

# Biomedical Text Mining

## Hands-on Session

AI4Health Winter School  
7-8th January 2021

Chryssa Zerva  
Fenia Christopoulou

[ai4healthschool.org](http://ai4healthschool.org)



The University of Manchester



# NaCTeM

## Research Staff



## Students



# Current Research Topics

- Named Entity Recognition
  - Entity Normalisation
  - Coreference Resolution
- Relation Extraction
  - Supervised Sentence/Document-level
  - Detection of Temporal links
  - Unsupervised/Open
- Event Extraction
  - Nested Event Extraction
  - Uncertainty Identification
  - Document level EE
- Cancer omics text mining
  - Screening for systematic reviews
  - Scientific citation extraction
    - Scientific summarisation
  - Misleading information detection
    - Corroborating/contradicting evidence
    - Rank by trust/confidence
  - Development of annotated corpora
    - Active/Pro-active learning
    - Crowdsourcing
  - Emotion Detection
    - Social media for ADR
  - Risk Assessment

# Overview

1st part	Introduction to BioNLP Tasks and related applications	Application demos and try-out
2nd part	Theoretical background: <ul style="list-style-type: none"><li>- language representations</li><li>- model architectures</li></ul>	
3rd part	Hands-on session: NE, RE & EE on biomedical data	Move to colab!
4th part	Hands-on session: NE, RE & EE on biomedical data	Move to colab!

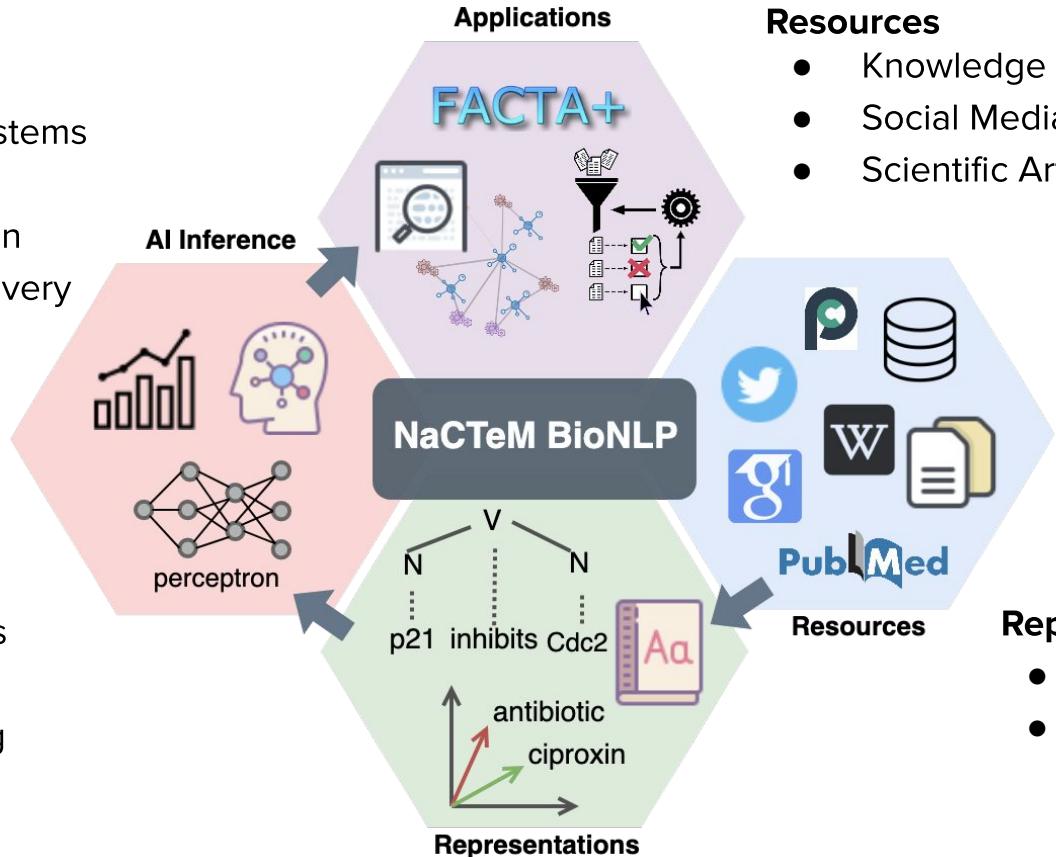
# Introduction

## Applications

- Search-based Systems
- Pathway Models
- Automatic curation
- Knowledge Discovery

## Inference

- Statistic Models
- Language Models
- NLP
- Machine Learning



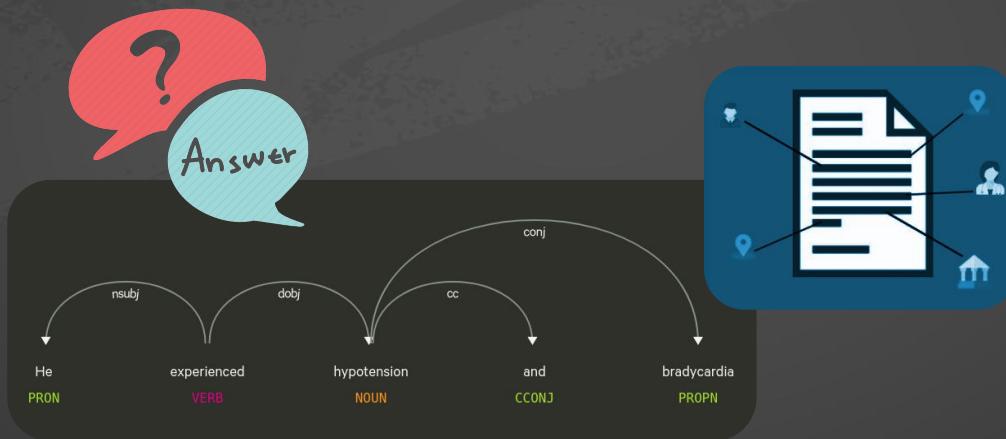
## Resources

- Knowledge Bases
- Social Media
- Scientific Articles

## Representations

- Embeddings
- Language Models

# BioNLP Tasks



# BioNLP Tasks

## Key Tasks

*Covered in this session*

- Named Entity Recognition
- Relation Extraction
- Event Extraction

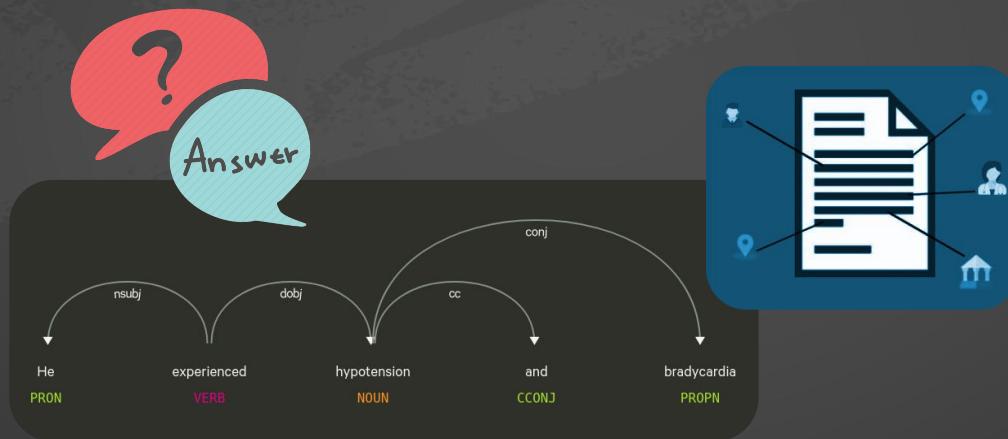
## Additional Tasks

*Briefly mentioned in this session*

- Entity Normalisation
- Metaknowledge/Modality/Negation
- Text Summarisation
- Question Answering

Key

# BioNLP Tasks



# Named Entity Recognition

→ The task that aims to **identify named entities** in **text**

EGTA inhibited down-regulation of alpha-AR mRNA by NE.

Chemical

Protein

**Input:** A textual snippet (typically a sentence)

**Output:** A label for each token in the sentence

## Definitions:

- **Named Entity:** A word or a group of words that constitute a proper name (e.g. EGTA)
- **Entity Type:** Semantic category in which a named entity belongs (e.g. Chemical)
- **Entity Span:** The words that are included in the entity (e.g. alpha-AR: 57-65)

## Notes:

- Named Entity types are typically *pre-defined*
- Current systems have reached almost human-level performance on NER  
But, when using very *general, high-level categories* ([check](#))
- *Tokenization* is essential before NER!

# NER: Tagging Schemes

- NER is treated as **token classification problem**
- Each word needs to have a certain label

**Input:** A textual snippet (typically a sentence)

**Output:** A label for each token in the sentence

1. **BIO** (Begin, Inside, Outside) ([Sang ET AL., 2000](#)) → CoNLL Shared Task (also found as IOB)

<b>EGTA</b>	inhibited down - regulation of <b>alpha</b>	-	<b>AR</b>	mRNA by NE
B-Chemical	O O O O O	B-Protein	I-Protein I-Protein	O O O

2. **IOE** (Inside, Outside, End)

<b>EGTA</b>	inhibited down - regulation of <b>alpha</b>	-	<b>AR</b>	mRNA by NE
E-Chemical	O O O O O	I-Protein	I-Protein E-Protein	O O O

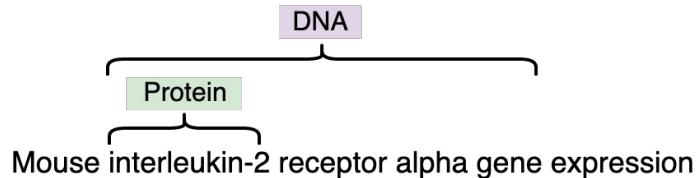
3. **IOBES** (Inside, Outside, Begin, End, Single) or **BILOU** (Begin, Inside, Last, Outside, Unit)

<b>EGTA</b>	inhibited down - regulation of <b>alpha</b>	-	<b>AR</b>	mRNA by NE
S-Chemical	O O O O O	B-Protein	I-Protein E-Protein	O O O

# NER: Usefulness & Challenges

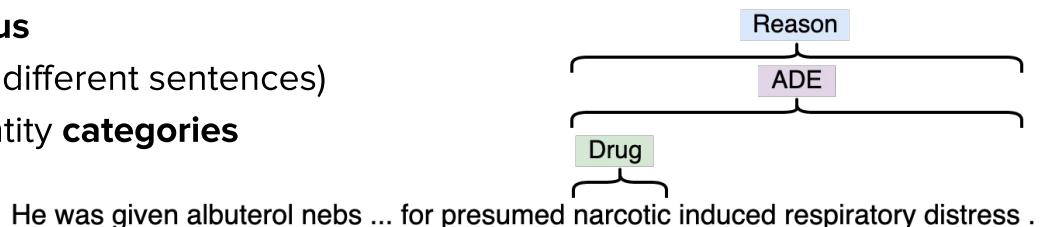
## Usefulness:

- Initial component of many downstream tasks (Relation Extraction, QnA, etc)
- All existing applications use some kind of NER algorithm (recommendation systems, etc)
- Customer support → find what the customer requests
- Classification → based on contained entities



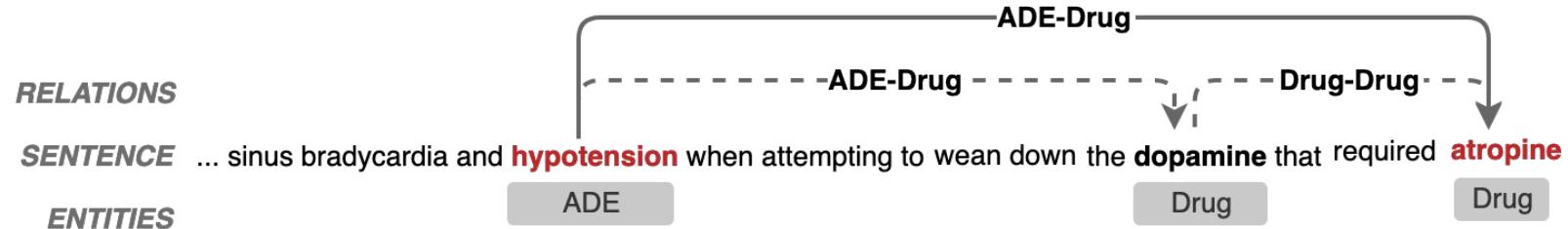
## Challenges:

- Biomedical text contains a lot of **nested entities**
- Biomedical entities are very **ambiguous**
- Additional **context** required (e.g. from different sentences)
- Difficult to detect very **fine-grained** entity **categories**



# (Textual) Relation Extraction

An **umbrella term** that describes the **identification of interactions between elements in text**



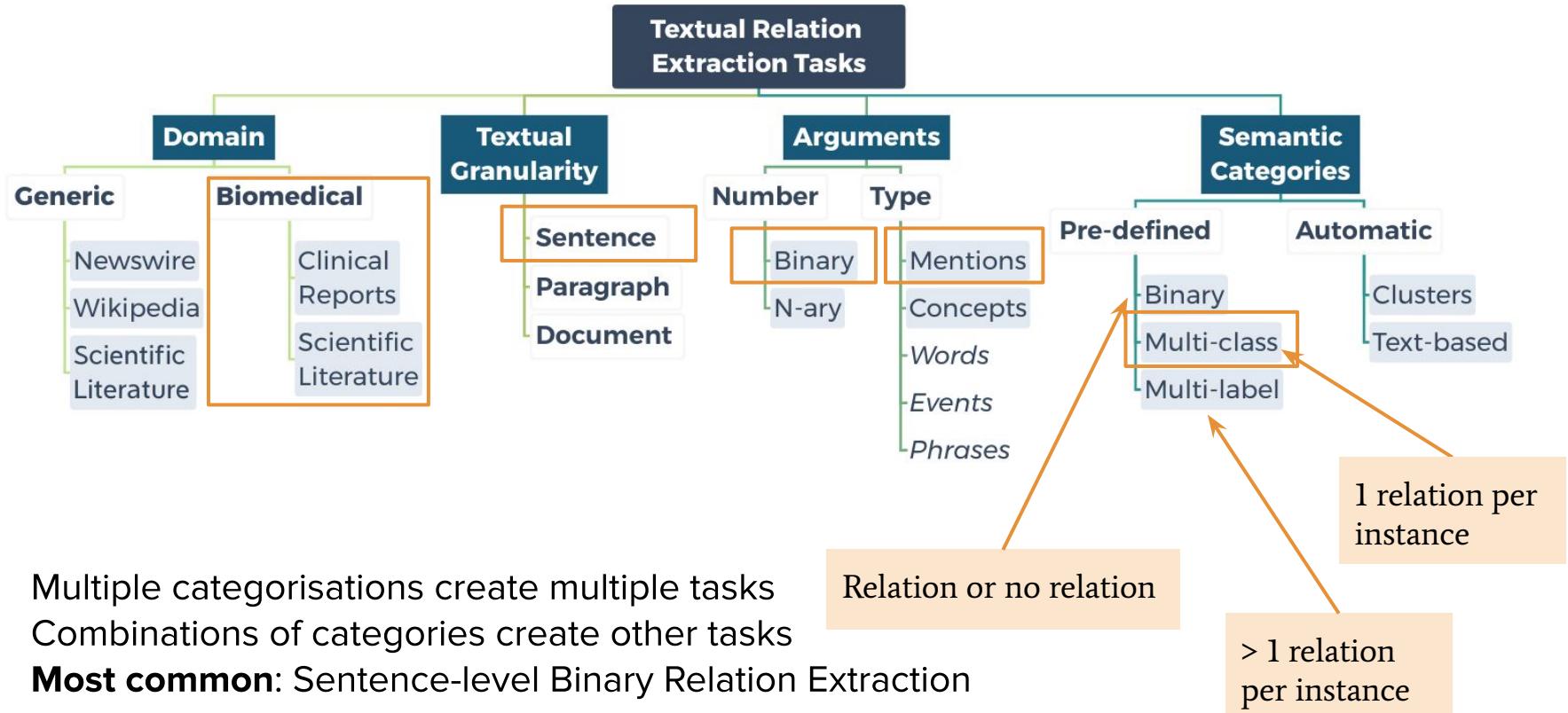
## Definitions

- **Named Entity:** A word or a group of words that constitute a proper name (e.g. hypotension)
- **Entity Type:** Semantic category in which a named entity belongs (e.g. ADE, Drug)
- **Relation Type:** Semantic category of the relation (e.g. ADE-Drug)
- **Relation Argument:** Named entity that participates in a relation
- **Relation Instance:** A group of named entities participating in a relation along with their relation type

**Input:** A textual snippet (typically a sentence)

**Output:** A set of entity pairs with their relation label(s)

# Relation Extraction: Taxonomy



# Relation Extraction: Usefulness

- Identification of patterns from raw text

Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

- Creation of structured **Knowledge Bases**, useful for any application
- Automatic extraction of **complex biomedical associations**
- Important for **Question-Answering**
  - One can imagine that answering questions requires factual knowledge
  - Questions and answers can be formed as a relation extraction task ([Lewis et al., 2019](#))

# Relation Extraction: Challenges



- Linguistic phenomena due to the diversity of language
- Grammatical phenomena (e.g. negation)
- Semantic challenges, e.g. implicit statements, common knowledge, etc

# Relation Extraction: Challenges

- Named entities have many **aliases** - especially in the biomedical domain
  - Named Entity Linking / Normalisation is often used to resolve such cases

What is another word for  
**paracetamol?**

Noun

A pharmaceutical drug used to treat pain and fever

acetaminophen    para-acetylaminophenol    tylenol    APAP    Tylenol

N-acetyl-p-aminophenol    p-acetylaminophenol



# Relation Extraction: Challenges

- Named entities have many **aliases** - especially in the biomedical domain
  - Named Entity Linking / Normalisation is often used to resolve such cases
- Presence of hierarchies between entities: **Hyponymy & Hypernymy**
  - **Ontologies** can help identify such associations. E.g. [WordNet](#)

**IS-A** (hypernym): Subsumption between classes

Giraffe IS-A ruminant IS-A ungulate IS-A mammal IS-A animal ...

# Relation Extraction: Challenges

- Named entities have many **aliases** - especially in the biomedical domain
  - Named Entity Linking / Normalisation is often used to resolve such cases
- Presence of hierarchies between entities: **Hyponymy & Hypernymy**
  - **Ontologies** can help identify such associations
- **Negation** alternates meaning and is expressed in various forms

Neither **ryanodine** nor EGTA inhibited down-regulation of **alpha-AR** mRNA by NE.

Further, **AICAR** pretreatment blocked **PAR-1**-induced increase in permeability of mouse-lung microvessels.

# Relation Extraction: Challenges

- Named entities have many **aliases** - especially in the biomedical domain
  - Named Entity Linking / Normalisation is often used to resolve such cases
- Presence of hierarchies between entities: **Hyponymy & Hypernymy**
  - **Ontologies** can help identify such associations
- **Negation** alternates meaning and is expressed in various forms
- Detecting **distant arguments** is difficult
  - E.g. intermediate parenthetical phrases

The case of a 40-year-old patient with **bilateral optic neuropathy**, who underwent unsuccessful cadaver kidney transplantation surgery, was treated with ethambutol and **isoniazid**.

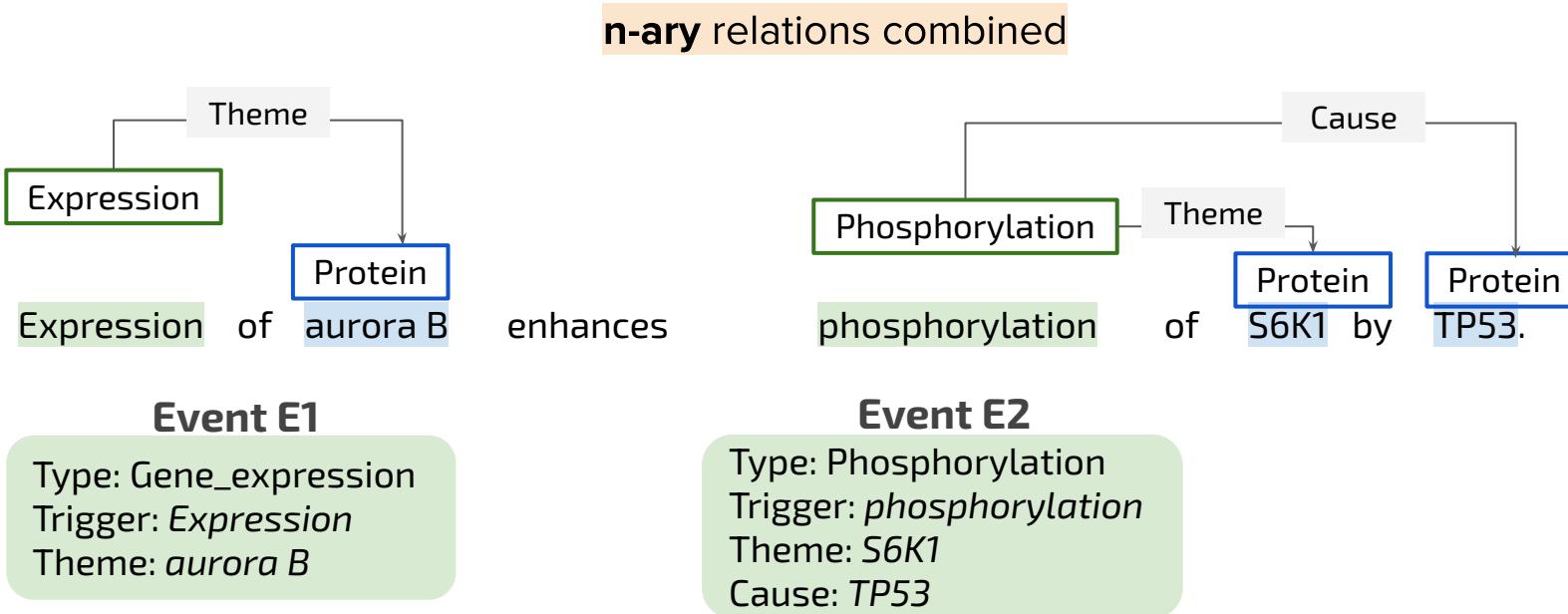
# Relation Extraction: Challenges

- Named entities have many **aliases** - especially in the biomedical domain
  - Named Entity Linking / Normalisation is often used to resolve such cases
- Presence of hierarchies between entities: **Hyponymy & Hypernymy**
  - **Ontologies** can help identify such associations
- **Negation** alternates meaning and is expressed in various forms
- Detecting **distant arguments** is difficult
  - E.g. intermediate parenthetical phrases
- Discourse contains a lot of **implicit relations, temporal relations**

Allergies: **Bactrim (rash)**

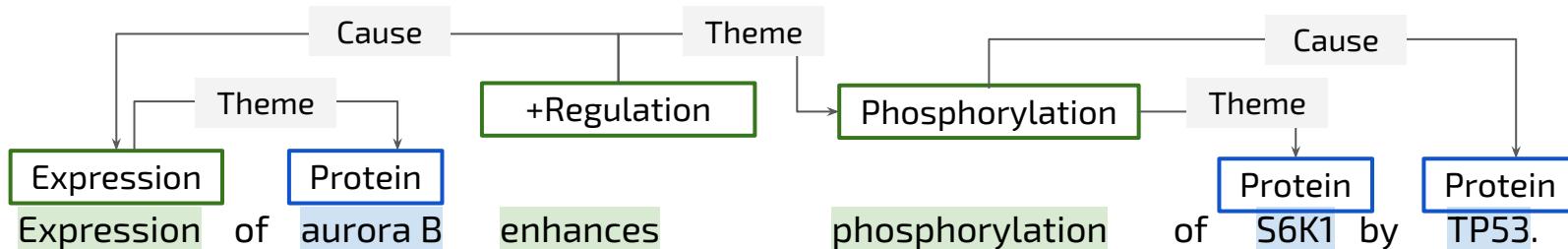
The patient had **two transfusion reactions** to platelets when first admitted . She was premedicated with **antihistamines** .

