in Alzheimer Disease
Nelson Arispe, Harvey B. Pollard and Eduar
Laboratory of Cell Biology and Genetics, NIDDK, National Institutes of He
Abstract
Introdu
MEDICAMENTS FOR ALZHEIMER
GME0000
Smith,
Caroline
Patents
PubMed
EHRs
Journals,
Books



Subject

Predicate

Object

a(CHEBI:corticosteroid)

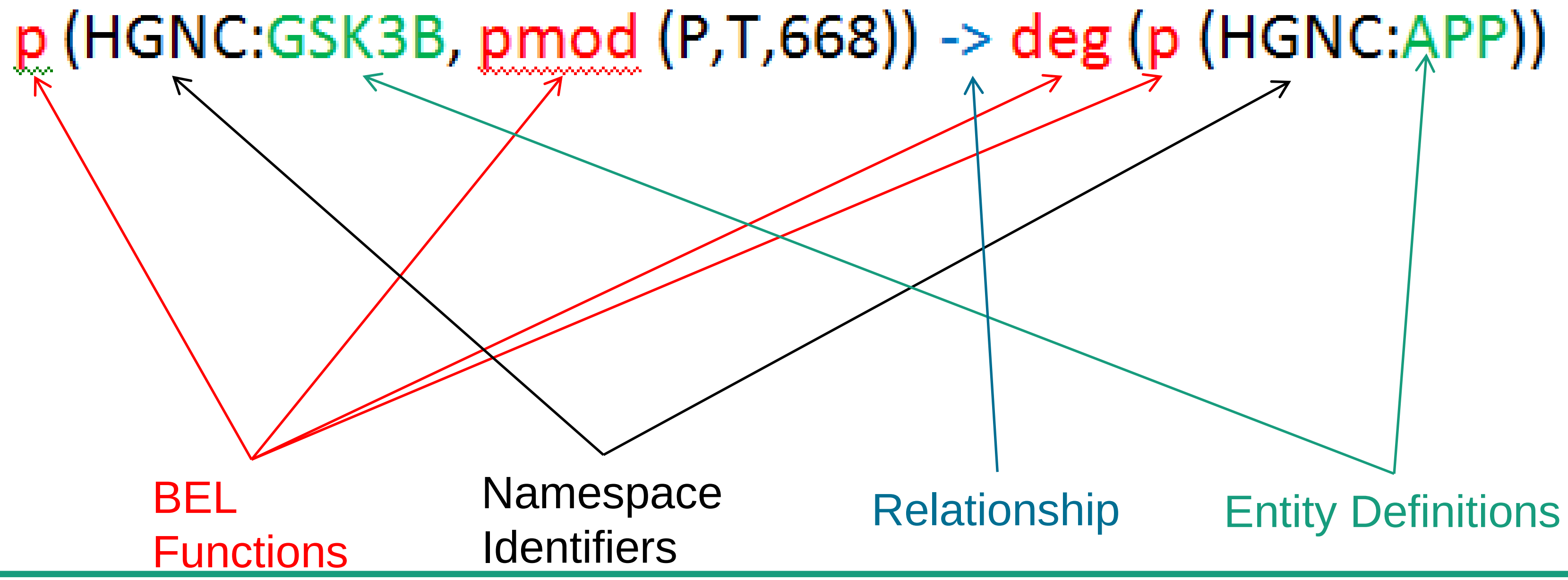
decreases

bp(MESHPP:"Oxidative Stress")

What is Bio-Medical Text Mining?