# Event Extraction



# Event Extraction

n-ary relations combined and **nested**



## Event E1

Type: Gene\_expression  
Trigger: Expression  
Theme: *aurora B*

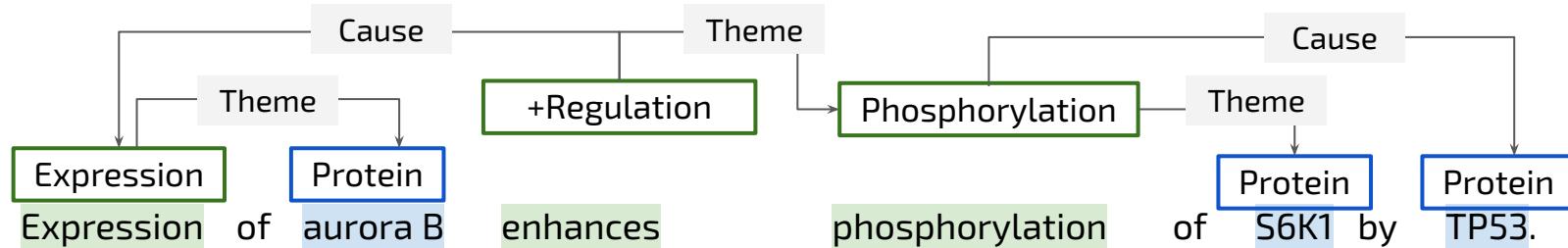
## Event E2

Type: Phosphorylation  
Trigger: *phosphorylation*  
Theme: *S6K1*  
Cause: *TP53*

## Event E3

Type: Positive\_regulation  
Trigger: *enhances*  
Cause: **E1**  
Theme: **E2**

# Event Anatomy



**Event mention:** The words comprising the event in a document

**Event type:** The class label of an event

**Event trigger:** The word(s) that expresses the event mention and indicates its **type**

**Event argument(s):** Entity mention or other event that serves as a **participant** or attribute

**Argument role:** The **relationship** between an argument and the event

## Sample Event Template

**Type:** Phosphorylation

**Trigger type:** <Phosphorylation>

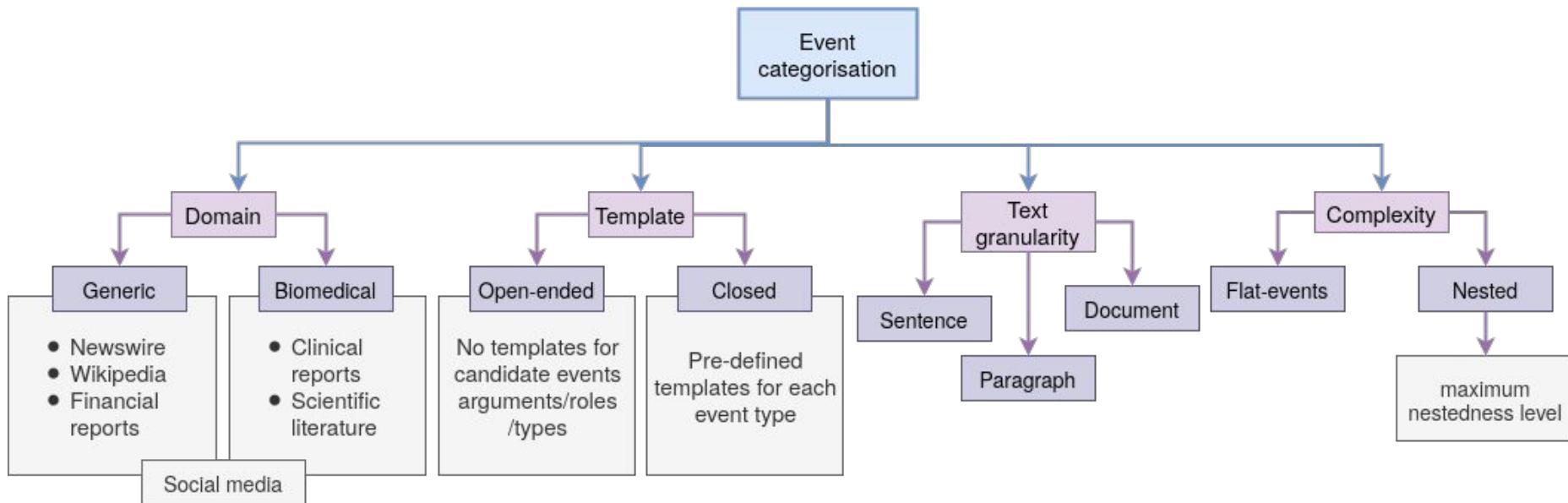
**Argument 1 Type:** <Theme>

**Argument 1 NEs:** <Protein | Expression>

**Argument 2 Type:** <Cause>

**Argument 2 NEs:** <Protein | Drug>

# Event Extraction categories



# Event Extraction: Usefulness

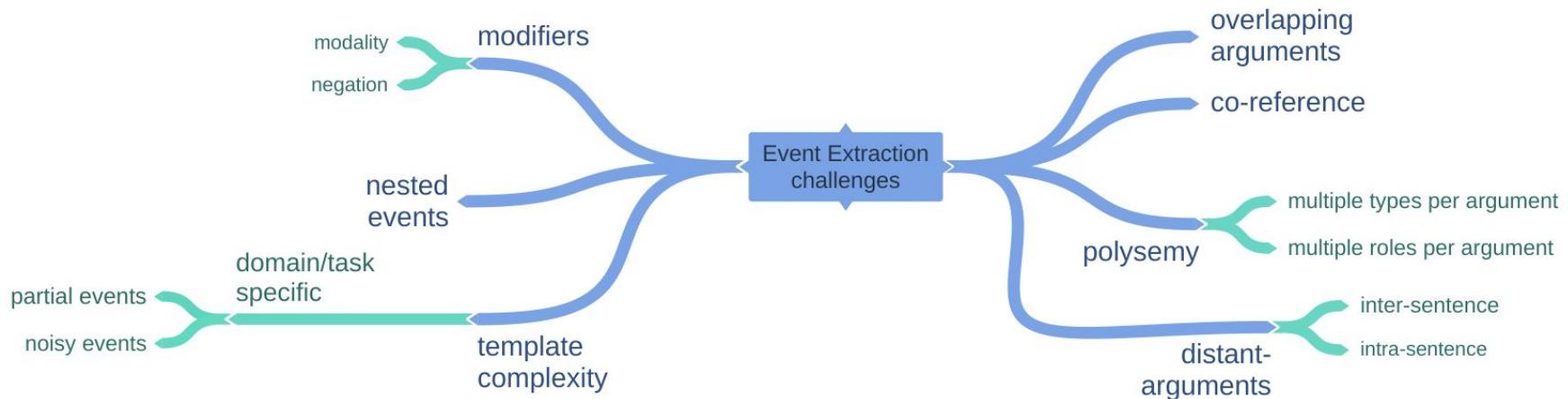
## Why do we need them:

- Events are structured information units:
  - Structured Information Extraction
  - Faceted search
  - Question answering
- Event structure can be mapped to biomolecular interaction structure
  - Constructing (extending) biomolecular networks
  - Semi-automated curation of pathways
- Linking of event mentions across texts can help inferring new associations
  - Knowledge discovery
  - Hypothesis formulation

## Impacted areas:

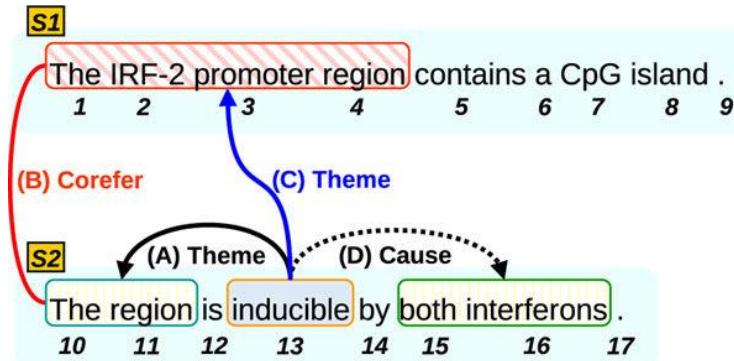
- Pharmacogenetics
  - New drug-protein interactions
- Translational genomics
  - Genome-based tests for standard clinical care
- Cancer research
  - Pathway construction and simulations
- Systems biology
- Functional genomics

# Event Extraction: Challenges



# Event Extraction: Challenges

- **Coreference:** The entity is represented by a **pronoun** or **abstract mention**.  
The actual entity/trigger representation is mentioned in another sub-phrase, sentence

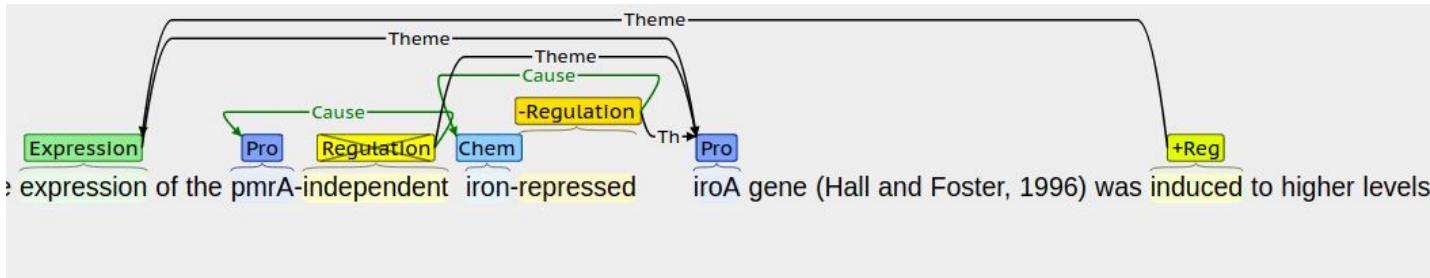


\* Credits to: Katsumasa Yoshikawa et al., 2011

<https://biomedsem.biomedcentral.com.manchester.idm.oclc.org/articles/10.1186/2041-1480-2-S5-S6>

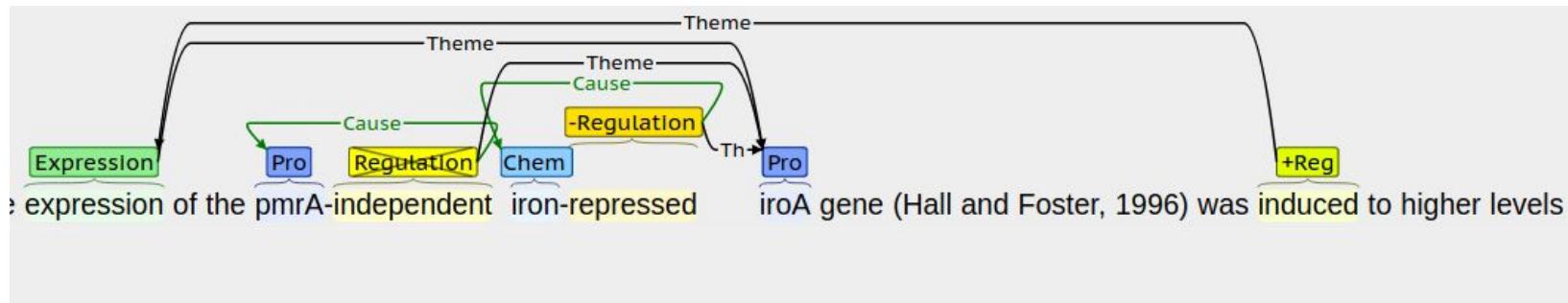
# Event Extraction: Challenges

- **Coreference:** The entity is represented by a **pronoun** or **abstract mention**.  
The actual entity/trigger representation is mentioned in another sub-phrase, sentence.
- An event trigger might **overlap fully or partially** with another event trigger, or event an entity. **Tokenization** might affect the event processing



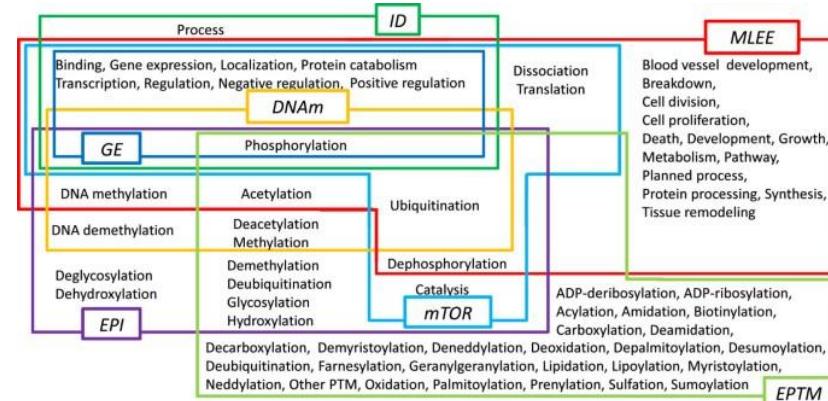
# Event Extraction: Challenges

- **Coreference:** The entity is represented by a **pronoun** or **abstract mention**.  
The actual entity/trigger representation is mentioned in another sub-phrase, sentence.
- An event trigger might **overlap fully or partially** with another event trigger, or event an entity. **Tokenization** might affect the event processing
- An event can (a) be an argument to other events (b) also have other events as arguments. The **depth of the underlying event graph** can expand rapidly!



# Event Extraction: Challenges

- **Coreference:** The entity is represented by a **pronoun** or **abstract mention**.  
The actual entity/trigger representation is mentioned in another sub-phrase, sentence.
- An event trigger might **overlap fully or partially** with another event trigger, or event an entity. **Tokenization** might affect the event processing
- An event can (a) be an argument to other events (b) also have other events as arguments. The **depth of the underlying event graph** can expand rapidly!
- **Generalisation** to other datasets within the domain can be challenging, due to **clashing event structures**.



# Event Extraction: Challenges

- **Coreference:** The entity is represented by a **pronoun** or **abstract mention**.  
The actual entity/trigger representation is mentioned in another sub-phrase, sentence.
- An event trigger might **overlap fully or partially** with another event trigger, or event an entity. **Tokenization** might affect the event processing
- An event can (a) be an argument to other events (b) also have other events as arguments. The **depth of the underlying event graph** can expand rapidly!
- **Generalisation** to other datasets within the domain can be challenging, due to **clashing event structures**.
- Extracting the event mention might be insufficient: context (meta-knowledge) matters!

BRAF affects the binding of MUC1 to PKM2.

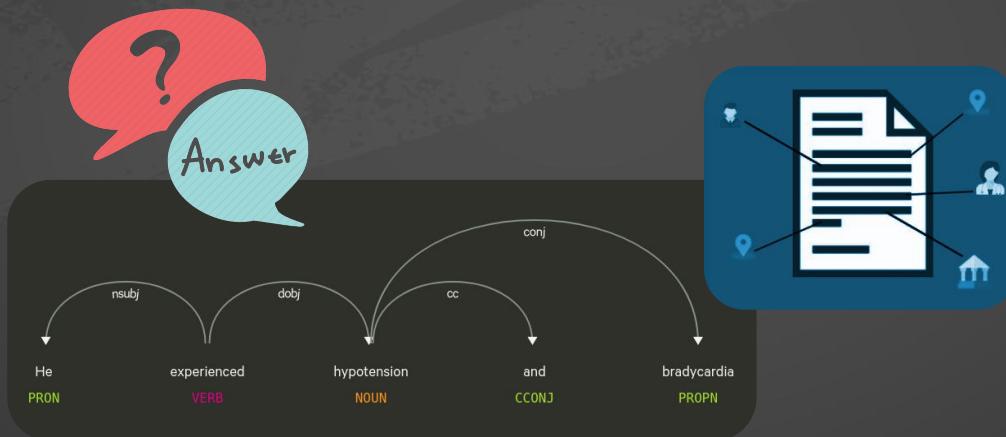
BRAF **does not** affect the binding of MUC1 to PKM2.

# Event Extraction: Challenges

- **Coreference:** The entity is represented by a **pronoun** or **abstract mention**.  
The actual entity/trigger representation is mentioned in another sub-phrase, sentence.
- An event trigger might **overlap fully or partially** with another event trigger, or event an entity. **Tokenization** might affect the event processing
- An event can (a) be an argument to other events (b) also have other events as arguments. The **depth of the underlying event graph** can expand rapidly!
- **Generalisation** to other datasets within the domain can be challenging, due to **clashing event structures**.
- Extracting the event mention might be insufficient: context (meta-knowledge) matters!
- The same event type might have **different argument roles** and/or **entity types** in each sentence

Additional

# BioNLP Tasks



# Metaknowledge Identification

Are all event mentions the same?

BRAF affects the binding of MUC1 to PKM2.

BRAF **does not** affect the binding of MUC1 to PKM2.

We **examined how** BRAF affects the binding of MUC1 to PKM2.

We **prove that** BRAF affects the binding of MUC1 to PKM2.

We **hypothesize that** BRAF affects the binding of MUC1 to PKM2.

BRAF affects the binding of MUC1 to PKM2 [4].

# Metaknowledge Identification

Are all event mentions the same?

BRAF affects the binding of MUC1 to PKM2.

- Contextual information within the same sentence
- Modifies the event
- Additional event classification layers

Based on **our experiments** MUC1 binds **weakly** to PKM2

Knowledge type: Observation      Manner: Low

We wanted to **investigate whether** MUC1 binds to PKM2

Knowledge type: Investigation → Certainty: Low

**According to Wong et al, (2014)** MUC1 **cannot** bind to PKM2

Source: Other      Polarity: Negative

# Entity Normalisation

Is TP53 same as p53? What does this mean?

Also known as:

- Named Entity Linking
- Named Entity Disambiguation

**Task:** assign **unique entity identifiers** to each NE, based on one or more **reference** resources

Reference resources:

- Dictionaries, Thesauri
  - **MeSH terms:** Controlled and hierarchically-organized vocabulary
- Ontologies
  - **ChEBI:** Ontology of molecular entities focused on 'small' chemical compounds.
  - **BioPortal:** Combination of >300 ontologies
- Knowledge bases - Databases
  - **UniProt:** A database of protein sequence and functional information
  - **UMLS:** A structured combination of >1M biomedical concepts from over 100 source vocabularies
    - Includes the Semantic Network of semantically organised concepts
  - **GENBank | DRUGBank**

# Entity Normalisation

Process:

NER

SARS-CoV was able to induce greater IL-1R when compared to Influenza-A.

Candidate generation

UniProtKB: [P14778](#)  
UniProtKB: [P05231](#)  
UniProtKB: [P18510](#)  
...

Interleukin 1 Alpha  
Interleukin 1 Receptor Antagonist  
Interleukin 1 Receptor Type 1  
...

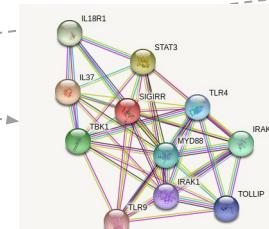
Aliases for IL1RN Gene

Aliases for IL1RN Gene  
Interleukin 1 Receptor Antagonist <sup>2 3 5</sup>  
Interleukin-1 Receptor Antagonist Protein <sup>2 3 4</sup>  
ICIL-1RA <sup>2 3 4</sup>  
**IL-1RN** <sup>2 3 4</sup>  
**IL1F3** <sup>2 3 4</sup>  
**IL1RA** <sup>2 3 4</sup>  
IRAP <sup>2 3 4</sup>  
IL1 Inhibitor <sup>3 4</sup>  
**IL-1ra** <sup>3 4</sup>

Interleukin 1 Receptor Antagonist

- UniProtKB: [P14778](#)
- Entrez Gene: [3557](#)
- Ensembl: [ENSG00000136689](#)

Link with additional context



External IDs for IL1RN Gene  
HGNC: 6000 | Entrez Gene: 3557 | Ensembl: ENSG00000136689 | OMIM: 147679  
Previous GeneCards Identifiers for IL1RN Gene  
GC02P111081, GC02P111796, GC02P113782, GC02P113970, GC02P113591  
GC02P114265, GC02P114906, GC02P113109, GC02P113241, GC02P113370  
Search aliases for IL1RN gene in PubMed and other databases

Summaries for IL1RN Gene

Entrez Gene Summary for IL1RN Gene 

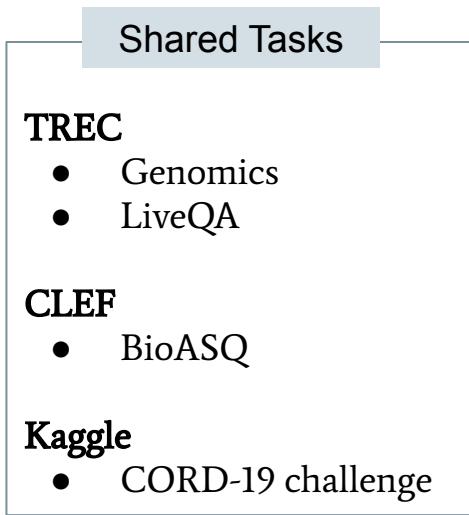
The protein encoded by this gene is a member of the interleukin 1 cytokine family, involved in immune and inflammatory responses. This gene and five other closely related cytokines increased risk of osteoporotic fractures and gastric cancer. Several alternatively spliced transcript variants have been described.

# Question Answering

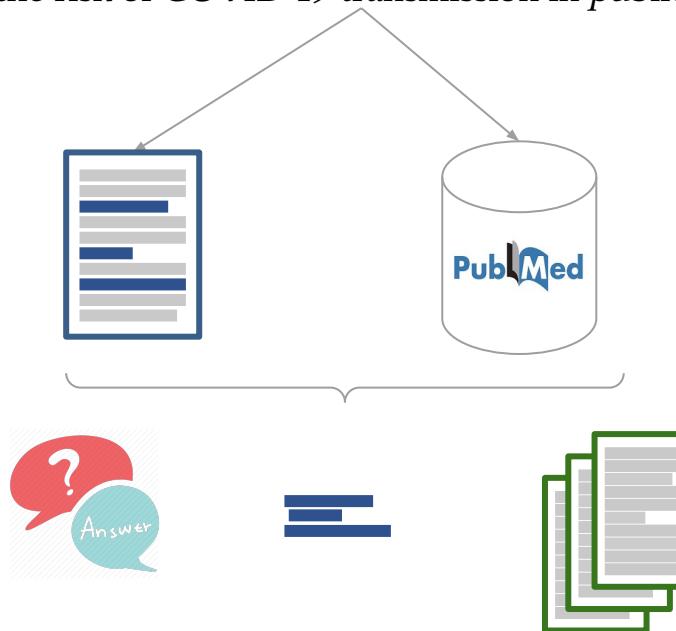
Given a query retrieve relevant information

Also known as:

- Information retrieval
- Document retrieval



*“What is the risk of COVID-19 transmission in public transport”*



# Text Summarisation

- The task that aims to produce a reduced version of a given snippet

## Methods:

- **Extractive:**

*Extract pieces of text and glue them together*

- The most common and easiest task

- **Abstractive:**

*Generate a new summary*

- Sequence-to-sequence models are incorporated
- Grammaticality is important

**Input:** A textual snippet

**Output:** A summary of it

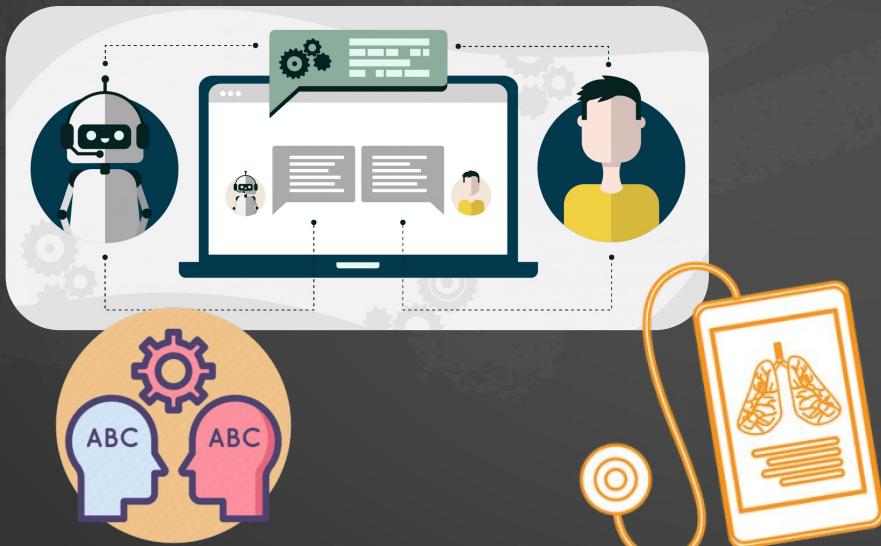
## CONTEXT:

*pyroptosis is an inflammasome-mediated programmed cell death pathway triggered in macrophages by a variety of stimuli, including intracellular bacterial pathogens. C. albicans triggers pyroptosis, a proinflammatory macrophage death. pyroptosis is a caspase-1 dependent pro-inflammatory form of programmed cell death associated with pyroptosis, the pro-inflammatory programmed cell death. our study here identified a novel cell death, pyroptosis in ox-LDL induced human macrophage, which may be implicated in lesion macrophages death and play an important role in lesion instability. caspase-1 induced pyroptosis is an innate immune effector mechanism against intracellular bacteria.*

## SUMMARY (Ground Truth):

*pyroptosis is an inflammasome-mediated programmed cell death pathway.*

# Applications



# Applications

## NaCTeM Tools

A large list can be found here: <http://nactem.ac.uk/software.php> & <http://www.nactem.ac.uk/services.php>

- **Thalia** (<http://www.nactem.ac.uk/Thalia/>)
  - A semantic search engine for Pubmed abstracts
- **LitPathExplorer** ([http://nactem.ac.uk/LitPathExplorer\\_BI/](http://nactem.ac.uk/LitPathExplorer_BI/))
  - A confidence-based visual text analytics tool for exploring literature-enriched pathway models
- **History of Medicine** (<http://www.nactem.ac.uk/MHM/>)
  - A semantic search system over historical medical archives
- **RobotAnalyst** (<http://www.nactem.ac.uk/robotanalyst/>)
  - A tool to minimise the human workload involved in the study identification phase of systematic reviews

# Thalia

Text mining for **H**ighlighting, **A**ggregating and **L**inking **I**nformation in **A**rticles

[http://nactem.ac.uk/Thalia\\_BI/](http://nactem.ac.uk/Thalia_BI/)

**Semantic Search Engine:** Recognition of named entities appearing in biomedical abstracts ([PubMed](#))

- 8 types of concepts:
  - Chemicals
  - Diseases
  - Drugs
  - Genes
  - Metabolites
  - Proteins
  - Species
  - Anatomical entities

<http://nactem.ac.uk/Thalia/>

## Publications

Axel J Soto, Piotr Przybyła and Sophia Ananiadou (2018). [Thalia: Semantic search engine for biomedical abstracts](#). *Bioinformatics*.

Piotr Przybyła, Axel J. Soto, and Sophia Ananiadou (2017). [Identifying Personalised Treatments and Clinical Trials for Precision Medicine using Semantic Search with Thalia](#), In *Proceedings of the Twenty-Fifth Text REtrieval Conference (TREC)*.

# Thalia: General Infrastructure



Search query box

Thalia

GAD

q

v

Advanced Search



Advanced Options

Article meta-data

Year  
Journal  
Author  
Type  
MeSH

Advanced options

Ranking: Newest first

Show ungrounded entities: Off

Click to export search articles as RIS

Export as RIS

Thalia last updated on: 31 Jan 2019

Click to show latest entities indexed

Search

Close

The screenshot shows the Thalia search interface. On the left, there's a sidebar for 'Article meta-data' with fields for Year, Journal, Author, Type, and MeSH. Below it is a 'Search' button. In the center, there's a search bar labeled 'GAD' with a magnifying glass icon and a dropdown arrow. To the right of the search bar is a large orange box labeled 'Advanced Options'. This box contains an 'Advanced options' section with a 'Ranking' dropdown set to 'Newest first', a 'Show ungrounded entities' switch set to 'Off', and a 'Click to export search articles as RIS' button. At the bottom of the dialog, it says 'Thalia last updated on: 31 Jan 2019' and 'Click to show latest entities indexed'. A 'Close' button is at the bottom right of the dialog.

# Thalia: General Infrastructure

The screenshot illustrates the Thalia search interface, which is divided into several sections:

- Search results:** A central area containing a search bar with the term "GAD", a message indicating "8228 abstracts found in 0.05 seconds", and a list of search results. An orange box highlights the search results count.
- Main search results:** A large orange box centered below the search results, containing the text "Main search results".
- Entity facets:** A large orange box on the right side listing entities categorized by type: Chemical, Disease, Drug, and Gene. An orange arrow points from this box to the "Entity facets" label.
- Meta-data facets:** A large orange box on the left side listing facets for Journal, Author, Type, and MeSH. An orange arrow points from this box to the "Meta-data facets" label.

**Search results (highlighted):**

GAD

8228 abstracts found in 0.05 seconds

Comparison of PROMIS Anxiety and Depression, PHQ-8, and GAD-7 to screen for anxiety and depression

...  
Journal of neurosurgery. Spine - 2019  
... Comparison of PROMIS Anxiety and Depression, PHQ-8, and **GAD-7** to screen for anxiety and depression ... at a single institution completed the 7-item Generalized Anxiety Disorder questionnaire (**GAD-7**), 8 ... and **GAD-7** and PHQ-8 scores. Published reference tables were used to determine the presence ...

Fear of disease progression in adult ambulatory patients with brain cancer: prevalence and clinical...

Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer - 2019  
... Scale, **GAD-7**), depression (Patient Health Questionnaire, PHQ-9), Quality of Life (Short Form Health ...

Psychological factors related to resilience and vulnerability among youth with HIV in an integrated...

AIDS care - 2019  
... as part of the standard protocol of care for youth aged 13-24 including information about anxiety (**GAD-7** ...

A Feasibility Open Trial of a Brief Internet-Delivered Written Exposure Therapy for Worry.

Behavioural and cognitive psychotherapy - 2019  
... three adults presenting clinical levels of anxiety (**GAD-7** ≥ 8) and worry (PSWQ ≥ 45) were recruited ... and preliminary efficacy. Primary outcome measures were the Generalized Anxiety Disorder-7 (**GAD-7**) and the Penn ... on the **GAD-7** ( $\eta^2 = 0.36$ ) and the PSWQ ( $\eta^2 = 0.23$ ) with similar findings among study completers ...

Prevalence and associated factors of comorbid anxiety disorders in late-life depression: findings f...

Neuropsychiatric disease and treatment - 2019

Page 1 of 412

**Entity facets:**

- Chemical
- Disease
- Drug
- Gene
  - glutamate decarboxylase (1063, HGNC:4092)
  - insulin (870, HGNC:6081)
  - GAD65 (444, HGNC:4093)
  - human (246, HGNC:7471)
  - tyrosine hydroxylase (172, HGNC:11782)
  - parvalbumin (133, HGNC:9704)
  - GABA(A) receptor (89, HGNC:4089)

**Meta-data facets:**

- Journal
  - Brain Res (278)
  - J Comp Neurol (244)
  - Diabetes Care (139)
  - J Anxiety Disord (134)
  - Diabetes (124)
  - J Affect Disord (124)
  - J Neurochem (106)
- Author
- Type
- MeSH

# Thalia: Querying



Thalia



GAD



8228 abstracts found in 0.14 seconds

▼ Year

▼ Journal

▼ Author

▼ Type

▼ MeSH

Comparison of PROMIS Anxiety and Depression, PHQ-8, and GAD-7 to screen for anxiety and depression ...

Journal of neurosurgery. Spine - 2019

... Comparison of PROMIS Anxiety and Depression, PHQ-8, and GAD-7 to screen for anxiety and depression ... at a single institution completed the 7-item Generalized Anxiety Disorder questionnaire (GAD-7), 8 ... and GAD-7 and PHQ-8 scores. Published reference tables were used to determine the presence ...

Fear of disease progression in adult ambulatory patients with brain cancer: prevalence and clinical...

Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer - 2019

... Scale, GAD-7), depression (Patient Health Questionnaire, PHQ-9), Quality of Life (Short Form Health ...

Psychological factors related to resilience and vulnerability among youth with HIV in an integrated...

AIDS care - 2019

... as part of the standard protocol of care for youth aged 13-24 including information about anxiety (GAD-7 ...

A Feasibility Open Trial of a Brief Internet-Delivered Written Exposure Therapy for Worry.

Behavioural and cognitive psychotherapy - 2019

... -three adults presenting clinical levels of anxiety (GAD-7  $\geq$  8) and worry (PSWQ  $\geq$  45) were recruited ... and preliminary efficacy. Primary outcome measures were the Generalized Anxiety Disorder-7 (GAD-7) and the Penn ... on the GAD-7 ( $r_{np2} = 0.36$ ) and the PSWQ ( $r_{np2} = 0.23$ ) with similar findings among study completers ...



▼ Chemical

▼ Disease

▼ Drug

▼ Gene

▼ Metabolite

▼ Protein

▼ Species

▼ Anatomical

# Thalia: Querying

We can access the PubMed article

PubMed.gov

Search PubMed

Advanced

Save En

> J Affect Disord. 2018 Jan 1;225:593-598. doi: 10.1016/j.jad.2017.08.082. Epub 2017 Aug 31.

## Peripheral proinflammatory cytokines in Chinese patients with generalised anxiety disorder

Zhen Tang <sup>1</sup>, Gang Ye <sup>1</sup>, Xinyun Chen <sup>2</sup>, Mingzhi Pan <sup>1</sup>, Jialin Fu <sup>1</sup>, Tian Fu <sup>1</sup>, Qichun Liu <sup>1</sup>,  
Zhenyong Gao <sup>1</sup>, David S Baldwin <sup>3</sup>, Ruihua Hou <sup>4</sup>

Affiliations + expand

PMID: 28886504 DOI: 10.1016/j.jad.2017.08.082

### Abstract

**Background:** Inflammatory responses and inflammatory cytokines have been implicated in the pathogenesis of affective disorders, particularly major depression. Given the limited evidence relating to the potential role of proinflammatory cytokines in generalised anxiety disorder (GAD), we aimed to examine peripheral proinflammatory cytokines in Chinese patients with GAD.

**Methods:** A case-controlled cross-sectional study design, with recruitment of 48 patients with first episode GAD and 48 matched healthy controls. All participants completed measures of anxiety using well-established questionnaires, and serum levels of pro-inflammatory cytokines were measured using multiplex technology.

### Full text view

## Peripheral proinflammatory cytokines in Chinese patients with generalised anxiety disorder.

Journal of affective disorders - 2018

Go to PubMed

**BACKGROUND:** Inflammatory responses and inflammatory cytokines have been in the pathogenesis of affective disorders, particularly major depression. Given the evidence relating to the potential role of proinflammatory cytokines in generalised disorder (GAD), we aimed to examine peripheral proinflammatory cytokines in Chinese patients with GAD.

**METHODS:** A case-controlled cross-sectional study design, with recruitment of 48 patients with first episode GAD and 48 matched healthy controls. All participants completed measures of anxiety using well-established questionnaires, and serum levels of inflammatory cytokines were measured using multiplex technology.

**RESULTS:** Serum levels of CRP, IL-1 $\alpha$ , IL-6, IL-8, IL-12, IFN- $\gamma$ , and GM-CSF were significantly higher in the GAD group in comparison to the control group ( $p < 0.05$ ). Pearson correlation revealed significant positive correlations between anxiety measures and serum levels of CRP, IL-1 $\alpha$ , IL-6, IL-8, IFN- $\gamma$ , and GM-CSF ( $p < 0.05$ ).

**LIMITATIONS:** The cross-sectional study design does not permit definite conclusions on causal directions between inflammation and GAD. The study was limited to a panel of 8 cytokines and does not exclude the possibility of other important cytokines being involved.

**CONCLUSIONS:** These findings indicate an elevated peripheral proinflammatory response, and provide further support for low grade inflammation in GAD. Further research may identify an 'inflammatory signature' for diagnosis and treatment response, and guide the search for novel pharmacological interventions.

Off Chemicals On Diseases Off Drugs On Genes  
Off Metabolites Off Proteins Off Species On Anatomic

HGNC

Gene data Tools Downloads VGNC Contact us More

Search all Search symbols, keywords or IDs

### Symbol report for CRP

Report HGNC homology predictions

#### HGNC data for CRP

Approved symbol CRP  
Approved name C-reactive protein  
Locus type gene with protein product  
HGNC ID HGNC:2367  
Symbol status Approved  
Previous names "C-reactive protein, pentraxin-related"  
Alias symbols PTX1  
Alias names "pentraxin 1"  
Chromosomal location 1q23.2  
Gene groups Short pentraxins

#### Gene resources for CRP

Ensembl ENSG00000132693 Curated

Ensembl region in detail, Ensembl gene sequence

NCBI Gene 1401 Curated

UCSC uc001ftw.3

Alliance of Genome Resources HGNC:2367

We can get information about an entity

Close

# Thalia: Filtering based on entity types



Thalia



GAD

8228 abstracts found in 0.14 seconds

Comparison of PROMIS Anxiety and Depression, PHQ-8, and GAD-7 to screen for anxiety and depression ...  
Journal of neurosurgery. Spine - 2019  
... Comparison of PROMIS Anxiety and Depression, PHQ-8, and **GAD-7** to screen for anxiety and depression ... at a single institution completed the 7-item Generalized Anxiety Disorder questionnaire (**GAD-7**), 8 ... and **GAD-7** and PHQ-8 scores. Published reference tables were used to determine the presence ...

Fear of disease progression in adult ambulatory patients with brain cancer: prevalence and clinical ...  
GAD

1063 abstracts found in 0.83 seconds

Exogenous  $\gamma$ -aminobutyric acid treatment improves the cold tolerance of zucchini fruit during postharvest ...  
Plant physiology and biochemistry : PPB - 2019  
... GABA transaminase (GABA-T) and glutamate decarboxylase (GAD). GABA-treated fruit contained higher ... t induced the GABA shunt by increasing the activities of the enzymes GABA transaminase (GABA-T) and glutamate decarboxylase (GAD). GABA-treated fruit contained higher levels of fumarate and malate than did non-treated fruit ... by increasing the activities of the enzymes GABA transaminase (GABA-T) and glutamate decarboxylase (GAD). GABA-treated fruit contained higher levels of fumarate and malate than did non-treated fruit, as ...

Energy saving and improvement of metabolism of cultured tobacco cells upon exposure to 2-D-quinotrop ...  
Journal of plant physiology - 2019  
... of glutamate decarboxylase (**GAD**) decreased. Regarding the role of **GAD** in initiation of gamma amino ... ity of glutamate producing enzyme, glutamate dehydrogenase (GDH) increased, whereas the activity of **glutamate decarboxylase** (**GAD**) decreased. Regarding the role of **GAD** in initiation of gamma amino butyric acid (GABA) shunt, ... g enzyme, glutamate dehydrogenase (GDH) increased, whereas the activity of glutamate decarboxylase (**GAD**) decreased. Regarding the role of **GAD** in initiation of gamma amino butyric acid (GABA) shunt, it is ...

Metabolite profiling of Listeria innocua for unravelling the inactivation mechanism of electrolysed ...  
International journal of food microbiology - 2018  
... decarboxylase (**GAD**) system and  $\gamma$ -aminobutyric acid (GABA) shunt for the protection against ... and amino acid metabolism. Elevated levels of  $\alpha$ -ketoglutarate and succinate implicated the enhanced **glutamate decarboxylase** (**GAD**) system and  $\gamma$ -aminobutyric acid (GABA) shunt for the protection against oxidative stress. These ... Elevated levels of  $\alpha$ -ketoglutarate and succinate implicated the enhanced glutamate decarboxylase (**GAD**) system and  $\gamma$ -aminobutyric acid (GABA) shunt for the protection against oxidative stress. These fin ...

▼ Year

▼ Journal

▼ Author

▼ Type

▼ MeSH

▼ Chemical

▼ Disease

▼ Drug

▼ Gene

Gene	Count
glutamate decarboxylase	1063
GAD65	129
insulin	90
human	40
tyrosine hydroxylase	39

HGNC:4093

HGNC:7471

HGNC:74081

# Thalia: Finding co-occurring entities

Search 🔍 ⏷

Article meta-data

Year  
Journal  
Author  
Type  
MeSH

Search

Let's see these and additional features in real time 😎

51 abstracts found in 0.22 seconds

**COMT and GAD1 gene polymorphisms are associated with impaired antisaccade task performance in schizophrenia.**  
European archives of psychiatry and clinical neuroscience - 2018  
... s a reduced proportion of COMT Val/Met heterozygotes and a significantly increased frequency of the **GAD1 rs3749034 C allele** in schizophrenic patients relative to controls. Patients had elevated error rate ... findings may well be derived from specific genetic associations with prefrontal cortex functioning in **schizophrenia**.

**▼ Year**

**▼ Journal**

**▼ Author**

**▼ Type**

**▼ MeSH**

**Chemical**

**Disease**

Entity	Count	UMLS ID
schizophrenia	51	UMLS:C0036341
generalized anxiety disorder	37	UMLS:C0270549
autoimmune diseases	35	UMLS:C0004364
ID	31	UMLS:C0011854
tumor	28	UMLS:C0027651

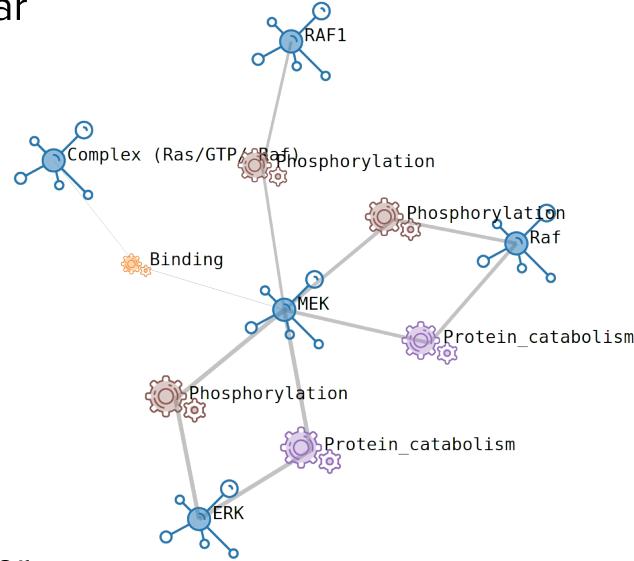
48

# LitPathExplorer

[http://nactem.ac.uk/LitPathExplorer\\_BI/](http://nactem.ac.uk/LitPathExplorer_BI/)

**Bio-events in use:** Enhancing interactions in biomolecular pathway networks

1. **Flexible** search and exploration
  - o multi-view
  - o **interactive** functionalities
2. Identification and visualisation of corroborating/contradicting **textual evidence**
3. **Discovery** and integration of new interactions
4. **Human-in-the-loop** analytical process - adapting to user needs



# LitPathExplorer: Contextualisation

Mapping pathway interactions to textual evidence requires a TM pipeline:

→ NER → EE → NE Normalisation → Metaknowledge → Bibliometrics → Event Mapping

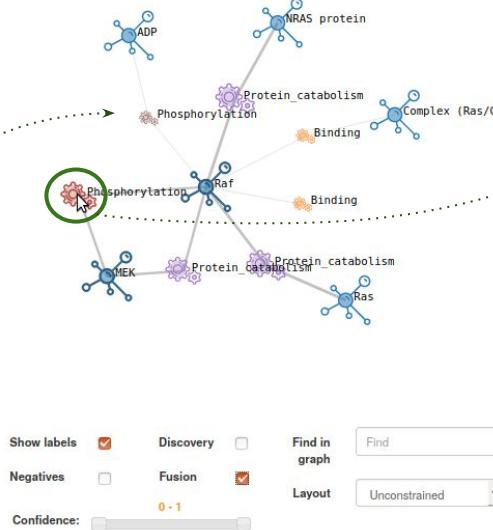
Quantified confidence values  
based on literature

- Metaknowledge
- Bibliometrics
- Citations

LitPathExplorer

Event	Raf
Role	+ -

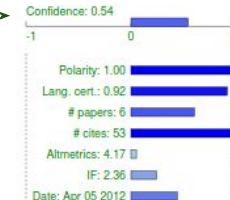
13 entities and events found



## Inspector

Node clicked: Event

Phosphorylation (E6939)



Entity  
normalisation

## Entities related:

GeneOrProtein:  
Raf (PANTHER PATHWAY COMPONENT:P04556)

Role: agent

GeneOrProtein:  
MEK (PANTHER PATHWAY COMPONENT:P04557)

Role: theme

# LitPathExplorer: Textual evidence

Text Analyzer [top](#)

Sentences Word Tree Trigger visualization

Article 1 "Down-regulation of c-Myc following MEK/ERK inhibition halts the expression of malignant phenotype in rhabdomyosarcoma and in non muscle-derived human tumors." - Published in Molecular cancer - 2006 Aug 9  
Cited by 30 articles - Altmetric score: 1 (33%) - Journal Impact Factor: 4.257

Paper confidence: 1.00  
Positive event

As shown in Figure 10A (see additional file 3), **U0126** efficiently inhibited ERK phosphorylation in all the tumor cell lines tested and induced a decrease in c-Myc expression as well as in its phosphorylation throughout the treatment period (6 hrs-4 days).

Sentence confidence: 1.00  
Positive sentence

Article 2 "A novel variant of ER-alpha, ER-alpha36 mediates testosterone-stimulated ERK and Akt activation in endometrial cancer Hec1A cells." - Published in Reproductive biology and endocrinology : RB&E - 2009 Sep 24  
Cited by 16 articles - Altmetric score: 0 (0%) - Journal Impact Factor: 2.226

Paper confidence: 0.50  
Positive event

Testosterone induced ERK and Akt phosphorylation, which could be abrogated by ER-alpha 36 shRNA knockdown or the kinase inhibitors, **U0126** and LY294002, and the aromatase inhibitor letrozole.

Sentence confidence: 0.50  
Positive sentence

Article 3 "Elevated insulin-like growth factor 1 receptor signaling induces antiestrogen resistance through the MAPK/ERK and PI3K/Akt signaling routes." - Published in Breast cancer research : BCR - 2011 May 19  
Cited by 37 articles - Altmetric score: 0 (0%) - Journal Impact Factor: 5.49

Paper confidence: 1.00

## Textual confidence:

- Polarity + uncertainty
- Multi-level
  - Sentence
  - Paper

## Interactive-adaptive:

- User feedback
- **propagated** (upwards aggregation)
- Used as data for **online learning** (2 layer NN)

# LitPathExplorer: Textual evidence

Alternative views:



# LitPathExplorer: Textual evidence

Alternative views:

Text Analyzer [top](#)

Sentences Word Tree Trigger visualization

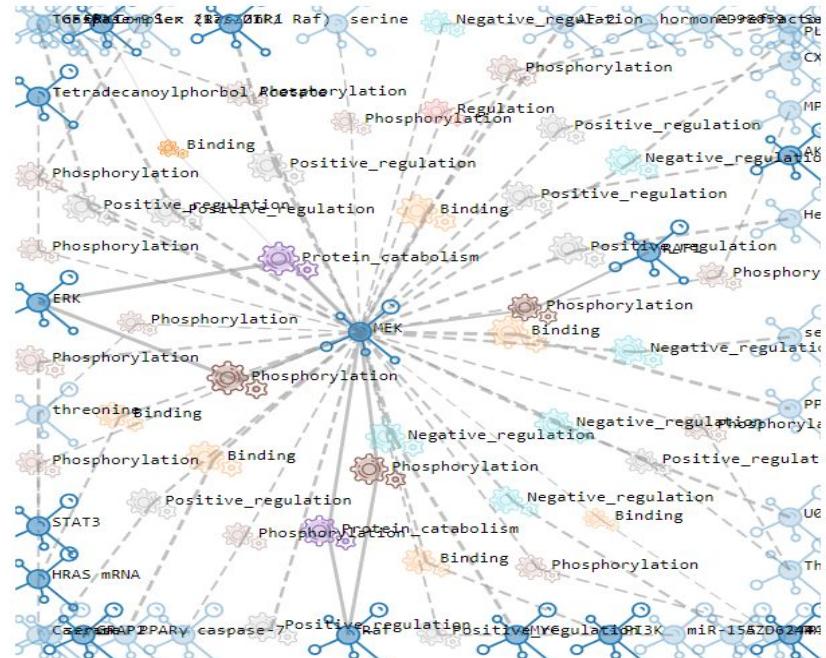
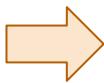
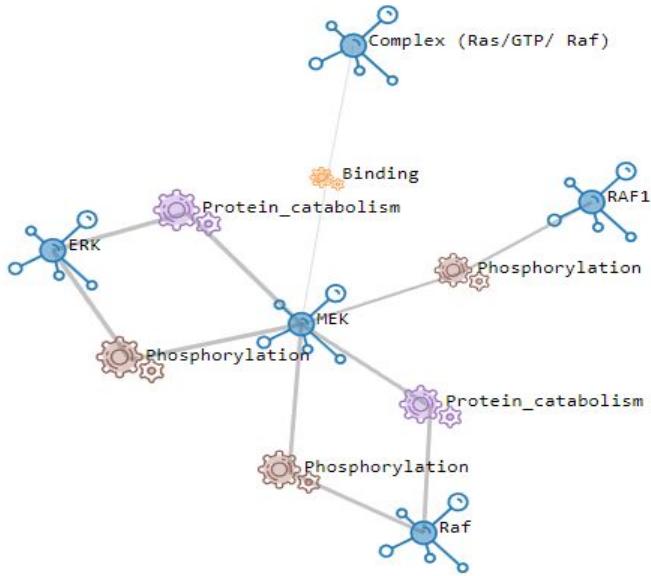
U0126

Starts  Ends

U0126

efficiently inhibited ERK phosphorylation in all the tumor cell lines tested and induced a decrease in c-Myc expression .  
(10 µM) largely inhibited ERK phosphorylation without interfering with activation of either IGF-1R or Akt .  
inhibited phosphorylation of ERK as expected .  
and LY294002 , and the aromatase inhibitor letrozole .

# LitPathExplorer: Discovery



Show labels  Discovery  Find in graph  
Negatives  Fusion  Layout Unconstrained  
Confidence:

filter

Show labels  Discovery  Find in graph  
Negatives  Fusion  Layout Unconstrained  
Confidence:



# History of Medicine

Discover and extract information automatically from medical historical archives

- Archives:
  - The British Medical Journal (BMJ) (1840-2013)
  - The London Medical Officer of Health reports (MOH) (1848-1972)
- **Bibliographic information**
  - Author name, publication date, index (BMJ or MOH), journal
- **Semantic Information**
  - **Entities:** medical conditions, signs or symptoms, environmental factors
  - **Events:** Relationships between entities
- **Suggestions of related query terms**
- **Historical tracking of query terms**

<http://nactem.ac.uk/MHM/>

<http://www.nactem.ac.uk/hom/>

## Publications

Thompson, P., Batista-Navarro, R. T. B., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmermann, C., Worboys, M. and Ananiadou, S. (2016). [Text Mining the History of Medicine](#). *PLOS One*.

Thompson, P., Carter, J., McNaught, J. and Ananiadou, S. (2015). [Semantically Enhanced Search System for Historical Medical Archives](#). In *Proceedings of Digital Heritage*.

Thompson, P., McNaught, J. and Ananiadou, S. (2015). [Customised OCR Correction for Historical Medical Text](#). In *Proceedings of Digital Heritage*.

Miwa, M., Thompson, P., Korkontzelos, I. and Ananiadou, S. (2014). [Comparable Study of Event Extraction in Newswire and Biomedical Domains](#). In *Proceedings of Coling*.

# History of Medicine: Documents

Search My Documents

Register User Sign In 

## Search Query

You are currently viewing the entire set of documents. To find specific documents please refine your search.

 Start refining search

## Search Results

Showing page 1 of 385409 results



### Cardiomyopathy associated with Wegener's granulomatosis



Cardiomyopathy associated with Wegener's granulomatosis Andrew To Janak De Zoysa Jonathan P Christiansen Jonathan.Christiansen@WaitemataDHB.govt.nz KEYWORDS: A 35-year-old man with no history of cardiovascular disease presented with severe biventricular heart failure, after a 6-week history ...

*bmj\_3106061*



### Recurrent wheezing ... is it only asthma?

Recurrent wheezing ... is it only asthma? Luis Vaz Rodrigues Cristina Lopes a Castel-Branco 1 Centro Hospitalar de Coimbra, Pulmunology, Quinta dos Vales, S Martinho do Bispo, Coimbra 3041-801, Portugal 2 Hospital de São João, EPE, Immuno-allergology Division, Alameda Prof Hernâni Monteiro, Porto...

*bmj\_3027324*



### Hepatitis, rhabdomyolysis and multi-organ failure resulting from statin use

Hepatitis, rhabdomyolysis and multi-organ failure resulting from statin use Muthuram Rajaram St Helens and Knowsley Hospitals NHS Trust, Medicine-Gastro, Whiston Hospital, Warrington Prescot, L35 5DR, UK drmuthuram@yahoo.co.in Objective: Abstract To report a case of hepatitis, rhabdomyolysis...

LWW 2008;100

# History of Medicine: Documents

Search My Documents

Register User Sign In 

## Cardiomyopathy associated with Wegener's granulomatosis

Showing page 1 of 385409 results

### Record

#### Reference

3106061

#### Title

Cardiomyopathy associated with Wegener's granulomatosis

#### Year

2009

#### Page Info

bcr2006098038

#### Index

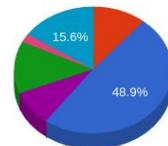
bmj

### Content

Due to licensing conditions we are unable to provide the full text of this document, however the article is available externally in the following repositories.