Phosphorylation of glycogen synthase kinase 3 beta at Threonine, 668 increases the degradation of amyloid precursor protein

Biological Expression Language (BEL)



p (HGNC:GSK3B, pmod (P,T,668)) -> deg (p (HGNC:APP))

BEL Functions

Namespace Identifiers

Relationship

Entity Definitions

Biological Entity Classes

Protein / Gene

- HGNC
- ENTREZGENE
- UNIPROT
- MGI
- RGD
- …

Small RNA

- MIRBASE
- piRNABank
- …

Further Classes

- **Organism** (NCBITaxonomy)
- **Anatomy** (UBERON)
- **Phenotype** (HP)
- **Biological Process** (GO)
- **Cell** (CL)
- **Cell lines** (CLO)
- **Several clinical features**
- **…**

Chemical

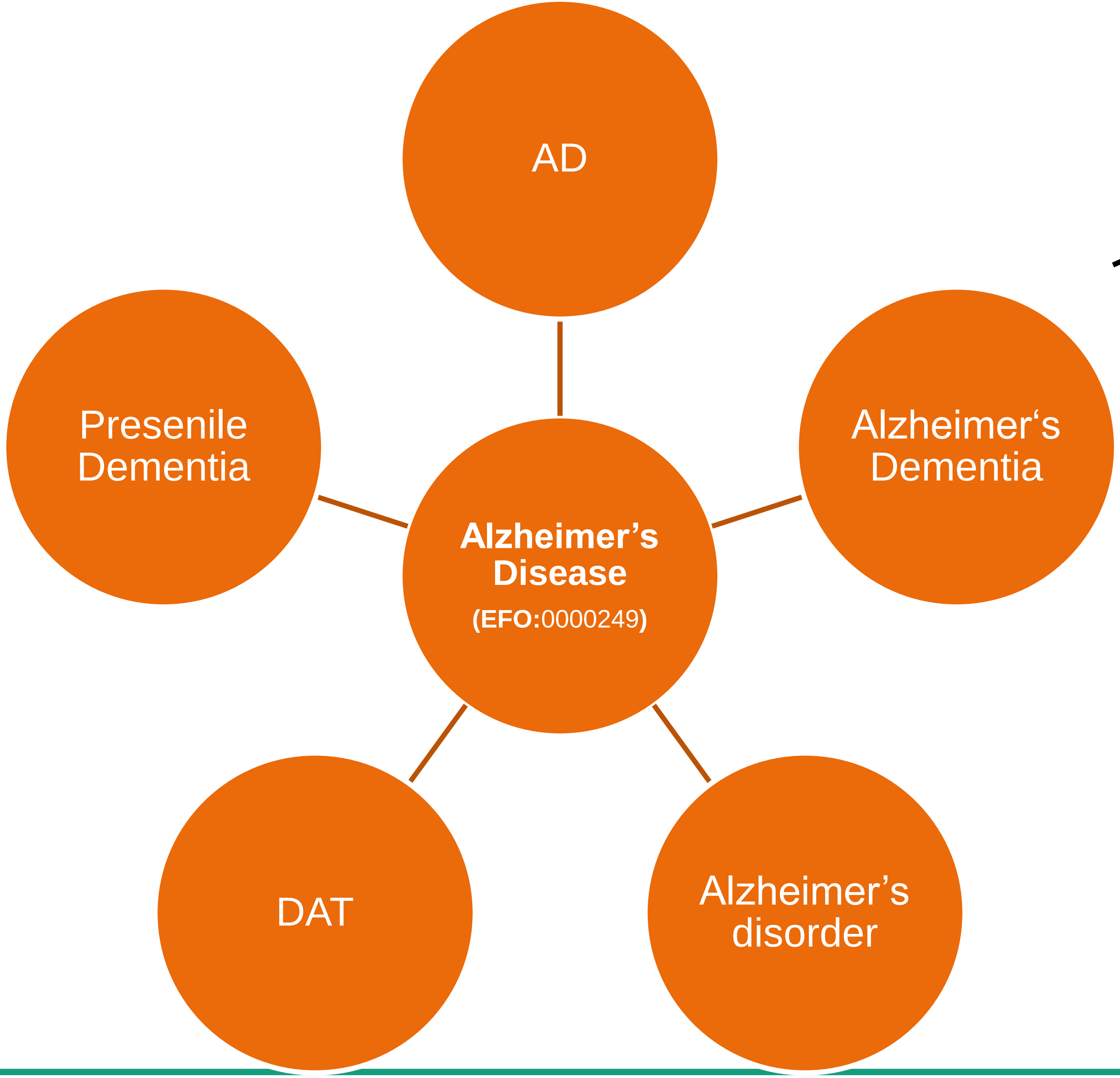
- MESH
- ChEBI
- ChEMBL
- …

Disease

- MESH
- DO
- ICD10
- …

Bold: Entity Class
Normal: Controlled Vocabulary

Several Spelling Variants (Synonyms) for One Single Concept



AD

Presenile Dementia

Alzheimer‘s Dementia

Alzheimer’s Disease

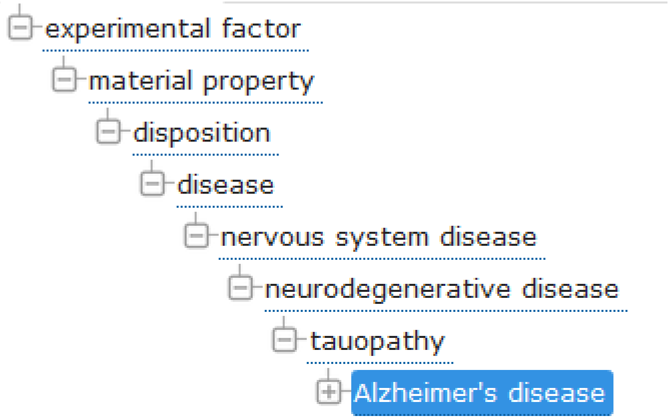
(EFO:0000249)

DAT

Alzheimer’s disorder

Over 28 Synonyms available just in Experimental Factor Ontology (EFO).

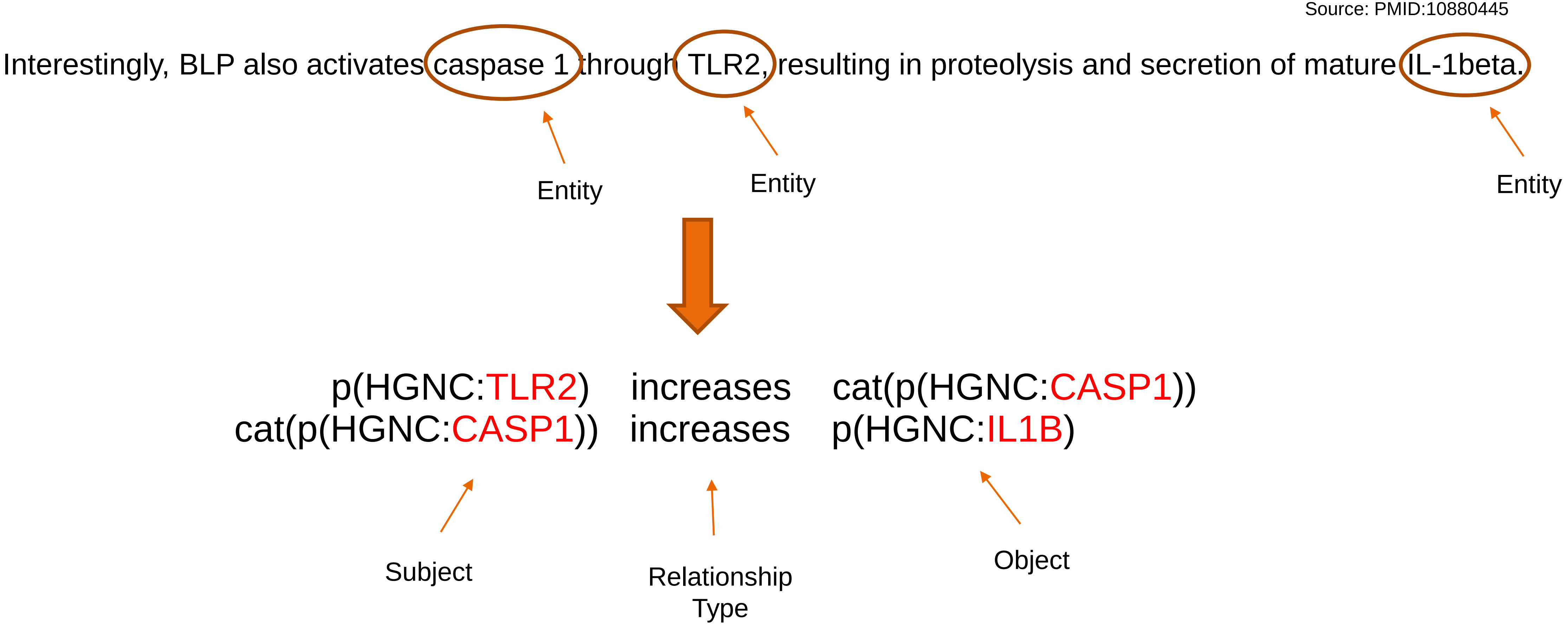
Alzheimer’s Disease (EFO:0000249)