- Europe PubMed Central
- PubMed Central

### Entity Types



### Entities

- Anatomical (5)
- Condition (22)
- Environmental (4)
- Sign or Symptom (6)
- Subject (1)
- Therapeutic or Investigational (7)

### Events

- Affect (2)

iatosis

To Janak De Zoysa Jonathan P Christiansen  
year-old man with no history of cardiovascular  
risk history ...

bmj\_3106061

a Lopes a Castel-Branco 1 Centro Hospitalar de  
mbra 3041-801, Portugal 2 Hospital de São João,  
Porto...

bmj\_3027324

lting from statin use

use Muthuram Rajaram St Helens and Knowsley  
Prescot, L35 5DR, UK drmuthuram@yahoo.co.in

bmj\_3027324

# History of Medicine: Search Entities

Search My Documents

Register User Sign In ?

## Search Query

You are currently viewing the entire set of documents. To find specific documents please refine your search.

What type of entity do you want to find?

- Anatomical
- Biological
- Condition
- Environmental
- Sign or Symptom
- Subject
- Therapeutic or Investigational



## Search Results ?

Showing page 1 of 385409 results



### Cardiomyopathy associated

Cardiomyopathy associated with  
Jonathan.Christiansen@Waitemata

Please enter a sign or symptom to search for

- symptom (83713)
- pain (61132)
- sign (53829)
- attack (48373)
- diarrhoea (21620)
- vomiting (21575)
- fever (19677)
- sickness (17136)
- ill (15852)
- weakness (15253)
- cough (14416)
- headache (14350)
- irritation (14045)
- enlarged (13976)
- paralysis (13654)

Back

?

Refine search

# History of Medicine: Search Entities

Search My Documents

Register User Sign In 

## Search Query

You are currently viewing the entire set of documents. To find specific documents please refine your search.

 Start refining search

## Search Results

Showing page 1 of 385409 results



### Cardiomyopathy associated with Wegener's granulomatosis

Cardiomyopathy associated with Wegener's granulomatosis Andrew To Janak De Zoysa Jonathan P Christiansen Jonathan.Christiansen@WaitemataDHB.govt.nz KEYWORDS: A 35-year-old man with no history of cardiovascular disease presented with severe biventricular heart failure, after a 6-week history ...

*bmj\_3106061*



### Recurrent wheezing ... is it only asthma?

Recurrent wheezing ... is it only asthma? Luis Vaz Rodrigues Cristina Lopes a Castel-Branco 1 Centro Hospitalar de

Search My Documents

Register User Sign In 

## Search Query

Sign or Sympton   
paralysis

New Search

Refine Search

## Search Results

Showing page 1 of 13654 results



### The benefits of steroids versus steroids plus antivirals for treatment of Bell's palsy: a meta-analysis

The benefits of steroids versus steroids plus antivirals for treatment of Bell's palsy: a meta-analysis Eudocia C Quant - neuro-oncology fellow Shafali S Jeste - neurologist Rajeev H Muni - ophthalmologist Alison V Cape - maternal fetal medicine fellow Manveen K Bhussar - clinical research ass...

*bmj\_2739281*



### Ecstasy-induced thyrotoxic periodic paralysis

Ecstasy-induced thyrotoxic periodic paralysis Lisa Forrest Julia Platts 1 Prince Phillip Hospital, Diabetes and Endocrinology, Dafan, Llanelli, Wales, SA14 8QF, UK 2 Diabetes and Endocrinology, Diabetes Centre, Llandough University Hospital, Cardiff, CF64 2XX, UK Lisa Forrest, lisaforrest@doctor...

# History of Medicine: Search Events

Search My Documents

Register User Sign In

Search Query

Sign or Sympton

New Search Refine Search

Please enter causality event arguments

CAUSE   
causes   
paralysis   
in   
SUBJ

What type of event do you want to find?

Affect  
Causality

13654 results

a meta-

ia C Quant -  
ernal fetal

bmj\_2739281

nd  
ndough

bmi\_3028000

# History of Medicine: Search Events

Search My Documents

Register User Sign In 

## Search Query

Sign or Sympton



paralysis



New Search

Refine Search

## Search Results

Showing page 1 of 13654 results



The benefits of steroids versus steroids plus antivirals for treatment of Bell's palsy: a meta-analy...

The benefits of steroids versus steroids plus antivirals for treatment of Bell's palsy: a meta-analysis Eudocia C Quant - neuro-oncology fellow Shafali S Jeste - neurologist Rajeev H Muni - ophthalmologist Alison V Cape - maternal fetal medicine fellow Manveen K Bhussar - clinical research ass...

bmj\_2739281

Search My Documents

Register User Sign In 

## Search Query

Sign or Sympton



paralysis



Event



Cause of paralysis

New Search

Refine Search

## Search Results

Showing page 1 of 1400 results



The benefits of steroids versus steroids plus antivirals for treatment of Bell's palsy: a meta-analy...

The benefits of steroids versus steroids plus antivirals for treatment of Bell's palsy: a meta-analysis Eudocia C Quant - neuro-oncology fellow Shafali S Jeste - neurologist Rajeev H Muni - ophthalmologist Alison V Cape - maternal fetal medicine fellow Manveen K Bhussar - clinical research ass...

bmj\_2739281



Recurring paralysis

Recurring paralysis Hung-Wei Lin Tom Chau Chin-Sheng Lin Shih-Hua Lin Tri-Service General Hospital, Department of Medicine, Number 325, Section 2, Cheng-Kung Road, Taipei, Neihu 114, Taiwan Shih-Hua Lin, shihhualin@yahoo.com Abstract A 22-year-old Chinese man presented with sudden onset of g...

bmj\_3027774

# History of Medicine: Tracking

Search My Documents

Register User Sign In ?

Search Query

Term  x

New Search Refine Search

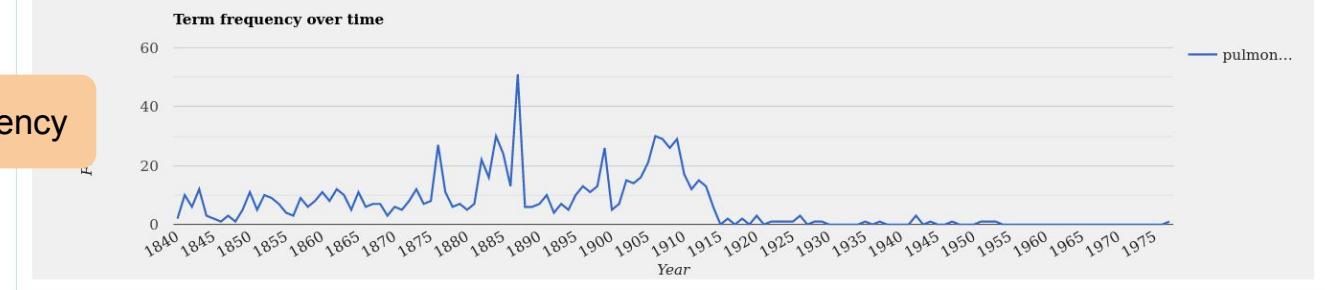
Related Terms

pulmonary cedema phthisis pulmonalis  
pulmonary tuberculosis pulmonary syphilis  
pulmonary affection  
tuberculous pulmonary phthisis  
infectious pulmonary Pulmonary tuberculous pulmonary  
tuberculous phthisis  
pulmonary form  
tubercular consumption Infectious pulmonary tuberculosis  
tuberculous consumption pulmonary heart disease  
pulmonary tubercular disease consumption  
pulmonary phthisis consumption  
tuberculous pulmonary disease chronic pulmonary tuberculosis

## Search Results ?

Showing page 1 of 870 results

### Term Frequencies



Frequency



### Rest in Pulmonary Consumption

REST IN PULMONARY CONSUMPTION. SIR,-In a paragraph which appears in the BRITISH1 MEDICAL JOURNAL of Oct. 4th, it is mentioned that Dr. Berkart has instituted a novel plan of treating pulmonary consumption by the application of mechanical compression made in such a manner by strapping and...

bmj\_2294575



### GRADUATED LABOUR IN PULMONARY CONSUMPTION

GRADUATED LABOUR IN PULMONARY CONSUMPTION. Sir,-Dr. M. S. Paterson's paper on graduated labour in pulmonary consumption will be read with great interest by every one engaged in treating consumption. For the last two years, since a short visit to Sir A. E. Wright's laboratory, endeavour ha...

bmj\_2435881



### Thomas Ballman

THOMAS BARMAN, M.D., OF LIVERPOOL WE (Liverpool A4bo'n) regret to have to announce the death of Thomas Ballman, M.D., late consulting-

Documents containing the term

# History of Medicine: Tracking

Search My Documents

Register User Sign In

Search Query

Term **pulmonary consumption**

Term **pulmonary tuberculosis**

New Search Refine Search

## Related Terms

**pulmonary y tuberculosis**

tuberculous pulmonary disease pulmonary cedema

**bilateral pulmonary tuberculosis**

**miliary pulmonary tuberculosis**

phthisis pulmonalis

**nonpulmonary tuberculosis**

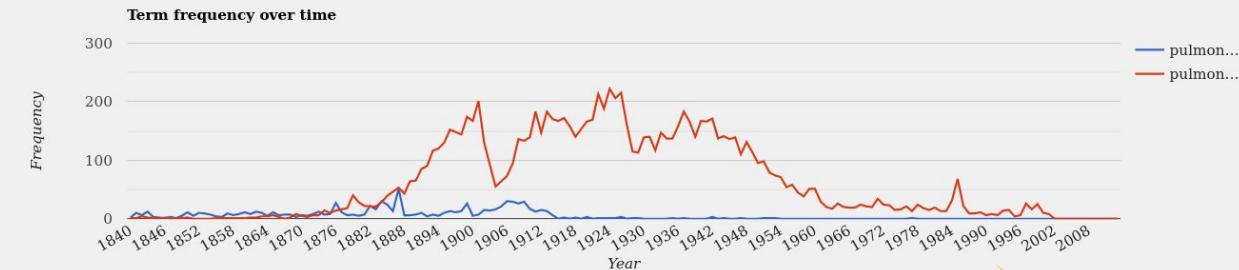
**extrapulmonary tuberculosis**

**inactive pulmonary tuberculosis**

## Search Results

Showing page 1 of 11460 results

### Term Frequencies



### The Naval Manœuvres: Danger of Dissemination of Consumption

THE NAVAL MANŒUVRES: DANGER OF DIS SEMINATION OF CONSUMPTION. Sir,-The time for mobilisation of the fleet for the naval manœuvres being at hand, it may not be amiss to draw attention to the necessity for careful examination of the men, with a view to prevent any with pulmonary tuberculosis be...

bmj\_2403738



### PREVENTION OF PULMONARY TUBERCULOSIS

PREVENTION OF PULMONARY TUBERCULOSIS. A CONFERENCE of representatives of twenty-eight out of three twenty-nine city and metropolitan borough councils was held on June 6th at the Paddington Town Hall to consider the question of taking measures to limit the spread of pulmonary tuberculosis. The chal...

bmj\_2357676



### Scotland

[FROM OUR SPECIAL CORRESPONDENTS.] GLASGOW OBSTETRICAL AND GYNAECOLOGICAL SOCIETY. ON Friday, June 28th, Professor Whitridge Williams, of Baltimore, delivered an address as Honorary President of the society. Taking as his subject the pernicious vomiting of pregnancy, he grouped the cases into tw...

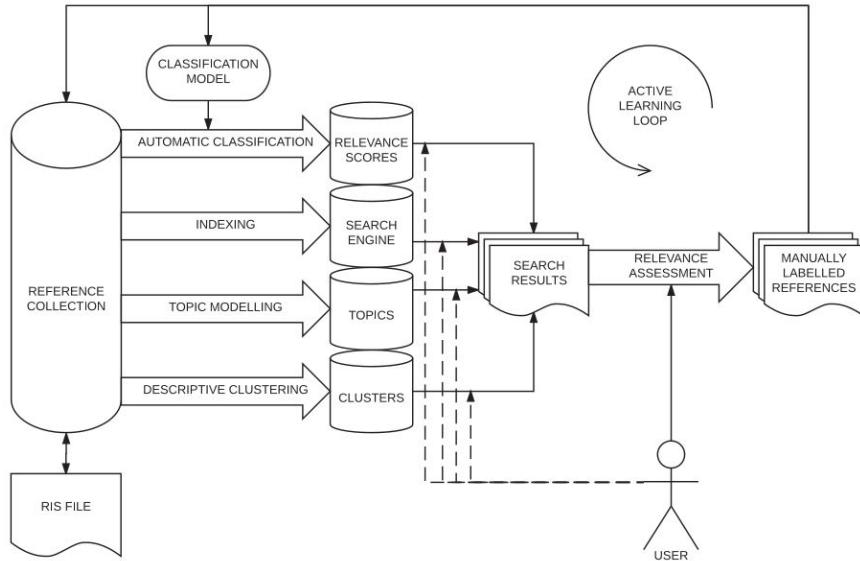
bmj\_2333945

The second term is more frequently used after 1876

# RobotAnalyst

## Systematic review helper:

- Support for multiple collections
  - Easily updatable - tuned
- Multi-faceted & multi-level search
  - Quick search and integration of new publications
  - Search within a collection
  - Search within a cluster or topic
  - Faceted search
- Facilitated screening
  - Active learning
- Faceted search
- Topic analysis
- Automated clustering



# RobotAnalyst

Upload a set of initial references (or leave empty)

Specify main research question

Create Collection

✓ Collection successfully created.  
✓ Documents added to the database.

Preprocessing documents. This may take a few minutes... 

**Collection Name**  
Cancer collection

**Files**  
 cancer.ris  
Supported file extensions: .txt & .ris  
**cancer.ris**

**Default Label**  
As marked

**Research Question**  
Cancer diagnosis and therapeutics

**Disclaimer**

# RobotAnalyst: Populate

Q Search Collection Import From ▾ Exit Collection Preferences Chrysoula Zerva ▾

Collection: Cancer Research

Screening - Round 0

Included: 0  
Excluded: 0  
Undecided: 1

Screening - Overall

Included: 0  
Excluded: 0  
Undecided: 1

Clusters Refine Search Clear

Choose Algorithm SC ▾  
Number of Clusters 5 ▾  
Get Clusters

There are no clusters to show.

Overall Summary

100%

Please enter the term you wish to search for and choose how many documents you want to retrieve

Lung cancer women

examples: "text mining", "smoking or tobacco" or "tobacco and health"

1 500

Number of Documents to retrieve: 402

Back Search

Excluded

Descending Go

NaCTeM The National Centre for Text Mining

Uncluded I E U

# RobotAnalyst: Populate

External Search: PubMed



New Search ▾ Add Documents to Collection 8 documents selected

Showing page 1 of 27 (396 results)

Results	
Surgical treatment of peri-acetabular metastatic disease: Retrospective, multicentre study of 91 THA cases. Show More ▾	-
Overcoming resistance to anti-PD-1 immunotherapy in lymphoepithelioma-like carcinoma: A case report and review of the literature. Show More ▾	-
Sex Differences in Mortality Rates and Underlying Conditions for COVID-19 Deaths in England and Wales. Show More ▾	-
microRNA-877 contributes to decreased non-small cell lung cancer cell growth via the PI3K/AKT pathway by targeting tartrate resistant acid phosphatase 5 activity. Show More ▾	+
Medicaid Expansion and Mortality Among Patients With Breast, Lung, and Colorectal Cancer. Show More ▾	-
Mediastinal Metastasis of Breast Cancer Mimicking a Primary Mediastinal Tumor. Show More ▾	-
Cancer incidence in the vicinity of a waste incineration plant in the Nice area between 2005 and 2014. Show More ▾	-
Fears and Perception of the Impact of COVID-19 on Patients With Lung Cancer: A Mono-Institutional Survey. Show More ▾	+
[Transsternal occlusion of main bronchi fistulae after pneumonectomy]. Show More ▾	+
[Immediate results of surgical treatment of advanced age patients with non-small cell lung cancer]. Show More ▾	+
Education, income and risk of cancer: results from a Norwegian registry-based study. Show More ▾	-
SOX9 Is Essential for Triple-Negative Breast Cancer Cell Survival and Metastasis. Show More ▾	+
[Cancer Therapy with Cell Sheets]. Show More ▾	-
Measuring progress against cancer in the Azores, Portugal: Incidence, survival, and mortality trends and projections to 2025. Show More ▾	+
Chronic circadian disruption modulates breast cancer stemness and immune microenvironment to drive metastasis in mice. Show More ▾	+

« 1 2 3 4 5 6 7 8 9 10 »

# RobotAnalyst: Panel Overview

Collection: Cancer collection

Screening - Round 0

Included: 0  
Excluded: 0  
Undecided: 400

Screening - Overall

Included: 0  
Excluded: 0  
Undecided: 400

Clusters Refine Search

You are currently viewing the entire set of documents. To find specific documents please refine your search.

[Start refining search](#)

Overall Summary



100%

What are you interested in searching for?

- Text search
- Term
- Document ID
- Publication year
- Author
- Type of publication
- Journal
- Notes
- Keywords
- Time of screening decision
- Retrieval method
- Topic

Similar Articles

Prostate cancer immunotherapy, particularly in combination with androgen deprivation or radiation treatment. Customized pharmacogenomic approaches to overcome immunotherapy cancer resistance.

Document ID: 5

NaCTeM  
The National Centre for Text Mining

Excluded

Ascending Go

Undecided I E U

Undecided I E U

Undecided I E U

Undecided I E U

# RobotAnalyst: Reviewing

Collection: Cancer collection



**Screening - Round 0**

Included: 2  
Excluded: 1  
Undecided: 399

**Screening - Overall**

Included: 1  
Excluded: 0  
Undecided: 399

**Clusters** **Refine Search**

You are currently viewing the entire set of documents. To find specific documents please refine your search.

**Start refining search**

**Overall Summary**

99.8%

**Screen: All** **Decided** **Undecided** **Included** **Excluded**

Showing page 1 of 27 (400 results)  
1 included | 1 excluded | 0 predicted includes | 0 predicted excludes

Year Descending Go

**Studies on phytochemical, antioxidant, anti-inflammatory, hypoglycaemic and antiproliferative activities of Echinacea purpurea and Echinacea angustifolia extracts.**  
Document ID: 1  
Year:  
**Most Relevant Topic:** activities, extracts, antioxidant, extract, context  
**Time of Screening Decision:** Just Now  
**Retrieval Method:** Faceted Search: no facets applied, Sort: {Year, desc}  
**Similar Articles**

**Minimum lesion detectability as a measure of PET system performance.**  
Document ID: 2  
Year:  
**Most Relevant Topic:** background, uptake, imaging, small, lesion  
**Time of Screening Decision:** Just Now  
**Retrieval Method:** Faceted Search: no facets applied, Sort: {Year, desc}  
**Similar Articles**

**Detection of aberrant protein phosphorylation in cancer using direct gold-protein affinity interactions.**  
Document ID: 3  
Year:  
**Most Relevant Topic:** protein, increased, cell, effects, expression  
**Time of Screening Decision:** Just Now  
**Retrieval Method:** Faceted Search: no facets applied, Sort: {Year, desc}  
**Similar Articles**

**Design, synthesis and molecular docking of novel diarylcyclohexenone and diaryllindazole derivatives as tubulin polymerization inhibitors.**  
Document ID: 4  
Year:  
**Most Relevant Topic:** compounds, inhibitors, derivatives, hca, ranging  
**Time of Screening Decision:** 2020-12-27 18:36:27  
**Retrieval Method:** Initial upload  
**Similar Articles**

**Prostate cancer immunotherapy, particularly in combination with androgen deprivation or radiation treatment. Customized pharmacogenomic approaches to overcome immunotherapy cancer resistance.**  
Document ID: 5

The diagram shows a vertical stack of five colored boxes, each containing a 3x3 grid of letters I, E, and U, representing the screening status of individual documents. The boxes are color-coded: green, pink, light green, orange, and white. The green box contains the first three rows of the grid. The pink box contains the next two rows. The light green box contains the next row. The orange box contains the next row. The white box contains the final row. This visual representation likely corresponds to the 'Included', 'Excluded', 'Undecided', 'Excluded', and 'Included' categories shown in the interface above.

# RobotAnalyst: Screening

Collection: Cancer collection



Screening - Round 0

Included: 12  
Excluded: 3  
Undecided: 394

Update Predictions

Screen: All Decided Undecided Included Excluded

Showing page 1 of 28 (409 results)  
12 included | 3 excluded | 0 predicted includes | 0 predicted excludes

Year Descending Go

Surgical treatment of peri-acetabular metastatic disease: Retrospective, multicentre study of 91 THA cases.  
Document ID: 2  
Year: 2020  
Most Relevant Topic: mobility, surgical, cup, dual, tha  
Time of Screening Decision: Just Now  
Retrieval Method: Faceted Search: no facets applied, Sort: {Year, desc}  
Similar Articles

Overcoming resistance to anti-PD-1 Immunotherapy in lymphoepithelioma-like carcinoma: A case report and review of the literature.  
Document ID: 3  
Year: 2020  
Most Relevant Topic: ielc, markers, rare, sfts, carcinoma  
Time of Screening Decision: Just Now  
Retrieval Method: Faceted Search: no facets applied, Sort: {Year, desc}  
Similar Articles

Sex Differences in Mortality Rates and Underlying Conditions for COVID-19 Deaths in England and Wales.  
Document ID: 4  
Year: 2020  
Most Relevant Topic: covid, deaths, rates, mortality, women  
Time of Screening Decision: Just Now  
Retrieval Method: Faceted Search: no facets applied, Sort: {Year, desc}  
Similar Articles

Medicaid Expansion and Mortality Among Patients With Breast, Lung, and Colorectal Cancer.  
Document ID: 5  
Year: 2020  
Most Relevant Topic: expansion, medicaid, mortality, states, lung  
Time of Screening Decision: Just Now  
Retrieval Method: Faceted Search: no facets applied, Sort: {Year, desc}  
Similar Articles

Mediastinal Metastasis of Breast Cancer Mimicking a Primary Mediastinal Tumor.  
Document ID: 6  
Year: 2020  
Most Relevant Topic: cancer, case, tumor, patient, report  
Time of Screening Decision: 2020-12-29 09:52:29  
Retrieval Method: Faceted Search: no facets applied, Sort: {Year, desc}  
Similar Articles

Clusters Refine Search

Choose Algorithm SC Number of Clusters 5 Get Clusters

There are no clusters to show.

Overall Summary

Inspect distribution

# RobotAnalyst: Automated screening

Collection: Cancer Research



Screening - Round 1

Included: 0   Excluded: 0   Undecided: 394

Update Predictions

Screening - Overall

Included: 12   Excluded: 3   Undecided: 394

Apply Predictions

Retrieval Method: Faceted Search, no facets applied, Sort: {Year, desc}

Year: 2020  
Most Relevant Topic: mobility, surgical, cup, dual, tha  
Time of Screening Decision: 2020-12-29 09:53:32  
Retrieval Method: Faceted Search: no facets applied, Sort: {Year, desc}  
Similar Articles

[Overcoming resistance to anti-PD-1 immunotherapy in lymphoepithelioma-like carcinoma: A case report and review of the literature.](#)  
Document ID: 3  
Year: 2020  
Most Relevant Topic: ielc, markers, rare, sfts, carcinoma  
Time of Screening Decision: 2020-12-29 09:53:33  
Retrieval Method: Faceted Search: no facets applied, Sort: {Year, desc}  
Similar Articles

[Sex Differences in Mortality Rates and Underlying Conditions for COVID-19 Deaths in England and Wales.](#)  
Document ID: 4  
Year: 2020  
Most Relevant Topic: covid, deaths, rates, mortality, women  
Time of Screening Decision: 2020-12-29 09:53:36  
Retrieval Method: Faceted Search: no facets applied, Sort: {Year, desc}  
Similar Articles

Overall Summary

Inclusion Confidence: 0.49094

Inclusion Confidence: 0.51307

Inclusion Confidence: 0.33465

Inclusion Confidence: 0.4884

Included   Excluded

I E U   I E U   I E U   I E U

NaCTeM

The National Centre for Text Mining

# RobotAnalyst: Automated screening

Collection: Cancer Research

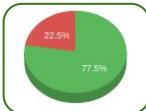


Clusters      Refine Search

Choose Algorithm: SC  
Number of Clusters: 5  
Get Clusters

There are no clusters to show.

Overall Summary



No undecided documents,  
but the user can still  
change the decisions

Screen: All      Decided      Undecided      Included      Excluded

Showing page 1 of 28 (409 results)  
317 included | 92 excluded | 0 predicted includes | 0 predicted excludes

Year      Descending      Go

Surgical treatment of peri-acetabular metastatic disease: Retrospective, multicentre study of 91 THA cases.  
Document ID: 2  
Year: 2020  
Most Relevant Topic: mobility, surgical, cup, dual, tha  
Time of Screening Decision: 2020-12-29 09:53:32  
Retrieval Method: Faceted Search: no facets applied, Sort: (Year, desc)  
Similar Articles

Overcoming resistance to anti-PD-1 immunotherapy in lymphoepithelioma-like carcinoma: A case report and review of the literature.  
Document ID: 3  
Year: 2020  
Most Relevant Topic: ielc, markers, rare, sfts, carcinoma  
Time of Screening Decision: 2020-12-29 09:53:33  
Retrieval Method: Faceted Search: no facets applied, Sort: (Year, desc)  
Similar Articles

Sex Differences in Mortality Rates and Underlying Conditions for COVID-19 Deaths in England and Wales.  
Document ID: 4  
Year: 2020  
Most Relevant Topic: covid, deaths, rates, mortality, women  
Time of Screening Decision: 2020-12-29 09:53:36  
Retrieval Method: Faceted Search: no facets applied, Sort: (Year, desc)  
Similar Articles

Medicaid Expansion and Mortality Among Patients With Breast, Lung, and Colorectal Cancer.  
Document ID: 5  
Year: 2020  
Most Relevant Topic: expansion, medicaid, mortality, states, lung  
Time of Screening Decision: 2020-12-29 09:53:38  
Retrieval Method: Faceted Search: no facets applied, Sort: (Year, desc)  
Similar Articles

Mediastinal Metastasis of Breast Cancer Mimicking a Primary Mediastinal Tumor.