experimental factor
material property
disposition
disease
nervous system disease
neurodegenerative disease
tauopathy
Alzheimer's disease

Source: http://www.ebi.ac.uk/efo/EFO_0000249

Biological Relations (Example)



Source: PMID:10880445

Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta.

p(HGNC:TLR2) increases cat(p(HGNC:CASP1))
cat(p(HGNC:CASP1)) increases p(HGNC:IL1B)

Relation Extraction Workflow based on Convolutional Neural Network

Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta (Source: PMID:10880445)

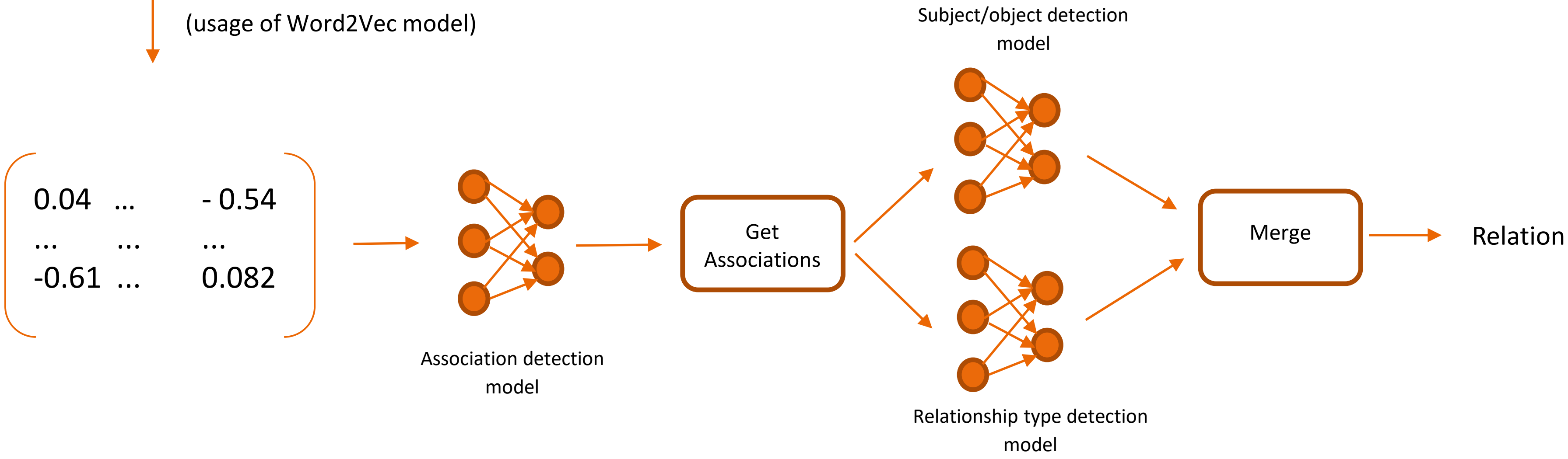
↓ Named Entity Recognition (NER)

Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta

↓ Pre-processing

Interestingly, BLP also activates ENTITY-1 through ENTITY-2, resulting in proteolysis and secretion of mature IL-1beta

↓ Sentence-matrix creation
(usage of Word2Vec model)



$$\begin{pmatrix} 0.04 & \dots & -0.54 \\ \dots & \dots & \dots \\ -0.61 & \dots & 0.082 \end{pmatrix}$$

Association detection model

Get Associations

Subject/object detection model

Relationship type detection model

Merge

Relation

Ali, M., Madan, S., Fischer, A., et al. (2017) **Automatic Extraction of BEL-Statements based on Neural Networks**, *Proc. BioCreative VI Chall. Work.*

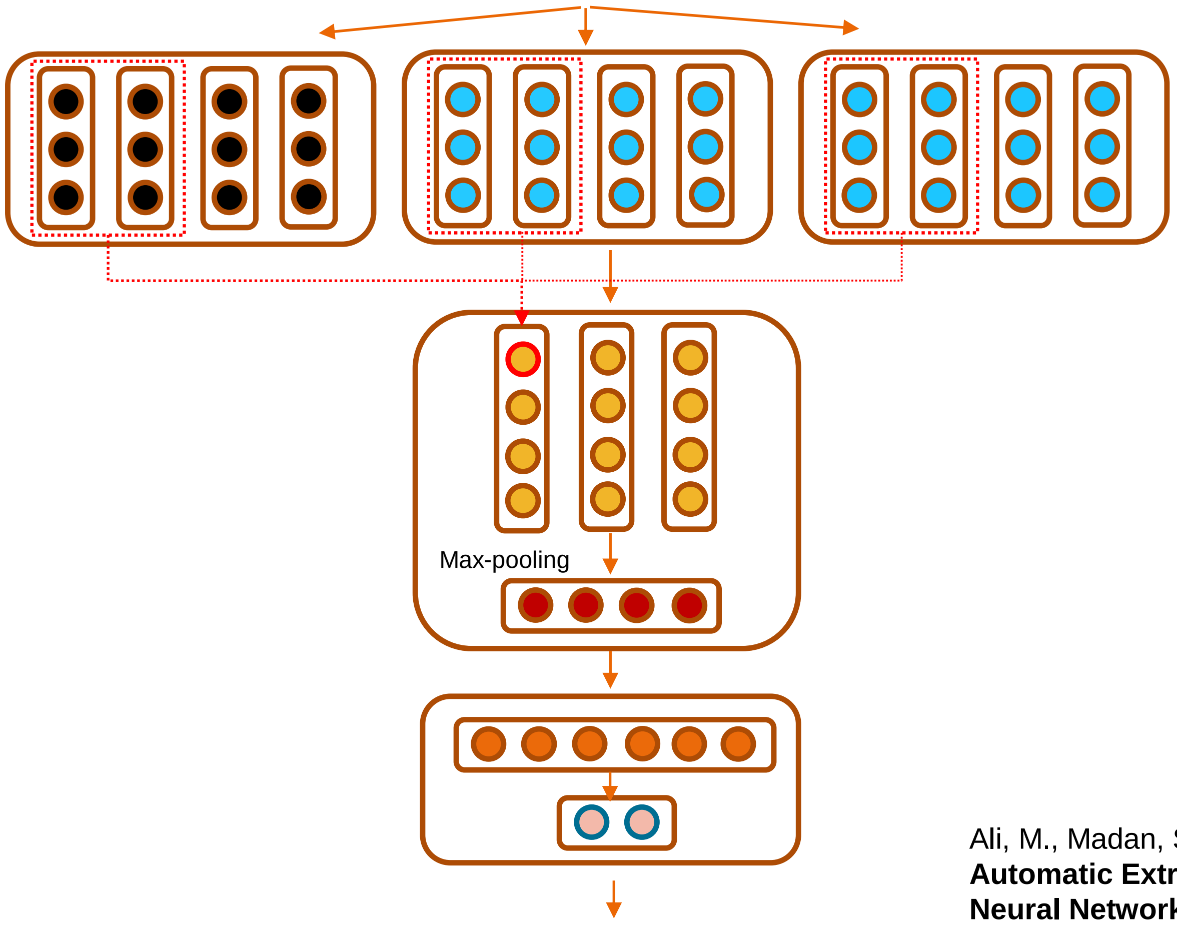
Multichannel Convolutional Neural Network (CNN) Architecture