Included      Excluded  
Included      Excluded  
Included      Excluded  
Included      Excluded

# RobotAnalyst: Clustering

**Username:** efstathiachristopoulou  
**Password:** thinarrangement57

Clusters    Refine Search

Choose Algorithm: SC

Number of Clusters: 5

Get Clusters

Size

Descending    Go

Cluster size

patient, background, year, month, undergo, ci, lymphocytic, multivariate analysis, median, woman, cohort, interval, rate, follow-up, man (115)

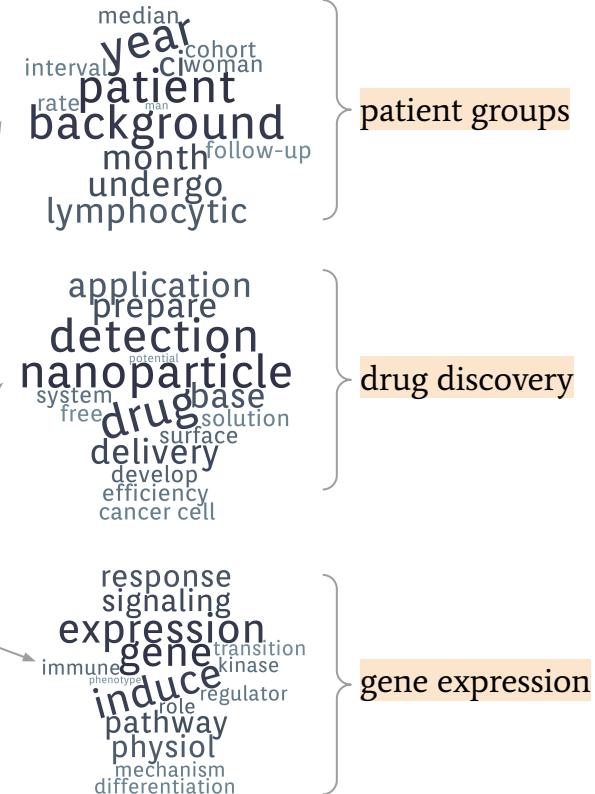
nanoparticle, detection, drug, delivery, prepare, base, application, develop, surface, system, efficiency, cancer cell, solution, free, potential (114)

gene, expression, induce, signaling, pathway, response, physiol, role, immune, kinase, regulator, mechanism, differentiation, transition, phenotype (91)

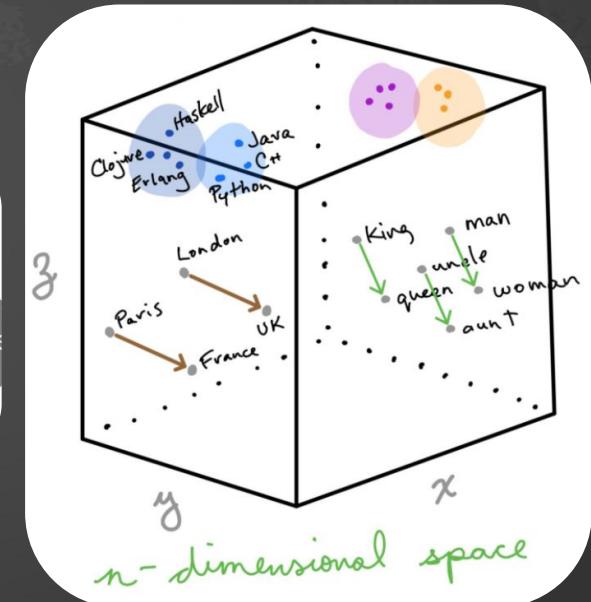
compound, extract, ic50, activity, context, antioxidant, cytotoxic, anti-inflammatory, acid, constituent (45)

management, smoking, association, health, corrigendum, hepatocellular, cardiovascular disease, individual, obesity, frequency (35)

Can be used  
for topic-based  
filtering bown  
(on click)



# Representations



Source: Jon Krohn [untapt](#) Safari Live Lessons

# Representations

... aka, how can we encode language in a machine-readable format? 🤔

- The straightforward way: **Rules**
  - Cumbersome
  - Fail to adapt to the complexity of language
  - Not as expressive as natural language can be
- The more elegant-flexible-intuitive way: **Embeddings**
  - Useful properties - closer to natural language
  - Implicit encoding of rules - easily generalisable
  - Demonstrably state-of-the-art performance on NLP tasks

# Language Models: Evolution

- Initial representations were discrete symbols, aka **One-hot vectors**
  - 1 in position i indicates the i-th word, rest positions are 0

---- Vocabulary ----	
amoxil	0 0 0 ... 0 0 1 0 0 ... 0 0 0 ... 0
mRNA	0 0 0 ... 0 0 0 0 ... 0 1 0 ... 0
surgery	0 0 0 ... 1 0 0 0 ... 0 0 0 ... 0

Size of the vector  
is equal to the size  
of the vocabulary

# Language Models: Evolution

- Initial representations were discrete symbols, aka **One-hot vectors**
  - 1 in position  $i$  indicates the  $i$ -th word, rest positions are 0
  - Drawbacks:** Requirement for VERY large vocabularies  
No knowledge about some form of word “meaning”

amoxil      0 0 0 ... 0 0 1 0 0 ... 0 0 0 ... 0

mRNA      0 0 0 ... 0 0 0 0 ... 0 1 0 ... 0

surgery      0 0 0 ... 1 0 0 0 ... 0 0 0 ... 0

Size of the vector  
is equal to the size  
of the vocabulary

# Language Models: Evolution

- Initial representations were discrete symbols, aka **One-hot vectors**
  - 1 in position  $i$  indicates the  $i$ -th word, rest positions are 0
  - **Drawbacks:** Requirement for VERY large vocabularies  
No knowledge about some form of word “meaning”
- Rely on context to understand “meaning”, aka **Distributional Semantics**

Words which frequently appear in similar contexts have similar meaning

([Harris, 1954](#))

# Language Models: Evolution

- Initial representations were discrete symbols, aka **One-hot vectors**
  - 1 in position  $i$  indicates the  $i$ -th word, rest positions are 0
  - **Drawbacks:** Requirement for VERY large vocabularies  
No knowledge about some form of word “meaning”
- Rely on context to understand “meaning”, aka **Distributional Semantics**

Words which frequently appear in similar contexts have similar meaning

([Harris, 1954](#))

- **Count-based methods:** Co-occurrence, Pointwise Mutual Information (PMI)
  - Encode information based on corpus statistics
- **Prediction-based methods:** Word2Vec ([Mikolov et al., 2013](#))
  - Learn to *predict* the context of a word (or the word from its context)
- **Hybrid:** GloVe ([Pennington et al., 2014](#))

# Count-based Methods

Count-based methods construct a **word-context matrix** with certain elements

- Co-occurrence ([Lund and Burgess, 1996](#))
  - **Words:** words
  - **Context:** A set of words in a fixed w-sized window (e.g.  $w = 3$ )
  - **Matrix elements:** Number of times word  $w$  appears in context  $c$

EGTA inhibited down-regulation of alpha-AR mRNA by NE .

	$w_1$	of	$w_3$	EGTA	$w_5$	$w_6$	NE	$w_8$	$w_9$	...
alpha-AR	0	1	0	1	0	0	0	0	0	...

- Typically, high co-occurrence → high similarity

# Count-based Methods

Count-based methods construct a **word-context matrix** with certain elements

- Pointwise Mutual Information

- **Words:** words
- **Context:** A set of words in a fixed w-sized window (e.g.  $w = 3$ )
- **Matrix elements:** PMI - a measure of association

Positive pmi indicates that words are correlated above chance!

$$pmi(x; y) = \log \frac{p(x,y)}{p(x)p(y)}$$

Probability words x and y co-occur

Probability of word x and word y

$$ppmi(x; y) = \max(0, pmi(x; y))$$

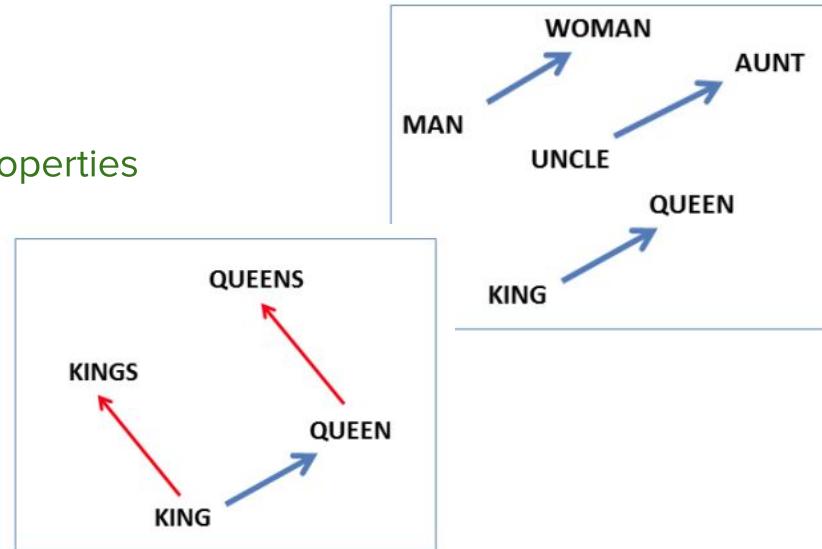
	$w_1$	of	$w_3$	EGTA	$w_5$	$w_6$	NE	$w_8$	$w_9$	...
alpha-AR	0	1.22	0	-3.78	0	0	0	0	0	...

# Prediction-based methods: Word2Vec

- Word embeddings are *learned* instead of constructed manually
  - Words:** words
  - Context:** A set of words in a fixed w-sized window (e.g.  $w = 3$ )
  - Probabilities:** Computed for each target word for each of their context
- Advantages**
  - Fast training
  - High quality word embeddings
  - Learning embeddings have some nice properties
  - Small dimensionality !

Algorithms:

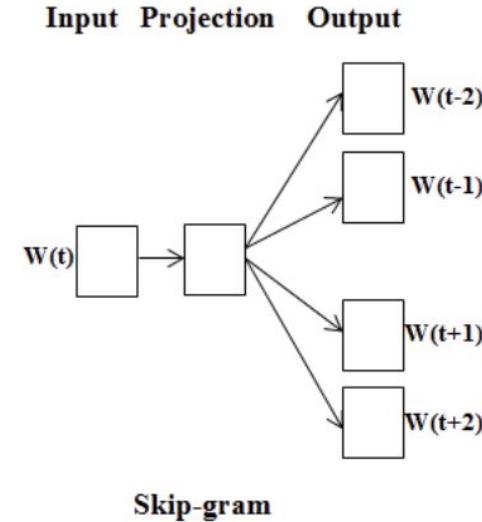
- Skip-Gram:* Predict the context given a word
- CBoW:* Predict a word given the context



# Prediction-based methods: Word2Vec

## Skip-gram

- Most famous and commonly used method
- Training is done in pairs using a **sliding window**  $w$
- Learnable embeddings via gradient descent
- 2 vectors are produced for each word
  - 1 as target word, 1 as context word
- **Negative Sampling** makes things faster
  - Use a subset of  $K$  less frequent words as context



Sentence

EGTA inhibited down-regulation of alpha-AR mRNA → (EGTA, inhibited) (EGTA, down-regulation)

EGTA inhibited down-regulation of alpha-AR mRNA → (inhibited, EGTA) (inhibited, down-regulation) (inhibited, of)

EGTA inhibited down-regulation of alpha-AR mRNA → (down-regulation, EGTA) (down-regulation, inhibited)  
(down-regulation, of) (down-regulation, alpha-AR)

# Hybrid: GloVe

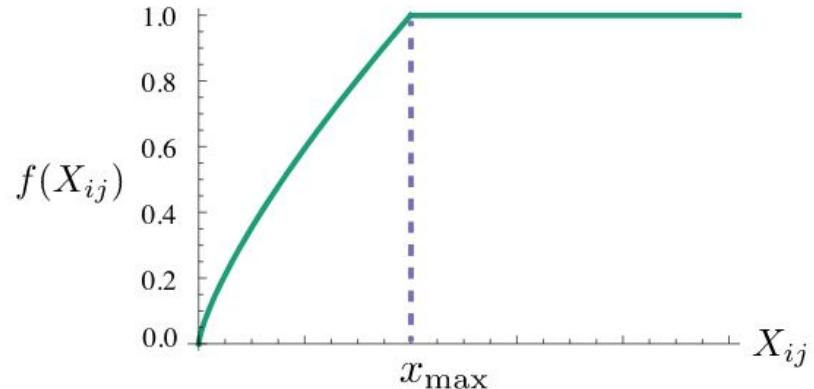
- Combination of **count-based** and **prediction-based** methods
- GloVe = **Global Vectors**
- Learned vectors via gradient descent

$$J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

Target word vector      Context word vector

Co-occurrence frequency

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$



Weighting function

The weighting function

- Penalises rare events
- Not overweight frequent events

# Language Models: Evolution

- Initial representations were discrete symbols, aka **One-hot vectors**
  - 1 in position i indicates the i-th word, rest positions are 0
  - Drawbacks:** Requirement for VERY large vocabularies  
No knowledge about some form of word “meaning”
- Rely on context to understand “meaning”, aka **Distributional Semantics**

Words which frequently appear in similar contexts have similar meaning

([Harris, 1954](#))

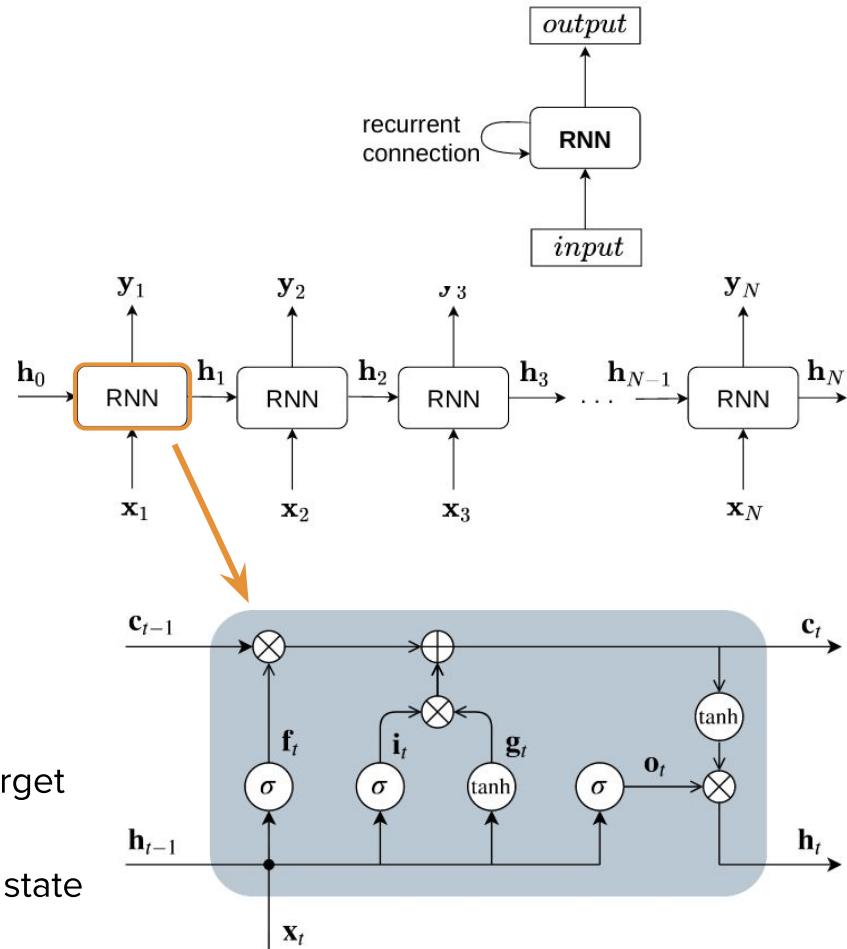
- Count-based methods:** Co-occurrence, Pointwise Mutual Information (PMI)
  - Encode information based on corpus statistics
- Prediction-based methods:** Word2Vec ([Mikolov et al., 2013](#))
  - Learn to *predict* the context of a word (or the word from its context)
- Hybrid:** GloVe ([Pennington et al., 2014](#))  
**Drawbacks:** The encoded “meaning” remains the same across different contexts

# Language Models: Evolution

- Different embeddings for different contexts, aka **Contextual Word Embeddings**
  - **ELMo** which is based on Bi-LSTMs ([Peters et al., 2018](#))
  - Predict the next word in a sentence given the current (and previous) words

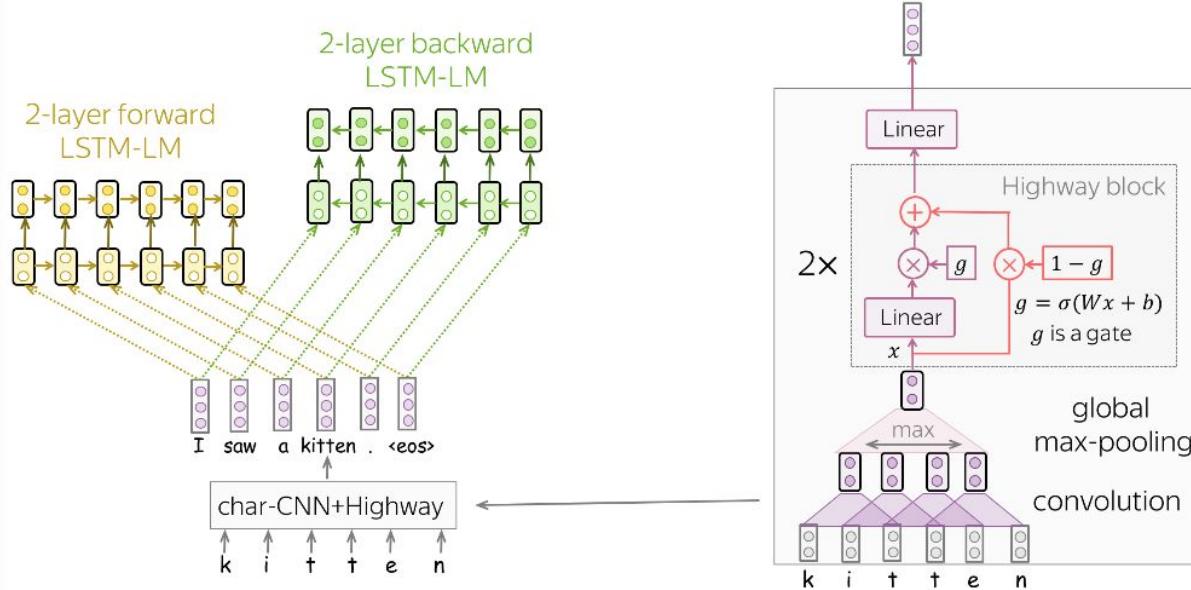
# Recap: RNNs & LSTMs

- Recurrent Neural Networks ([Elman, 1990](#))
  - Previous elements used as input
  - Have an internal “memory” aka hidden state ( $\mathbf{h}$ )
  - Contextual word embeddings ( $\mathbf{y}$ )
  - Drawback: Cannot model very long dependencies
- 
- Long-Short Term Memory ([Hochreiter et al., 1997](#))
  - Able to learn long dependencies
  - The LSTM cell has 4 layers called “gates”
  - LSTM cell is our “memory”
    - *Forget gate*: How much information we want to forget
    - *Input gate*: Amount of incoming information
    - *Output gate*: Amount of memory revealed to next state
    - *G gate*: Simple RNN cell



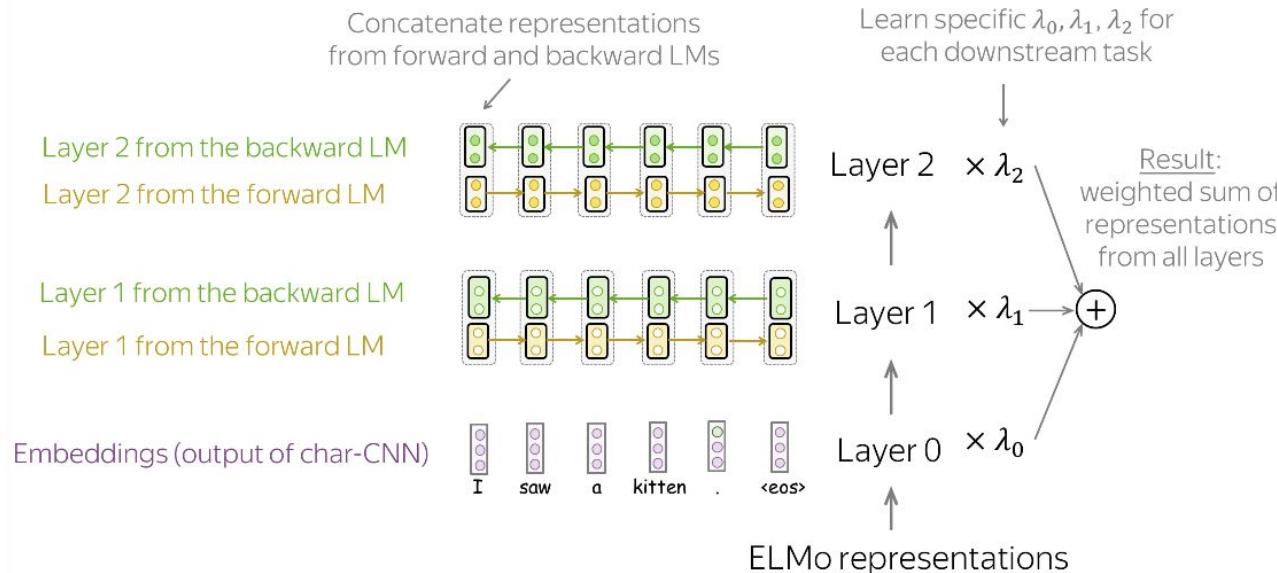
# ELMo: Embeddings for Language Models

- Produce contextualised word embeddings by looking at the entire sentence
- Different context - different embedding
- Successor of the CoVe model ([McCann et al., 2017](#)) introduced for Machine Translation
- Able to fine-tune it on downstream tasks !



# ELMo: Embeddings for Language Models

- Produce contextualised word embeddings by looking at the entire sentence
- Different context - different embedding
- Successor of the CoVe model ([McCann et al., 2017](#)) introduced for Machine Translation
- Able to fine-tune it on downstream tasks !



# Language Models: Evolution

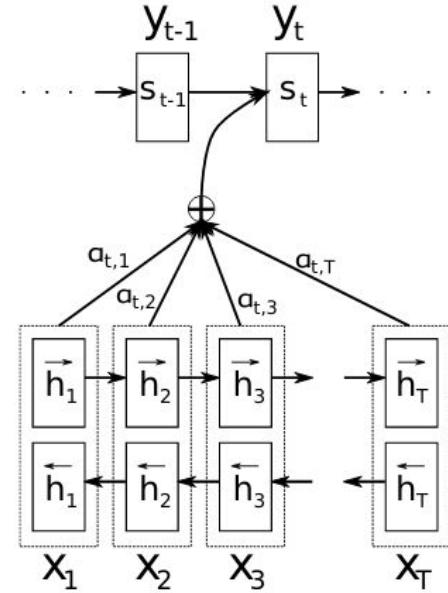
- Different embeddings for different contexts, aka **Contextual Word Embeddings**
  - **ELMo** which is based on Bi-LSTMs ([Peters et al., 2018](#))
  - Predict the next word in a sentence given the current (and previous) words
  - **Drawbacks:** Very *slow* to train  
Rely mostly on word order making it easy to “forget” long dependencies

# Language Models: Evolution

- Different embeddings for different contexts, aka **Contextual Word Embeddings**
  - **ELMo** which is based on Bi-LSTMs ([Peters et al., 2018](#))
  - Predict the next word in a sentence given the current (and previous) words
  - **Drawbacks:** Very *slow* to train  
Rely mostly on word order making it easy to “forget” long dependencies
- From word order to attention, aka **Transformers**
  - **Attention** heads learn which words are important when looking at a certain word
  - **Masked Language Modeling:** Fill-in the blank (i.e. find the word given the context)

# Attention in Neural Networks

- Concept originally adapted from computer vision
  - Learning to “attend” to specific places in the image
- Measures the “importance” of a representation towards another
- Utilised first in NLP for machine translation  
[\(Bahdanau et al., 2015\)](#)
- Attention comes in many flavours
  - Additive, Dot-product, ...



$$\mathbf{c}_t = \sum_{i=1}^n \alpha_{t,i} \mathbf{h}_i$$

; Context vector for output  $y_t$

$$\alpha_{t,i} = \text{align}(y_t, x_i)$$

; How well two words  $y_t$  and  $x_i$  are aligned.

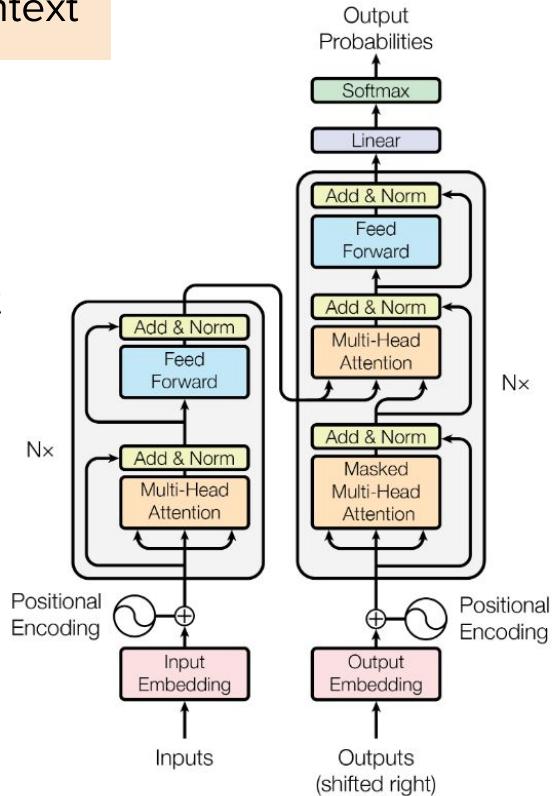
$$= \frac{\exp(\text{score}(\mathbf{s}_{t-1}, \mathbf{h}_i))}{\sum_{i'=1}^n \exp(\text{score}(\mathbf{s}_{t-1}, \mathbf{h}_{i'}))}$$

; Softmax of some predefined alignment score..

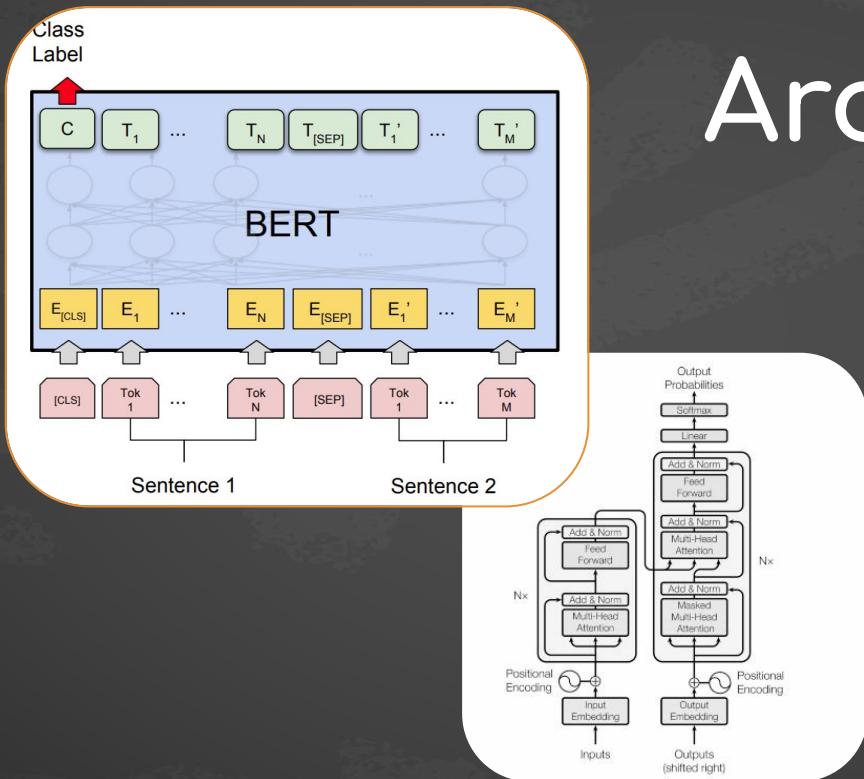
# Transformers

Use of **attention** to capture word context

- Introduced in [Vaswani et al. \(2018\)](#)
- Use of attention-heads to identify significant words for each concept representation
- Stack multiple transformers blocks to capture more distant dependencies
- Training: Masked Language Modelling
  - GPT ([Radford et al., 2019](#)): next token prediction
  - BERT ([Devlin et al., 2018](#)): next token prediction + next sentence prediction
  - ELECTRA ([Clark et al., 2020](#)): Valid/Invalid token prediction



# Architectures



# Positional Features

- Such features are mostly common in *Relation Extraction*
- They are used to encode information about the “place” of a word in a snippet

## 1. Entity Markers ([Zhang et al., 2015; Soares et al., 2019](#))

- Explicitly learn representations for them as additional words

`<e1>EGTA</e1>` inhibited down-regulation of `<e2>alpha-AR</e2>` mRNA by NE

## 2. Relative Positions ([Collobert et al., 2011; Zeng et al., 2014, Zhang et al., 2017](#))

- Measure the relative position of one word towards another

The diagram illustrates relative positions between words in a sentence. It shows two arrows pointing from the word "EGTA" to the word "alpha-AR". The first arrow points to the right and is labeled "4", indicating the word is 4 tokens away. The second arrow points to the left and is labeled "-3", indicating the word is 3 tokens away.

EGTA inhibited down-regulation of alpha-AR mRNA by NE

## 3. Position Indicators ([Soares et al., 2019](#))

- indicator embedding layer

EGTA inhibited down - regulation of alpha - AR mRNA by NE

1	0	0	0	0	2	2	2	0	0	0
---	---	---	---	---	---	---	---	---	---	---

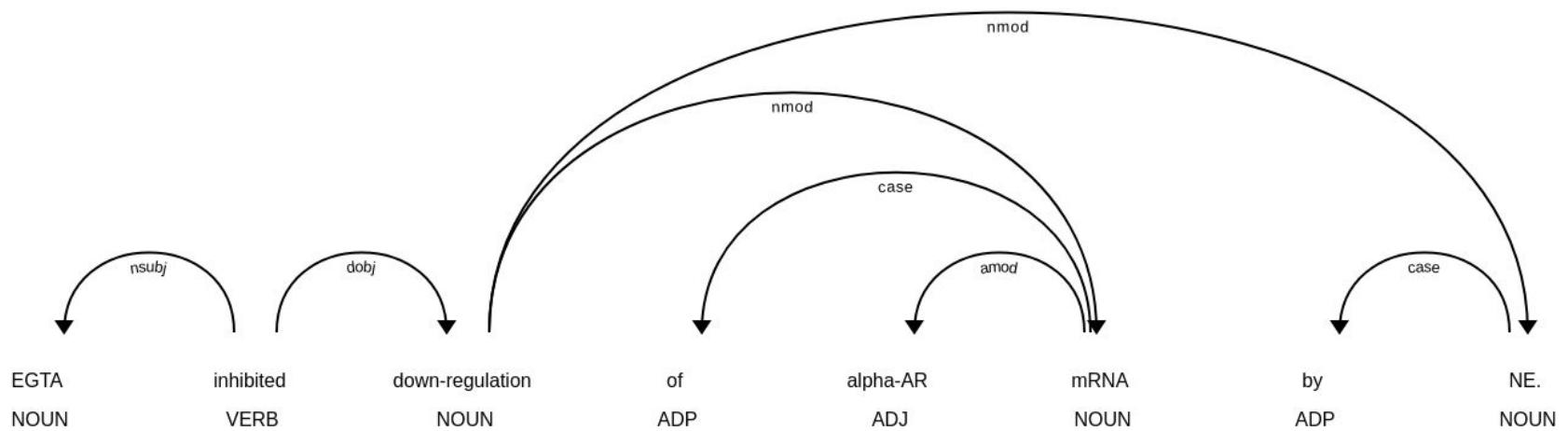
# Linguistic features

- Linguistic features are designed to help models encode linguistic information into their vector representations
- Most commonly used Linguistic features:
  - **Part-of-speech (POS) tags:** VERB, NOUN, etc

EGTA	inhibited	down-regulation	of	alpha-AR	mRNA	by	NE
<i>NOUN</i>	<i>VERB</i>	<i>NOUN</i>	<i>ADP</i>	<i>ADJ</i>	<i>NOUN</i>	<i>ADP</i>	<i>NOUN</i>

# Linguistic features

- Linguistic features are designed to help models encode linguistic information into their vector representations
- Most commonly used Linguistic features:
  - **Part-of-speech (POS) tags:** VERB, NOUN, etc
  - **Dependency paths** (most common to use the shortest dependency path: SDP)



# Linguistic features

- Linguistic features are designed to help models encode linguistic information into their vector representations
- Most commonly used Linguistic features:
  - **Part-of-speech (POS) tags:** VERB, NOUN, etc
  - **Dependency paths** (most common to use the shortest dependency path: SDP)
  - **Lemmas:** Assign the basic form of a word, e.g. “was” → “be”, “breathing” → “breathe”

EGTA	inhibited	down-regulation	of	alpha-AR	mRNA	by	NE
<i>EGTA</i>	<i>inhibit</i>	down-regulation	<i>of</i>	alpha-AR	<i>mRNA</i>	<i>by</i>	<i>NE</i>

# Linguistic features

- Linguistic features are designed to help models encode linguistic information into their vector representations
- Most commonly used Linguistic features:
  - **Part-of-speech (POS) tags:** VERB, NOUN, etc
  - **Dependency paths** (most common to use the shortest dependency path: SDP)
  - **Lemmas:** Assign the basic form of a word, e.g. “was” → “be”, “breathing” → “breathe”
  - **Surface Form:** The form in which a word appears
  - **Characters:** The characters that comprise a word
  - **Word Length**
  - **Semantics:** Different meanings of a word, e.g. [WordNet synsets](#)

<i>Synsets</i>	<i>Definition</i>
suppress	to put down by force or authority
inhibit	limit the range or extent of
inhibit	limit, block, or decrease the action or function of
inhibit	control and refrain from showing; of emotions, desires, impulses, or behavior

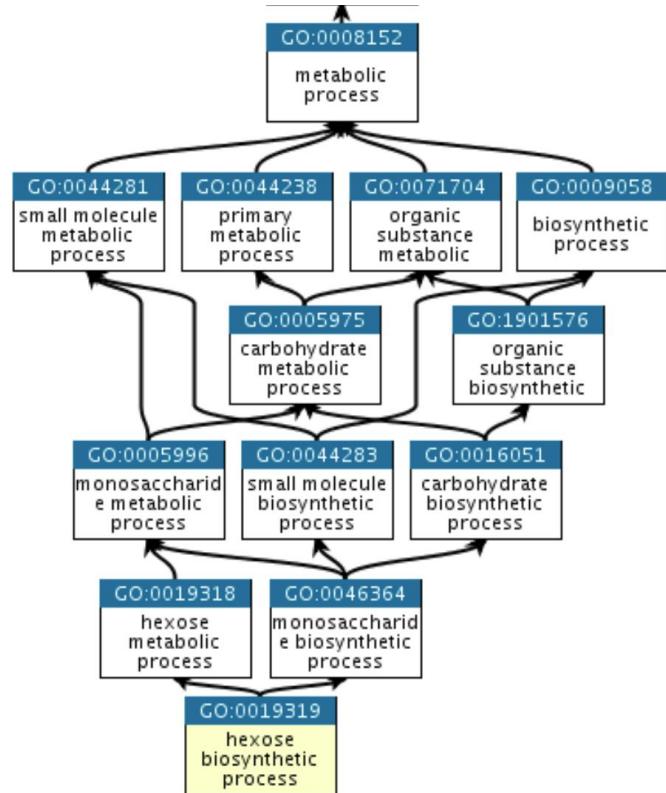
# External Features

- External features can be information derived from:
  - **Dictionaries:**
    - Included in the dictionary (1), not included (0)

Dictionary Category	Examples
<b>Causation</b>	because, effect
<b>Articles</b>	a, an, the
<b>Negation</b>	no, not, never
<b>Exclusives</b>	but, without, exclude
<b>Nonfluencies</b>	er, hm, umm
<b>Positive emotion</b>	love, nice, sweet
<b>Negative emotion</b>	hurt, ugly, nasty

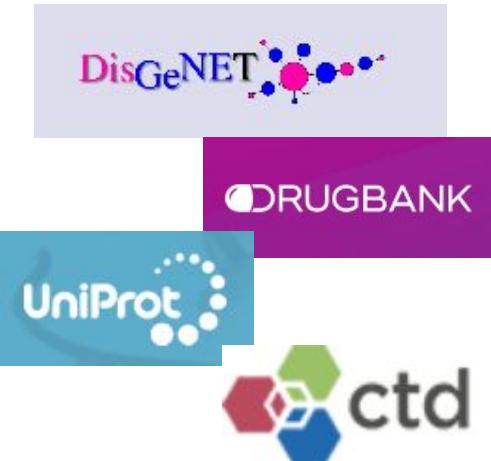
# External Features

- External features can be information derived from:
  - **Dictionaries:**
    - Included in the dictionary (1)
  - **Ontologies:**
    - N-step hypernyms of hyponyms of an entity

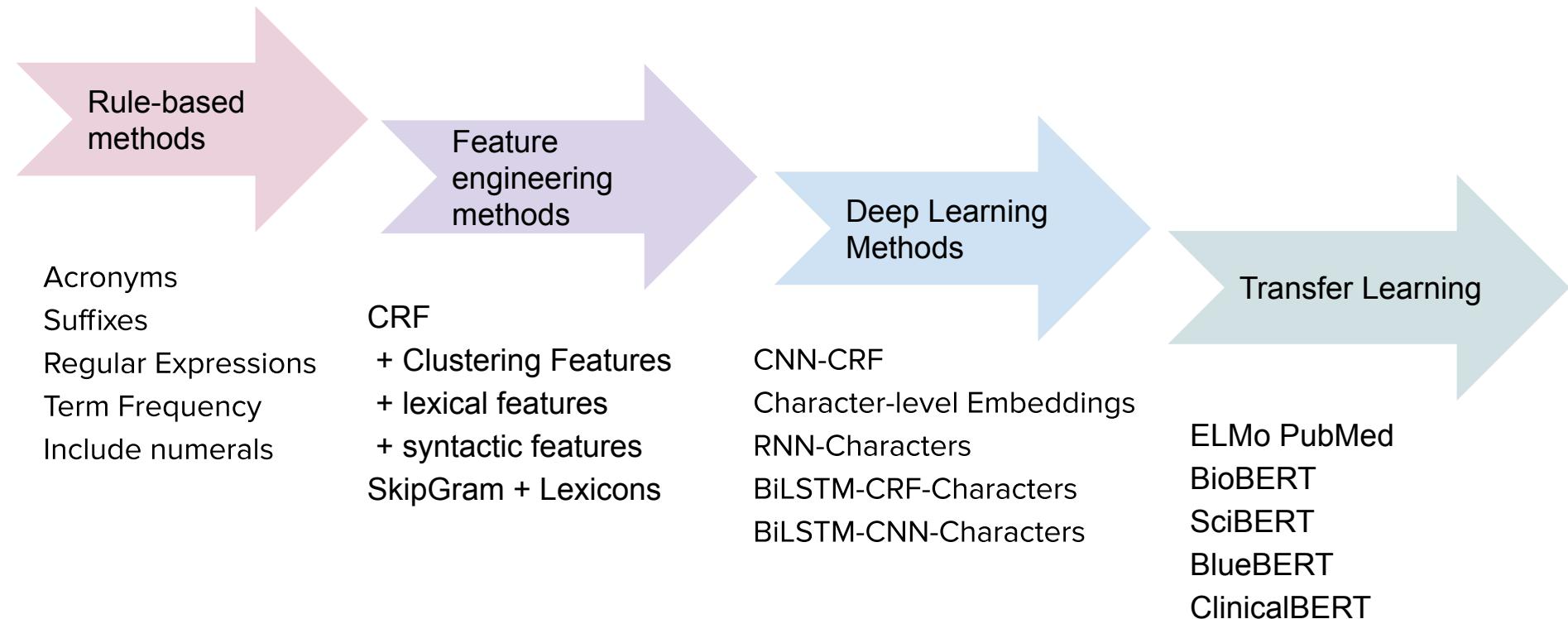


# External Features

- External features can be information derived from:
  - **Dictionaries:**
    - Included in the dictionary (1), not included (0)
  - **Ontologies:**
    - N-step hypernyms of hyponyms of an entity
  - **Knowledge Bases:** *Source of factual knowledge*
    - Associations with other objects in the KB
    - Semantic types if the entity exists in the KB



# Named Entity Recognition



# Named Entity Recognition

1. Initial approaches were **rule-based** ([Hettne et al., 2009](#))

Eg. *Detection of Acronyms:*

words with caps only, more than one cap, letters and digits, etc

2. Later on, **Machine Learning** models were used with **feature engineering**

- Tagging schemes, Gazetteers ([Ratinov and Roth, 2009](#))
  - Early indication that pre-training on large resources will be useful!
- CRF + Clustering features ([Lin and Wu, 2009](#))
- CRF + Phrase embeddings ([Passos et al., 2014](#))

Conditional Random Field (CRF) ([Lafferty et al., 2001](#))

- A probabilistic model that takes into account neighboring samples
- Handles dependencies between the samples
- Linear chain CRF is thus suitable for modelling the dependencies among sentential words

# NER: Neural Approaches

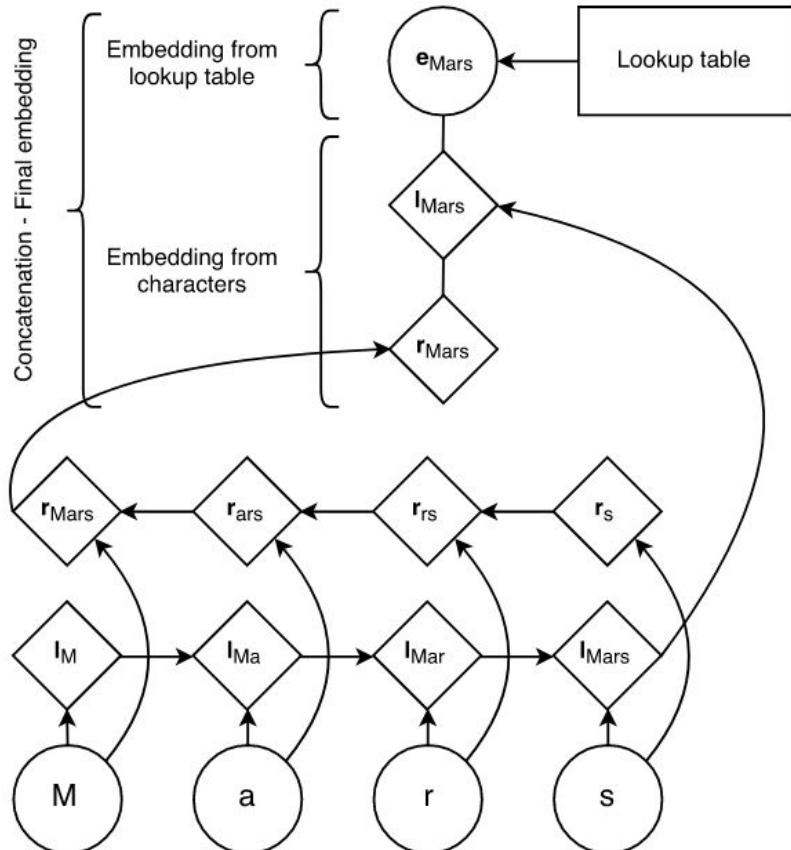
## → BiLSTM-CRF-Char

Proposed by [Lample et al. \(2016\)](#)

- [Nogueira and Guimaraes \(2015\)](#) proposed **character-level embeddings** based on CNN
- Useful for morphologically rich languages
- Helpful for out-of-vocabulary words

Steps for Character Embeddings:

1. Feed characters of a word into a BiLSTM
2. Concatenate the hidden states of forward and backward LSTM
3. Concatenate these with the word embedding



# NER: Neural Approaches

Word Embeddings as input:

- Created as we saw earlier

BiLSTM Layer:

- Concatenation of forward and backward states  $\rightarrow [h_{\text{forw}} ; h_{\text{back}}]$

CRF Layer:

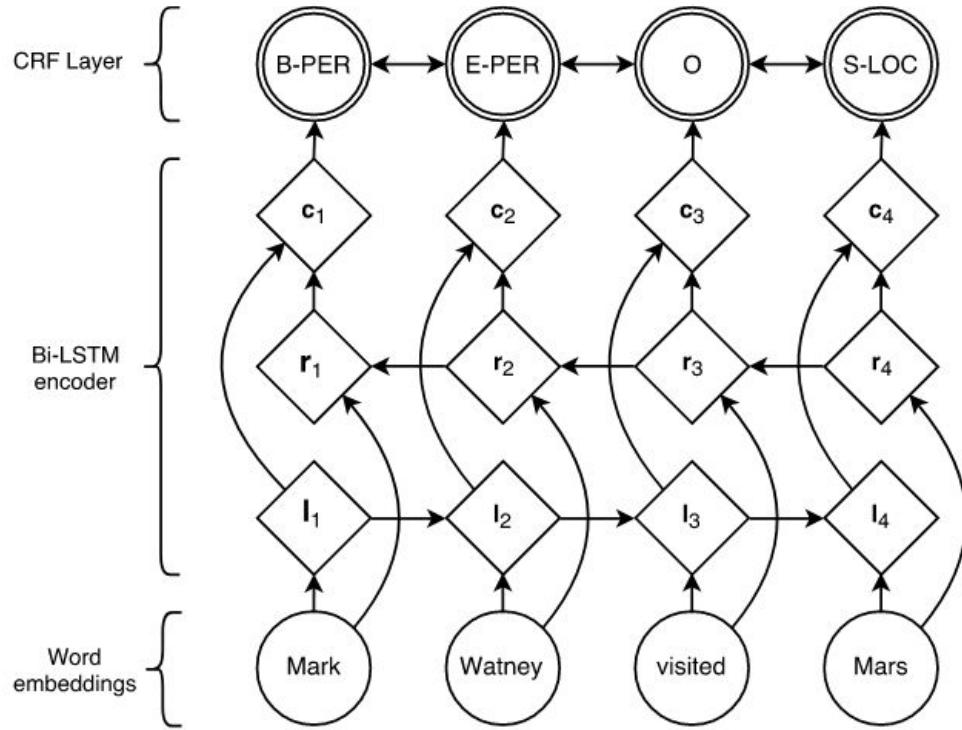
Transition scores matrix

$$s(\mathbf{X}, \mathbf{y}) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i}$$

sequence      labels      logits

$$\mathbf{y}^* = \operatorname{argmax}_{\tilde{\mathbf{y}} \in \mathbf{Y}_x} s(\mathbf{X}, \tilde{\mathbf{y}})$$

All possible tag sequences

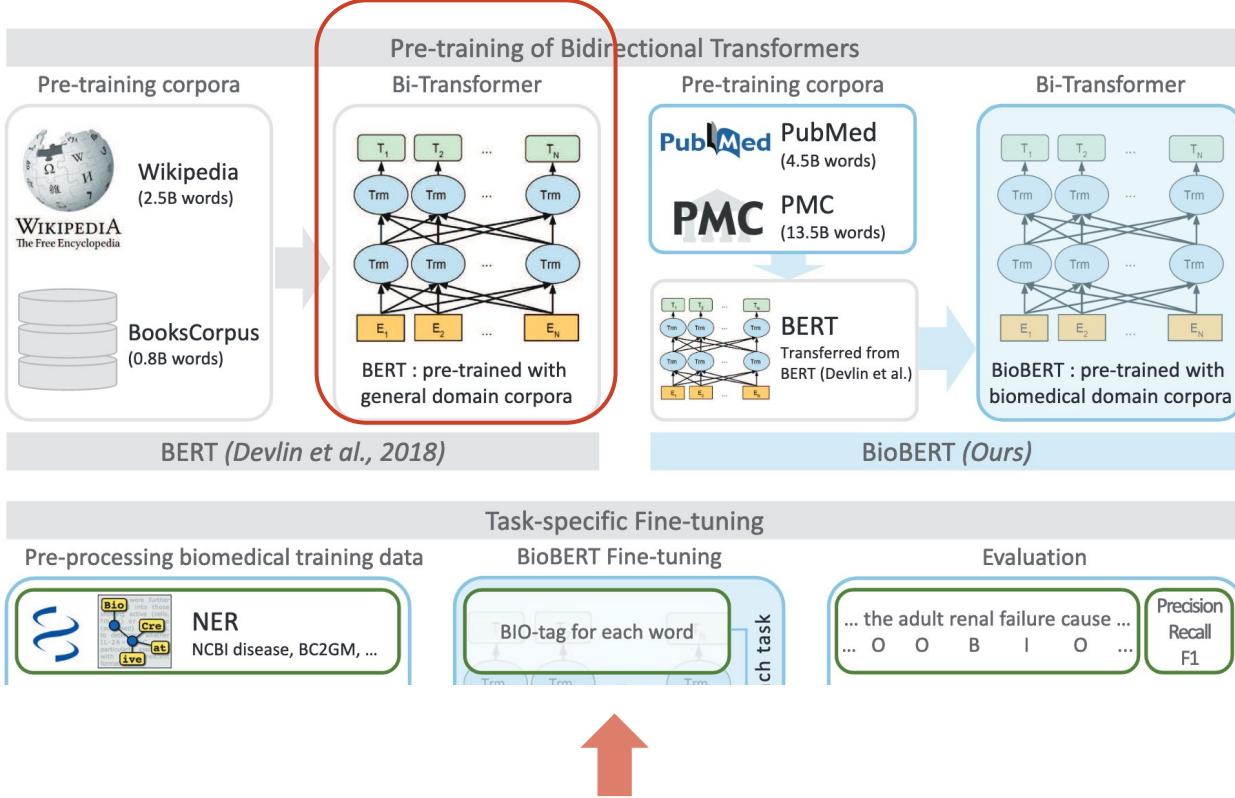


# NER: Neural Approaches

Subsequent approaches experimented with several combinations:

- RNN-Char ([Sahu and Anand, 2016](#))
  - Comparison of multiple RNN architectures
  - Showed the **RNNs perform better than CNNs on the NER task**
- BiLSTM-CNN-Char ([Chiu and Nichols, 2016](#); [Kocaman and Talby, 2020](#))
  - **Convolutional layer on characters** is better than sequential!
- BiLSTM-CRF with dependencies ([Jie and Lu, 2019](#))
  - Strong correlations between the entity types and dependency relations
- BiLSTM-CRF with ELMo ([Dai et al., 2019](#))
  - **CRF** continues to be used as an **effective feature**
  - **Pre-training** on large corpora benefits representations and performance

# NER: Pre-trained LMs



## BioBERT (Lee et al., 2019)

- Initialised on BERT, trained on:
  - English Wikipedia (2.5B words)
  - BooksCorpus (0.8B words)
- Further trained on:
  - PubMed Abstracts (4.5B words)
  - PMC Full-text Articles (13.5B words)

# NER: Pre-trained LMs

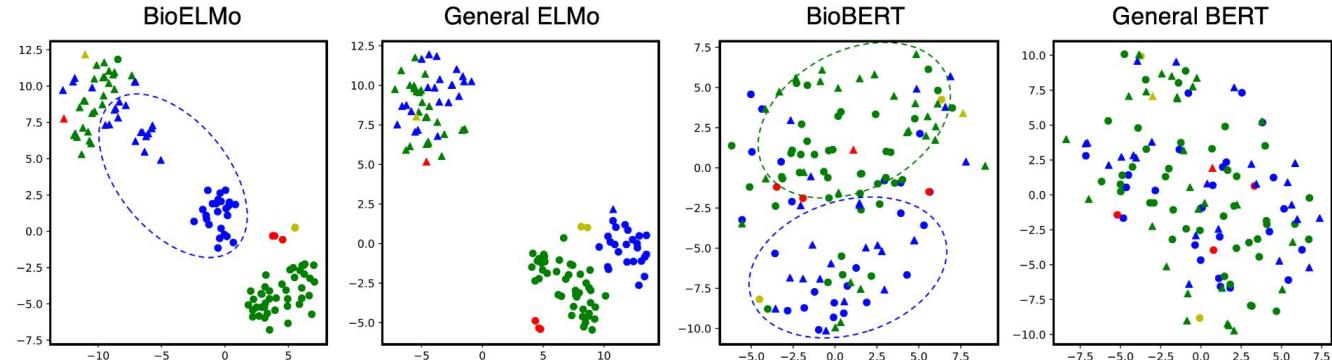
Clinical BERT ([Alsentzer et al., 2019](#)) → Trained on 2M nodes from [MIMIC-III](#)