Interestingly, BLP also activates ENTITY-1 through ENTITY-2, resulting in proteolysis and secretion of mature IL-1beta (PMID:10880445)

Embedding layer

Convolution and max-pooling

Fully-connected and classification

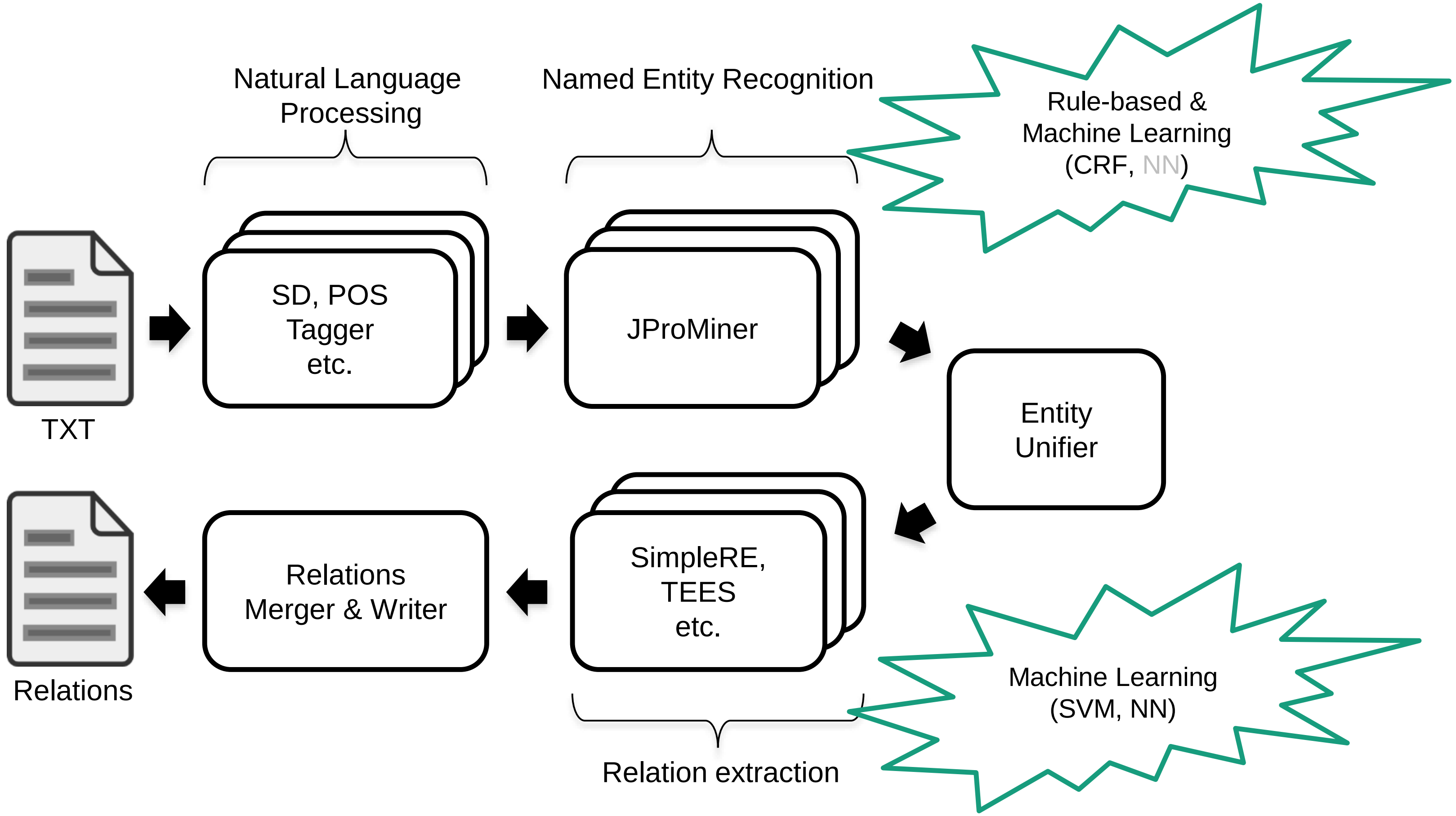


Max-pooling

Predictions

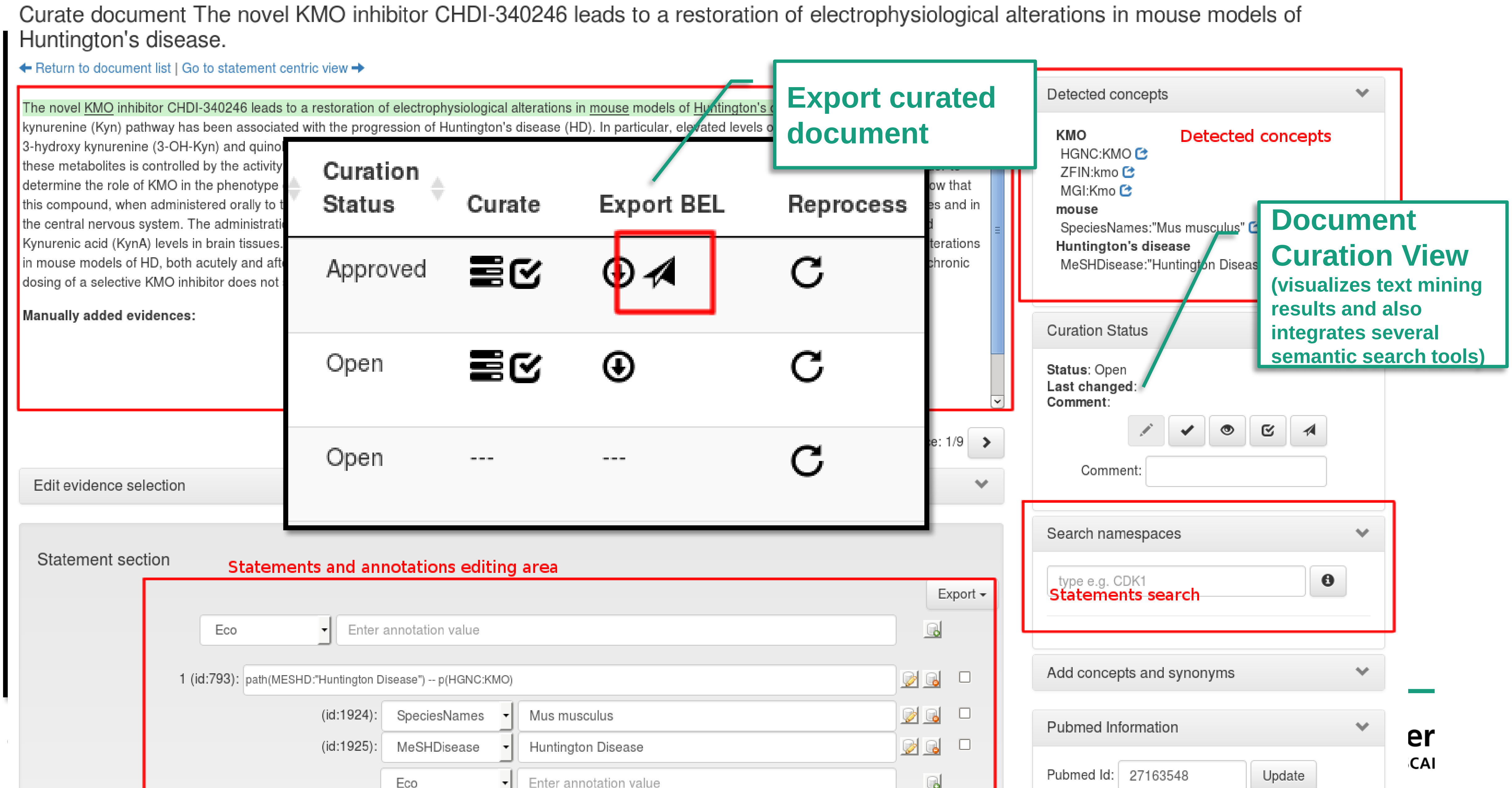
Ali, M., Madan, S., Fischer, A., et al. (2017)
Automatic Extraction of BEL-Statements based on Neural Networks, *Proc. BioCreative VI Chall. Work.*