Model	i2b2 2006	i2b2 2010	i2b2 2012	i2b2 2014
BERT	93.9	83.5	75.9	92.8
BioBERT	<b>94.8</b>	86.5	78.9	<b>93.0</b>
Clinical BERT	91.5	86.4	78.5	92.6
Discharge Summary BERT	91.9	86.4	78.4	92.8
Bio+Clinical BERT	94.7	87.2	<b>78.9</b>	92.5
Bio+Discharge Summary BERT	94.8	<b>87.8</b>	78.9	92.7

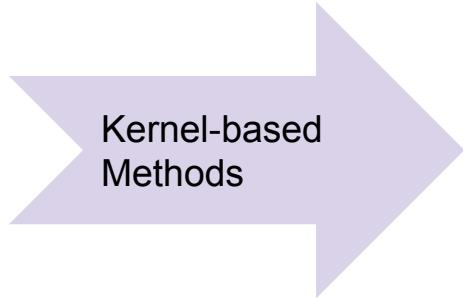
In-domain training vs out-of-domain training

**BioELMo**  
([Jin et al., 2019](#))

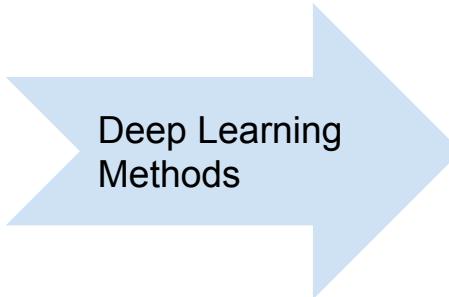
→ Trained on PubMed  
(10M recent abstracts,  
2.46B tokens)



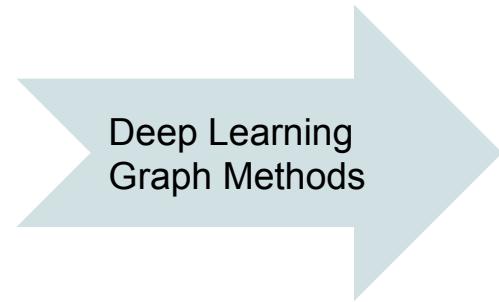
# Relation Extraction



Tree-based SVM kernel  
Sub-sequence kernel  
Shortest Dependency Path



LSTM-based  
CNN-based  
+ Positional Features  
+ Tree-structures  
+ Attention

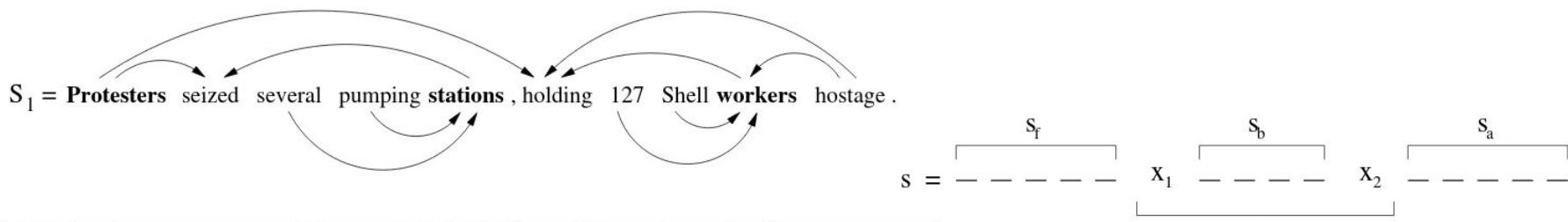


Pre-trained LMs  
o GPT-2  
o BERT

# Relation Extraction

## Kernel-based Methods

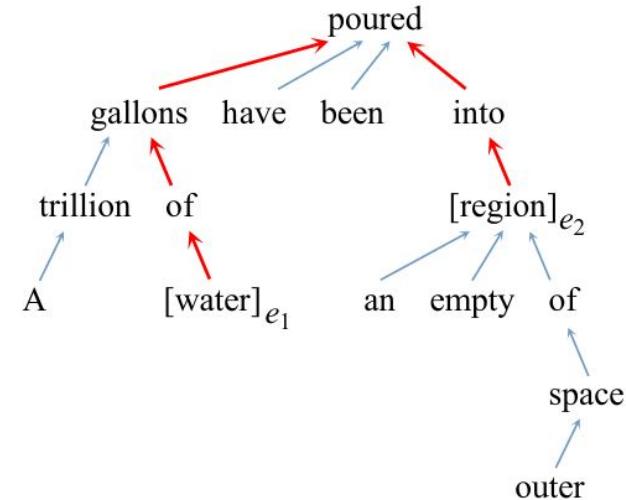
- Relying on using Support Vector Machines (SVMs) with different kernels
  - Sub-sequence Kernel ([Bunescu and Mooney 2005](#))
  - Tree-kernel ([Zelenko et al., 2003; Bunescu and Mooney, 2006](#))



Relation Instance	Shortest Path in Undirected Dependency Graph
$S_1: \text{protesters AT stations}$	<b>protesters</b> → seized ← <b>stations</b>
$S_1: \text{workers AT stations}$	<b>workers</b> → holding ← protesters → seized ← <b>stations</b>
$S_2: \text{troops AT churches}$	<b>troops</b> → raided ← <b>churches</b>
$S_2: \text{ministers AT churches}$	<b>ministers</b> → warning ← troops → raided ← <b>churches</b>

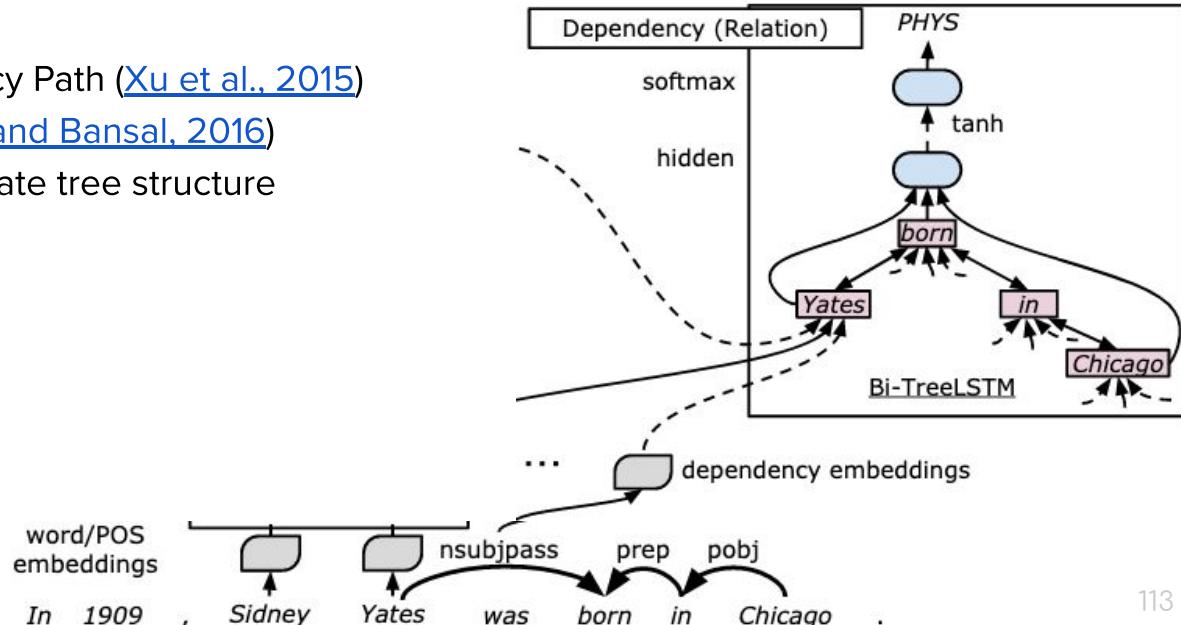
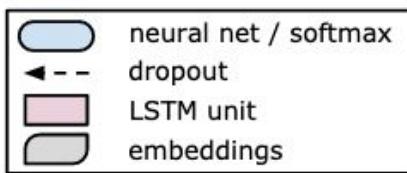
# Relation Extraction: Neural Approaches

- Using **position features**
  - CNN + positions ([Zeng et al., 2014](#))
  - RNN + max pooling + **entity markers** ([Zhang and Wang, 2015](#))
- Using **tree-structures**
  - LSTM + Shortest Dependency Path ([Xu et al., 2015](#))



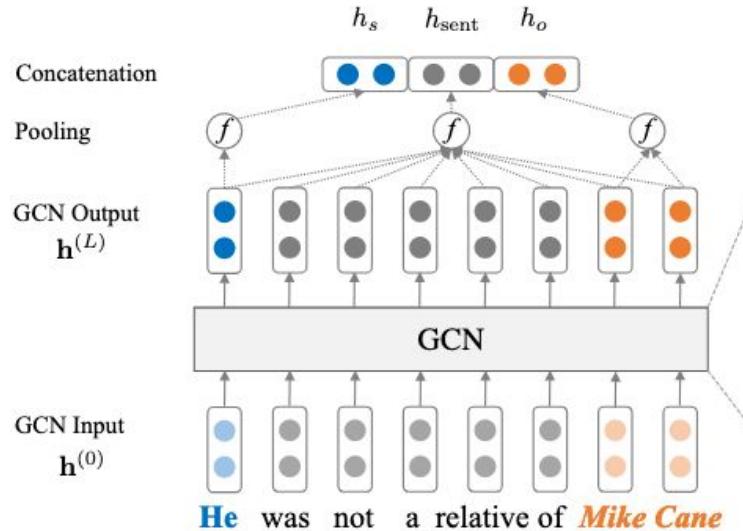
# Relation Extraction: Neural Approaches

- Using **position features**
  - CNN + positions ([Zeng et al., 2014](#))
  - RNN + max pooling + **entity markers** ([Zhang and Wang, 2015](#))
- Using **tree-structures**
  - LSTM + Shortest Dependency Path ([Xu et al., 2015](#))
  - BiLSTM + Tree LSTM ([Miwa and Bansal, 2016](#))
    - Selecting the appropriate tree structure



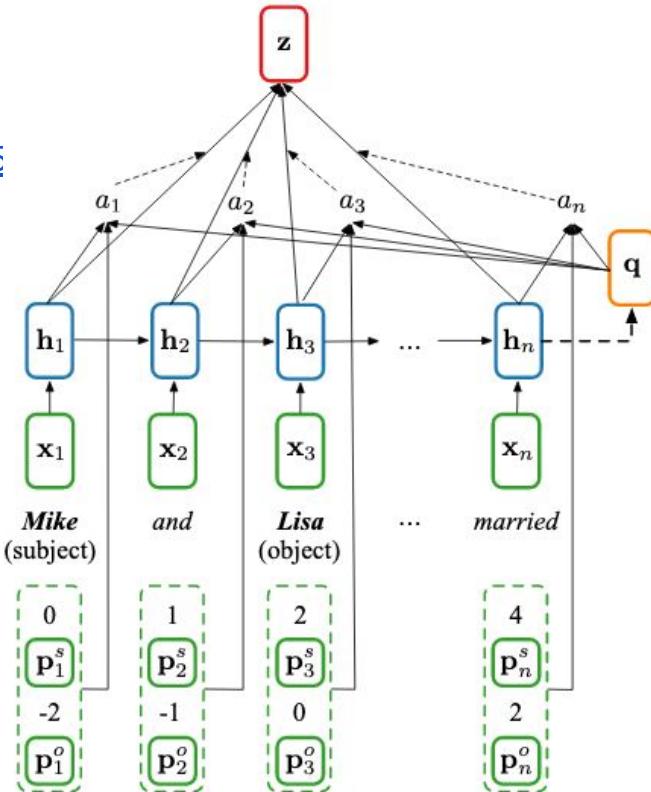
# Relation Extraction: Neural Approaches

- Using **position features**
  - CNN + positions ([Zeng et al., 2014](#))
  - RNN + max pooling + **entity markers** ([Zhang and Wang, 2015](#))
- Using **tree-structures**
  - LSTM + Shortest Dependency Path ([Xu et al., 2015](#))
  - BiLSTM + Tree LSTM ([Miwa and Bansal, 2016](#))
    - Selecting the appropriate tree structure
  - GCN on trees ([Zhang et al., 2018](#))
    - Not all edges are important
  - Graph LSTM ([Song et al., 2018](#))



# Relation Extraction: Neural Approaches

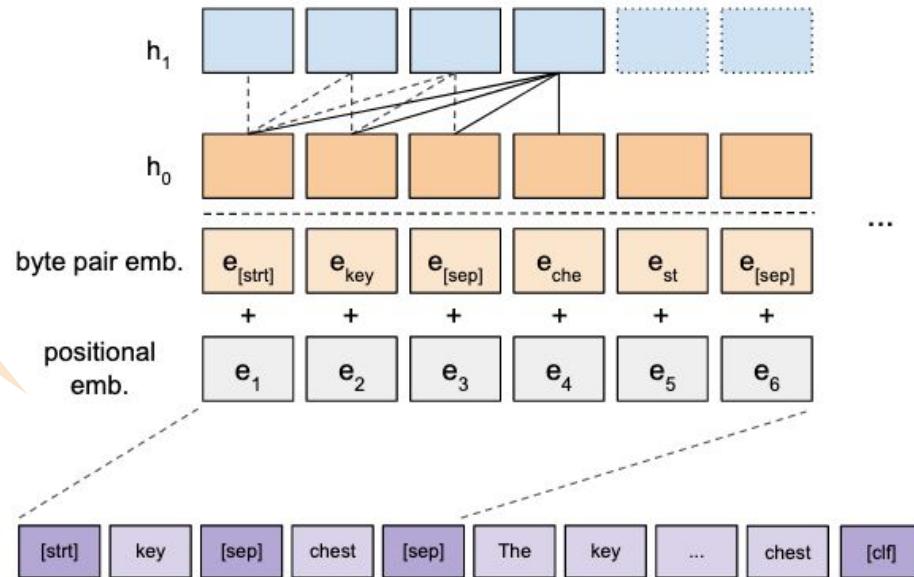
- Using **position features**
  - CNN + positions ([Zeng et al., 2014](#))
  - RNN + max pooling + **entity markers** ([Zhang and Wang, 2015](#))
- Using **tree-structures**
  - LSTM + Shortest Dependency Path ([Xu et al., 2015](#))
  - BiLSTM + Tree LSTM ([Miwa and Bansal, 2016](#))
    - Selecting the appropriate tree structure is important
  - GCN on trees ([Zhang et al., 2018](#))
    - Not all edges are important
  - Graph LSTM ([Song et al., 2018](#))
- Using **attention**
  - CNN + Attention ([Wang et al., 2016](#))
  - BiLSTM + Attention ([Zhou et al., 2016](#))
  - BiLSTM + Attention + Relative Positions ([Zhang et al., 2017](#))



# Relation Extraction: Pre-trained LMs

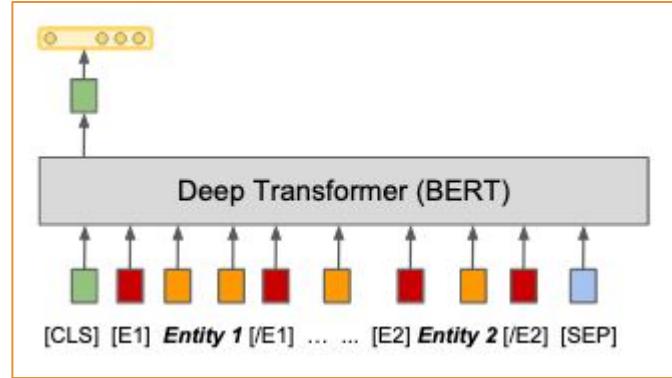
- GPT-2 ([Alt et al., 2019](#))

Arguments are appended at the beginning to attend to the rest of the sentence



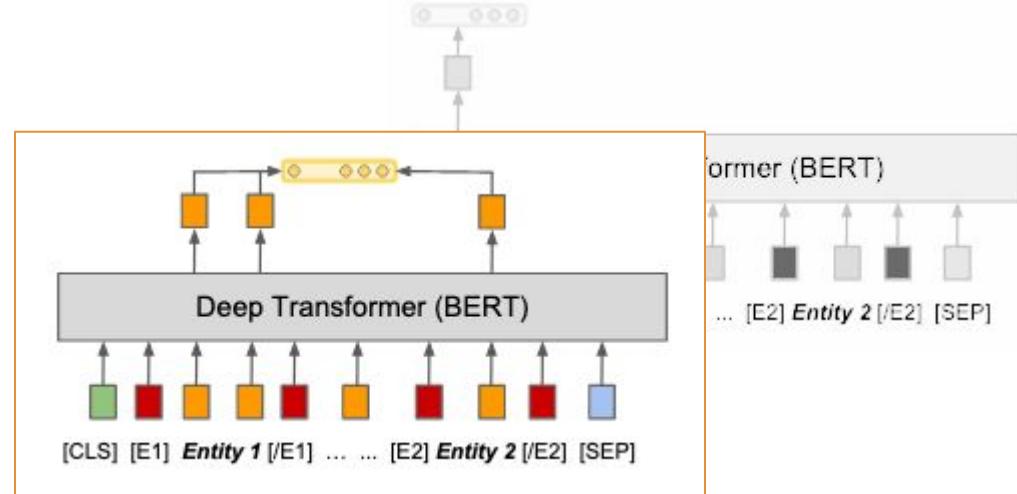
# Relation Extraction: Pre-trained LMs

- GPT-2 ([Alt et al., 2019](#))
- BERT ([Soares et al., 2019](#))
  - Task as “fill-in the blanks”



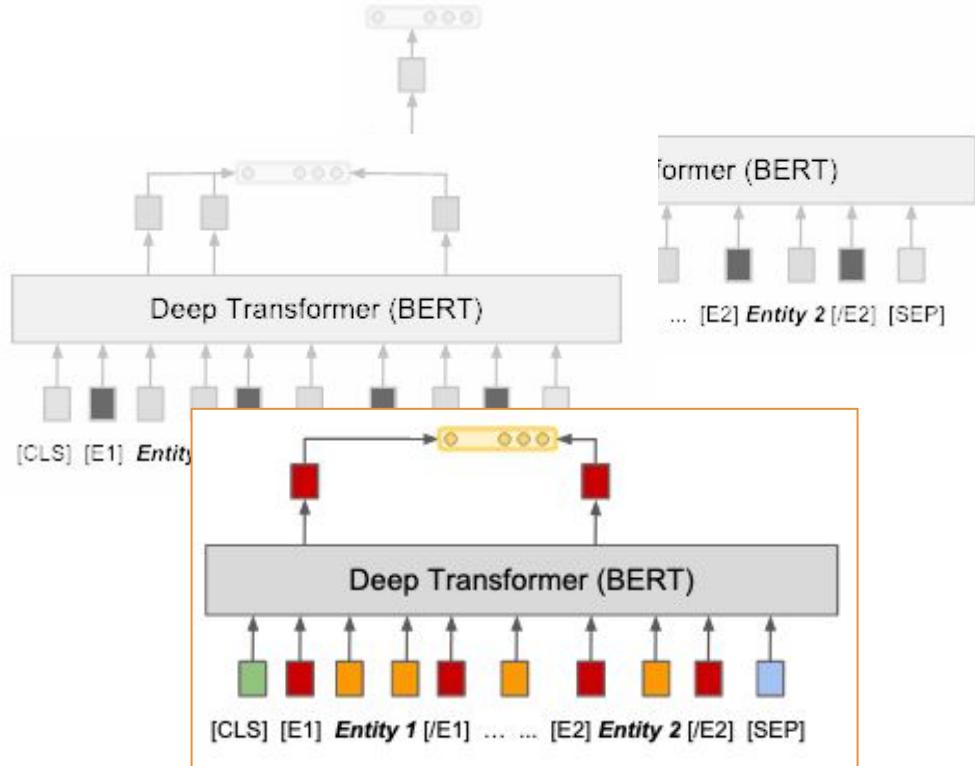
# Relation Extraction: Pre-trained LMs

- GPT-2 ([Alt et al., 2019](#))
- BERT ([Soares et al., 2019](#))
  - Task as “fill-in the blanks”



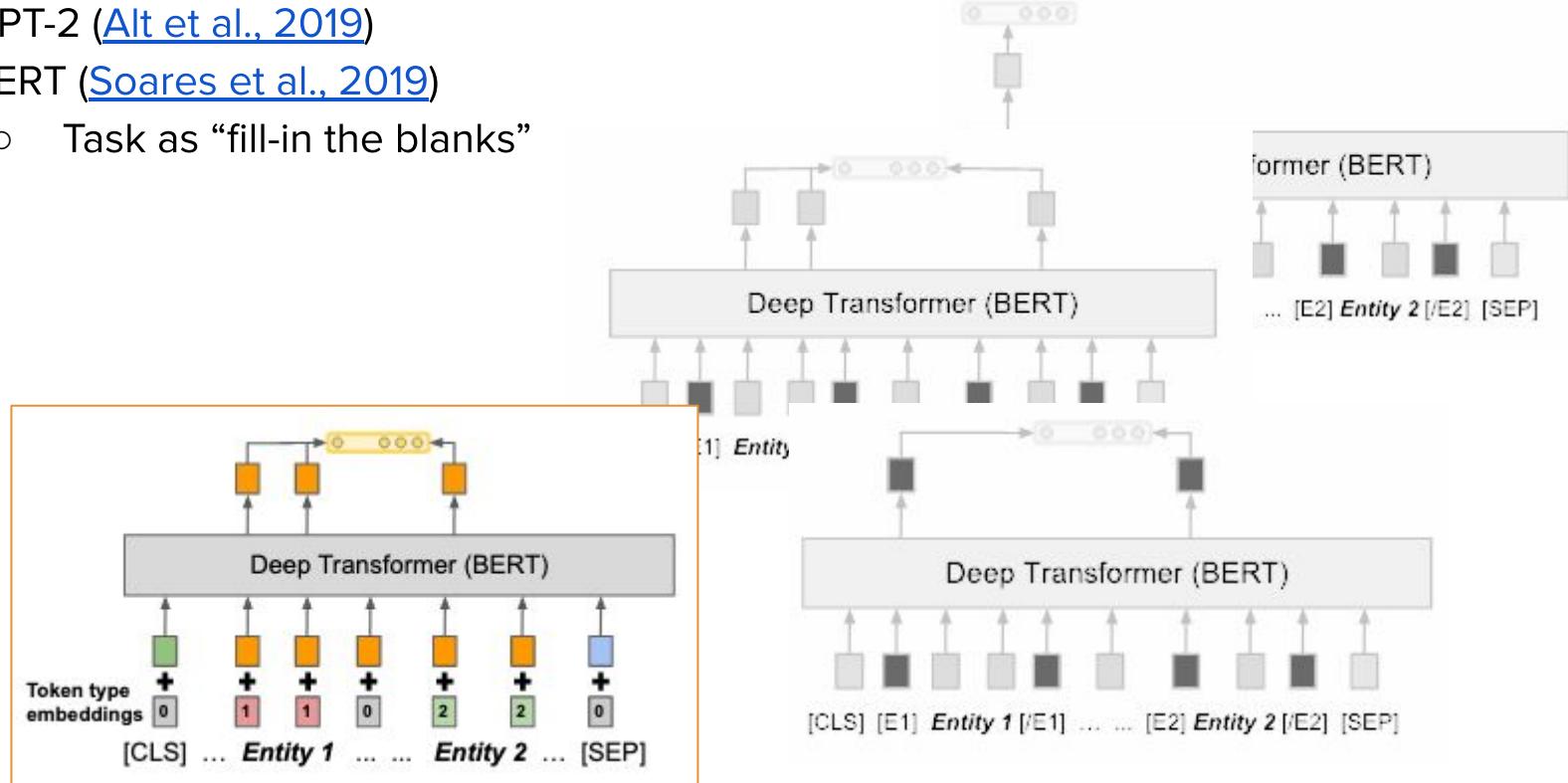
# Relation Extraction: Pre-trained LMs

- GPT-2 ([Alt et al., 2019](#))
- BERT ([Soares et al., 2019](#))
  - Task as “fill-in the blanks”



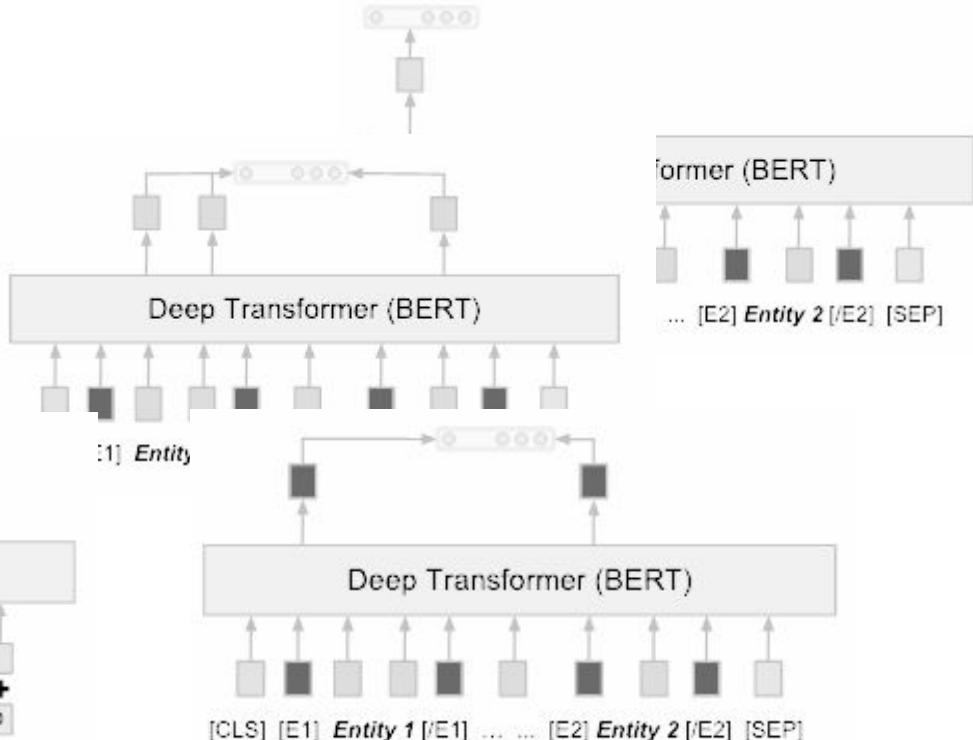
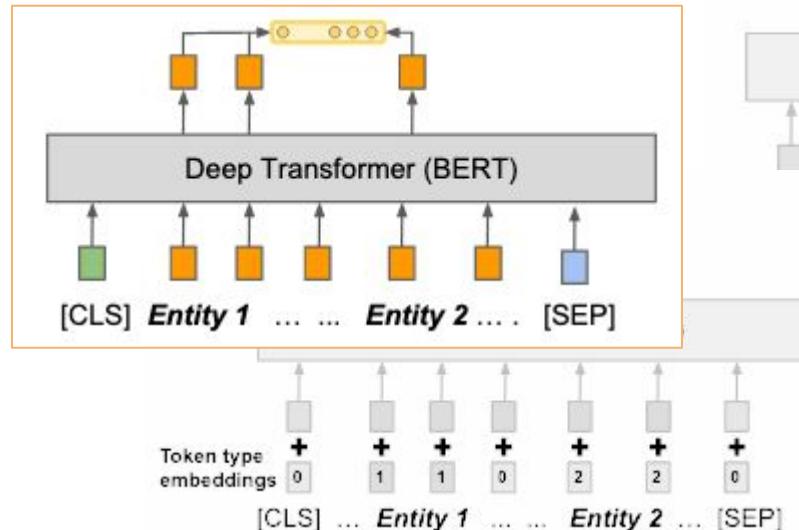
# Relation Extraction: Pre-trained LMs

- GPT-2 ([Alt et al., 2019](#))
- BERT ([Soares et al., 2019](#))
  - Task as “fill-in the blanks”

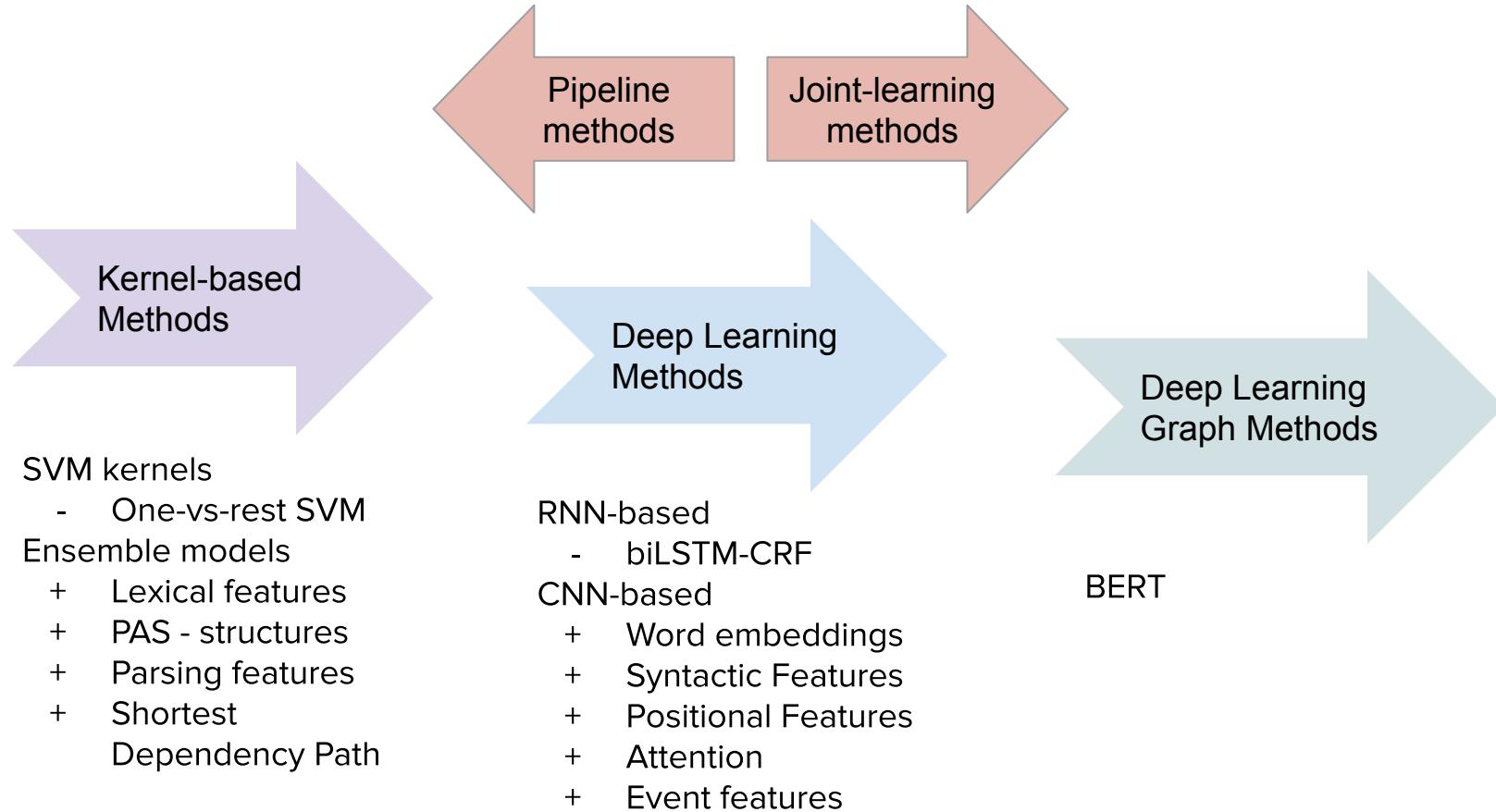


# Relation Extraction: Pre-trained LMs

- GPT-2 ([Alt et al., 2019](#))
- BERT ([Soares et al., 2019](#))
  - Task as “fill-in the blanks”



# Event Extraction



# Event Extraction

**General assumption:** EE can be broken down to sub-tasks:

NER | Trigger detection | Argument detection | Flat event prediction | Nested event prediction | Metaknowledge

## Pipeline methods

*train a set of independent classifiers for each subtask*

- Modular approach
- Assumption of subtask independence
- Error propagation

## Joint-learning methods

*learn (some) subtasks jointly*

- Shared parameters
- Assumption of subtask correlation
- Transformation of initial subtasks
  - Link prediction - MLN (Vlachos et al, 2011)
  - Information networks (Li et al., 2014)
  - Joint learning trigger-relation-EE with shared CNN parameters

# Event Extraction

## Kernel-based methods:

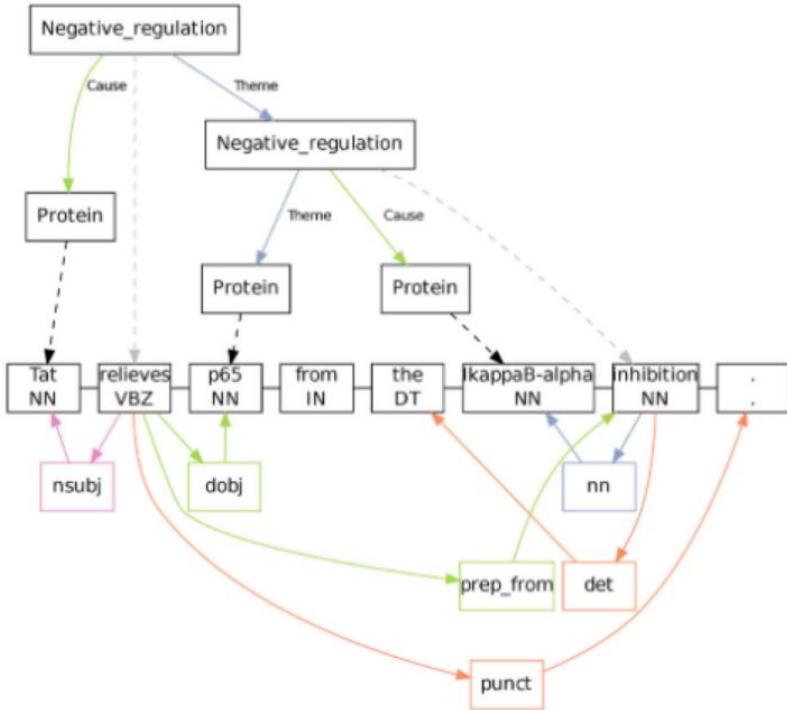
SVMs with different feature-sets:

([Bjorne et. al, 2013](#), [Miwa et al., 2013](#))

- Dependency parsing - tree structures
- Syntactic parsing
- Semantic features
- + External resources ([Miwa et. al, 2012](#))

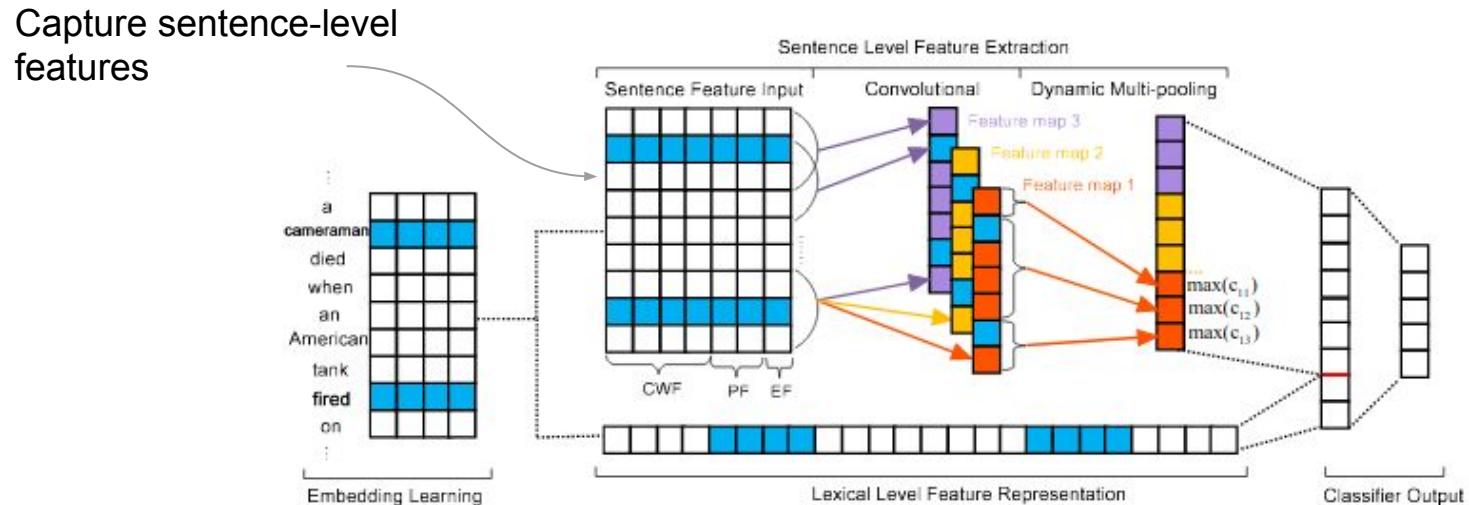
## Ensemble methods:

([Bali et al., 2020](#))



# Event Extraction: Neural methods

- Pipelined architectures:
  - CNN with dynamic pooling ([Chen et al., 2015](#))



# Event Extraction: Neural methods

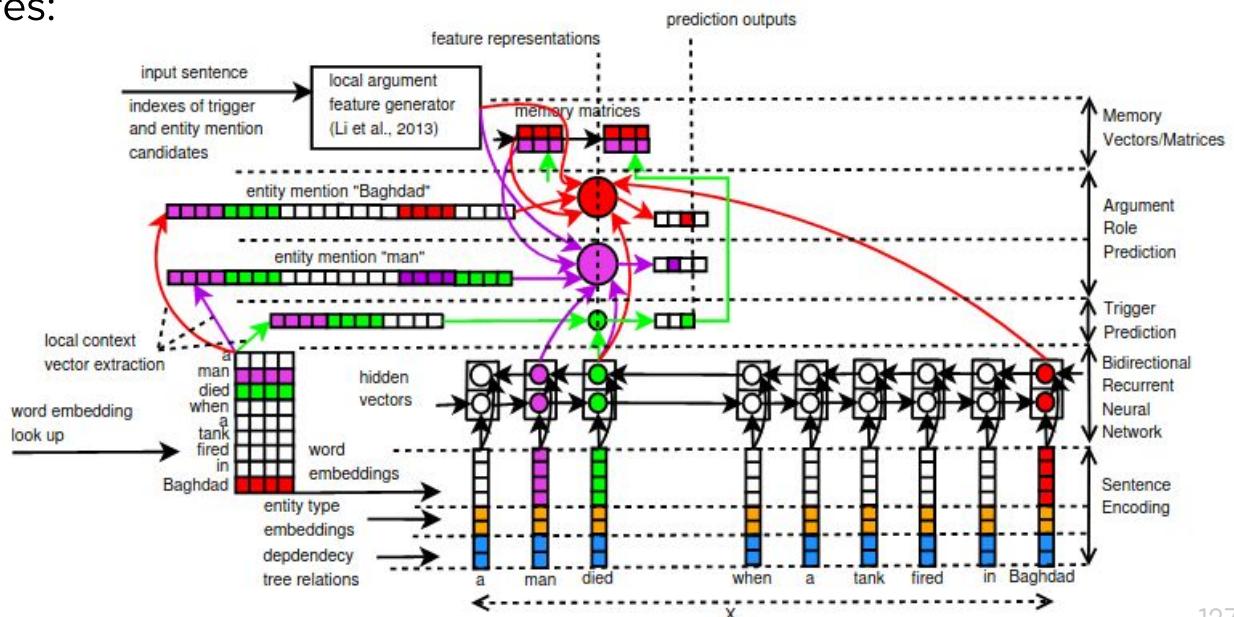
- Pipelined architectures:
  - CNN with dynamic pooling ([Chen et al., 2015](#))
  - CNN with dependency path embeddings ([Bjorne et al., 2018](#))

# Event Extraction: Neural methods

- Pipelined architectures:
  - CNN with dynamic pooling ([Chen et al., 2015](#)):
  - CNN with dependency path embeddings ([Bjorne et al., 2018](#))
- Joint-learning architectures:
  - RNN with memory ([Nguyen et al., 2016](#))

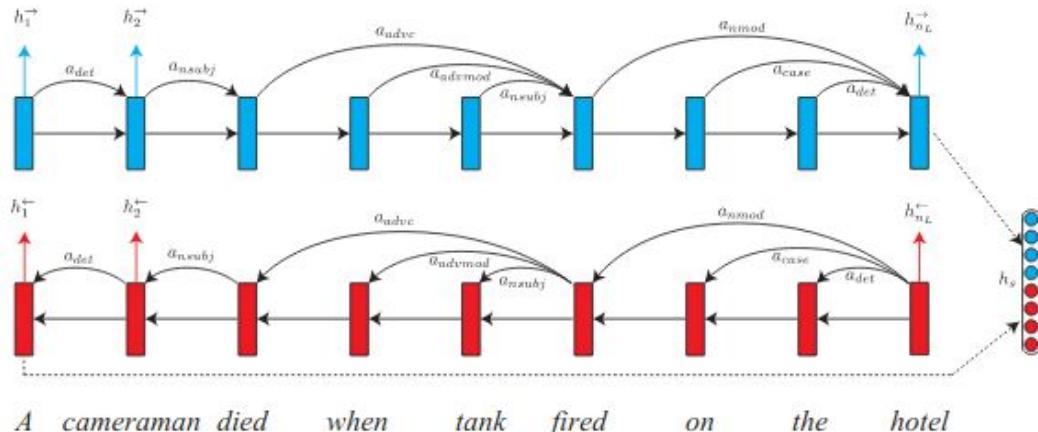
Assumption of task dependencies

- among trigger types
- among role types
- between argument roles and trigger subtypes



# Event Extraction: Neural methods

- Pipelined architectures:
  - CNN with dynamic pooling ([Chen et al., 2015](#)):
  - CNN with dependency path embeddings ([Bjorne et al., 2018](#))
- Joint-learning architectures:
  - RNN with memory  
([Nguyen et al., 2016](#))
  - RNN with dependency “Bridges” ([Sha et al., 2018](#))



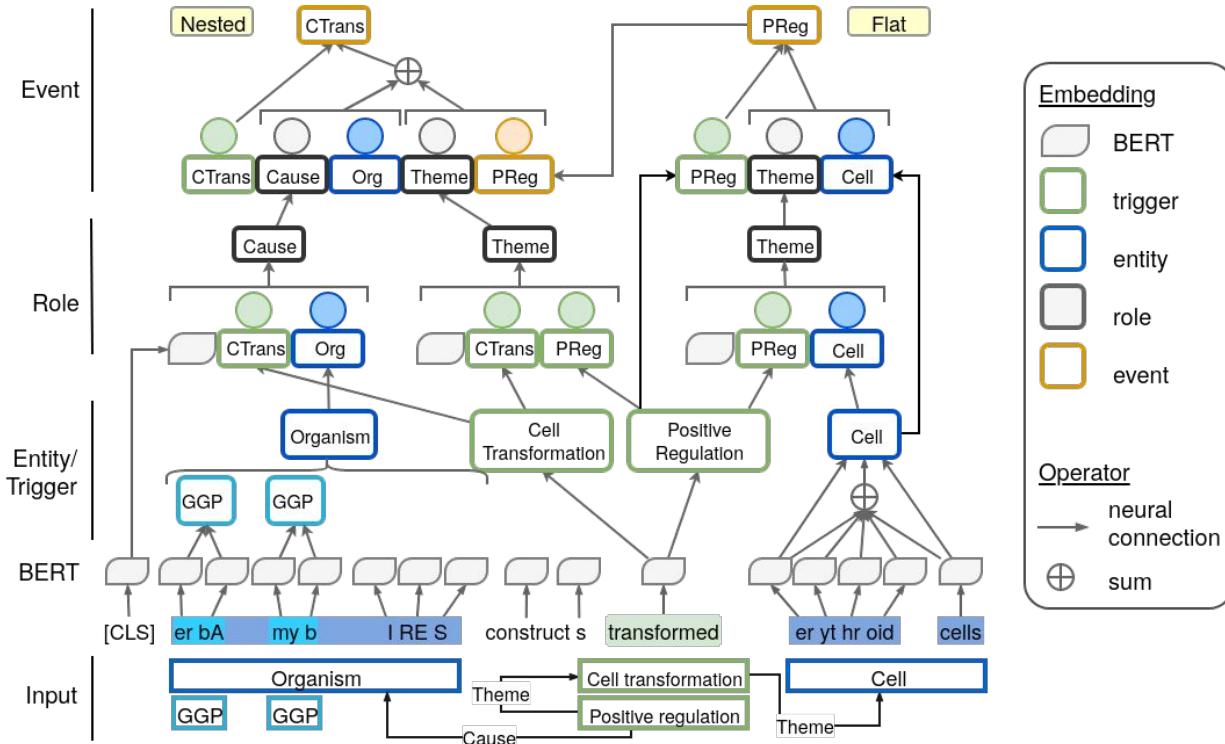
# Event Extraction: Neural methods

- Pipelined architectures:
  - CNN with dynamic pooling ([Chen et al., 2015](#)):
  - CNN with dependency path embeddings ([Bjorne et al., 2018](#))
- Joint-learning architectures:
  - RNN with memory  
([Nguyen et al., 2016](#))
  - RNN with dependency  
“Bridges” ([Sha et al., 2018](#))
  - CNN with dependency embeddings ([Li et. al, 2020](#))

# Event Extraction:

BERT-based (Trieu et. al, 2020)

- Two settings:
  - Pipelined
  - Shared parameters
- Sub-word embeddings



# Hands-on !



try:

```
    start = date(int(self.start_year.get()),  
                 self.months.index(self.start_month.get()),  
                 int(self.start_day.get()))
```

```
    end = date(int(self.end_year.get()),  
               self.months.index(self.end_month.get()),  
               int(self.end_day.get()))
```

# Subword Segmentation: Byte Pair Encoding

- **Learnable subword segmentation** → Byte Pair Encoding (BPE)
  - Introduced by [Sennrich et al. \(2016\)](#)
  - It originates in data compression ([Gage, 1994](#))
  - Replace common pairs of consecutive bytes with a byte that does not appear
- In order to perform **subword tokenization** BPE is modified
- Goal is to represent **rare** and **unseen words** as a sequence of **subword units**
- Hybrid between character- and word-level representations

This is a method to learn how to tokenize  
NOT learn embeddings

Frequently occurring **subword pairs** are *merged* together

athazagoraphobia

ath | az | agor | aphobia

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
a, t	19
m, a	10
t, s	2
t, e	6
e, a	2

<i>word</i>	<i>count</i>
c a t	4
m a t	5
m a t s	2
m a t e	3
a t e	3
e a t	2

Merges (total 5):

a t → at

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
a, t	19
m, a	10
t, s	2
t, e	6
e, a	2

<i>word</i>	<i>count</i>
c at	4
m at	5
m at s	2
m at e	3
at e	3
e at	2

Merges (total 5):

a t → at

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
c, at	4
m, at	10
at, s	2
at, e	6
e, at	2

<i>word</i>	<i>count</i>
c at	4
m at	5
m at s	2
m at e	3
at e	3
e at	2

Merges (total 5):

a t → at  
m at → mat

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
c, at	4
m, at	10
at, s	2
at, e	6
e, at	2

<i>word</i>	<i>count</i>
c at	4
mat	5
mat s	2
mat e	3
at e	3
e at	2

Merges (total 5):

a t → at  
m at → mat

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
c, at	4
mat, s	2
mat, e	3
at, e	3
e, at	2

<i>word</i>	<i>count</i>
c at	4
mat	5
mat s	2
mat e	3
at e	3
e at	2

Merges (total 5):

a t → at  
m at → mat  
c at → cat

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
c, at	4
mat, s	2
mat, e	3
at, e	3
e, at	2

<i>word</i>	<i>count</i>
cat	4
mat	5
mat s	2
mat e	3
at e	3
e at	2

Merges (total 5):

a t → at  
m at → mat  
c at → cat

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
mat, s	2
mat, e	3
at, e	3
e, at	2

<i>word</i>	<i>count</i>
cat	4
mat	5
mat s	2
mat e	3
at e	3
e at	2

Merges (total 5):

a t → at  
m at → mat  
c at → cat  
mat e → mate

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
mat, s	2
mat, e	3
at, e	3
e, at	2

<i>word</i>	<i>count</i>
cat	4
mat	5
mat s	2
mate	3
at e	3
e at	2

Merges (total 5):

a t → at  
m at → mat  
c at → cat  
mat e → mate

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
mat, s	2
at, e	3
e, at	2

<i>word</i>	<i>count</i>
cat	4
mat	5
mat s	2
mate	3
at e	3
e at	2

Merges (total 5):

a t → at  
m at → mat  
c at → cat  
mat e → mate  
at e → ate

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

<i>pairs</i>	<i>count</i>
mat, s	2
at, e	3
e, at	2

<i>word</i>	<i>count</i>
cat	4
mat	5
mat s	2
mate	3
ate	3
e at	2

Merges (total 5):

a t → at  
m at → mat  
c at → cat  
mat e → mate  
at e → ate

# Byte Pair Encoding: Algorithm

1. Initialize the vocabulary, e.g. 100,000 words
2. Split each word into **characters**
3. Iterate:
  - a. Count the **frequency** of each **consecutive pair of characters**
  - b. **Merge** the **most frequent** tokens into a single one

We add the ## symbol to distinguish segmented words

Final vocabulary:

cat  
mat  
mat## s  
mate  
at  
e## at

<i>word</i>	<i>count</i>
cat	4
mat	5
mat s	2
mate	3
ate	3
e at	2

LIMIT REACHED

Merges (total 5):

a t → at  
m at → mat  
c at → cat  
mat e → mate  
at e → ate

# Subword Segmentation: Unigram LM

- Drawbacks of **BPE**:

- Ambiguous resulting vocabulary
- What if there is more than one way to separate a word?

ma | t

m | a | t

m | at

Final vocabulary:

m, ma, at, a, t

- Solution: Probabilistic Unigram Language Model ([Kudo, 2018](#))

- Predict the probability of a certain output given an initial state (LM objective)
- *Unigram LM*: only considers the probability of the next word

Choose the most *likely* subword instead of the most *frequent*

- Algorithm

1. Define a vocabulary size and a seed set (e.g. the BPE output!)
2. Find the probability of each subword (assume: subwords occur independently)
3. Calculate the loss of a subword in case it was removed
4. Keep the top X% subwords with the smallest loss (typical 80%)

repeat

Bonus!  
Proposed a **subword regularisation** to take into account multiple subword segmentations

# Subword Segmentation: WordPiece

- Originally introduced by [Schuster and Nakajima \(2012\)](#) used in BERT LM ([Devlin et al., 2018](#))
- In the middle between BPE and Unigram LM
  - **BPE**
    - ✓ Choose what to merge → based on frequency
    - ✓ Merging → based on frequency
  - **Unigram LM**
    - ✓ Choose what to merge → based on probability
    - ✓ Merging → based on probability
  - **WordPiece**
    - ✓ Choose what to merge → based on frequency
    - ✓ Merging → based on probability

Merge the pair of characters that will result in the largest likelihood increase

# Thank you !

Fenia [fenia.christopoulou@gmail.com](mailto:fenia.christopoulou@gmail.com)

Chryssa [chryssa.zrv@gmail.com](mailto:chryssa.zrv@gmail.com)