Larger Workflow for Relation Extraction (Project BELIEF)


Natural Language Processing
Named Entity Recognition
Rule-based & Machine Learning (CRF, NN)
TXT
SD, POS Tagger etc.
JProMiner
Entity Unifier
Relations
Relations Merger & Writer
SimpleRE, TEES etc.
Machine Learning (SVM, NN)
Relation extraction

Madan, S., Hodapp, S., Senger, P., et al. (2016) **The BEL information extraction workflow (BELIEF): evaluation in the BioCreative V BEL and IAT track**, *Database*, 2016, baw136.
http://belief.scai.fraunhofer.de/BeliefDashboard/

BELIEF Dashboard – Functionalities of Web-based Curation Interface



Curate document The novel KMO inhibitor CHDI-340246 leads to a restoration of electrophysiological alterations in mouse models of Huntington's disease.

← Return to document list | Go to statement centric view →

The novel KMO inhibitor CHDI-340246 leads to a restoration of electrophysiological alterations in mouse models of Huntington's d
kynurenine (Kyn) pathway has been associated with the progression of Huntington's disease (HD). In particular, elevated levels o
3-hydroxy kynurenine (3-OH-Kyn) and quino
these metabolites is controlled by the activity
determine the role of KMO in the phenotype ... ow that
this compound, when administered orally to t ... es and in
the central nervous system. The administrati ... d
Kynurenic acid (KynA) levels in brain tissues. ... terations
in mouse models of HD, both acutely and aft ... chronic
dosing of a selective KMO inhibitor does not

Manually added evidences:

Curation Status	Curate	Export BEL	Reprocess
Approved	☰ ☑	⬇ ➤	↻
Open	☰ ☑	⬇	↻
Open	---	---	↻

Export curated document

e: 1/9 ›

Edit evidence selection

Statement section

Statements and annotations editing area

Export ▾

Eco | Enter annotation value

1 (id:793): path(MESHD:"Huntington Disease") -- p(HGNC:KMO) □

(id:1924): SpeciesNames | Mus musculus □

(id:1925): MeSHDisease | Huntington Disease □

Eco | Enter annotation value

Detected concepts

Detected concepts

KMO
HGNC:KMO
ZFIN:kmo
MGI:Kmo
mouse
SpeciesNames:"Mus musculus"
Huntington's disease
MeSHDisease:"Huntington Diseas

Curation Status

Status: Open
Last changed:
Comment:

Comment:

Document Curation View (visualizes text mining results and also integrates several semantic search tools)

Search namespaces

type e.g. CDK1

Statements search

Add concepts and synonyms

Pubmed Information

Pubmed Id: 27163548 Update

Outlook

- Usage of bidirectional recurrent neural networks (BiLSTM)

- Use more training data from different tasks (such as BioNLP, and also BioCreative)
- Create further models to predict biological functions
- Automate hyper-parameter optimization
- Train new optimized word2Vec models (use PubMed, PMC & ontologies)
- Build hybrid systems: machine learning + rules-based system (UIMA Ruta)