

Outline of the Tutorial

- Introduction to NLP and Information Extraction
- Introduction to GATE
- Social media analysis
- Text Analysis for Semantic Search
 - Semantic Search
 - Semantic Annotation
 - GATE MIMIR
- Example applications
 - ExoPatent, TNA Web Archive, BBC News, Envilod, Prospector, Political Futures Tracker, Brexit

1. NLP and Information Extraction

Oddly enough, people have successfully combined information and toast...



The weather-forecasting toaster



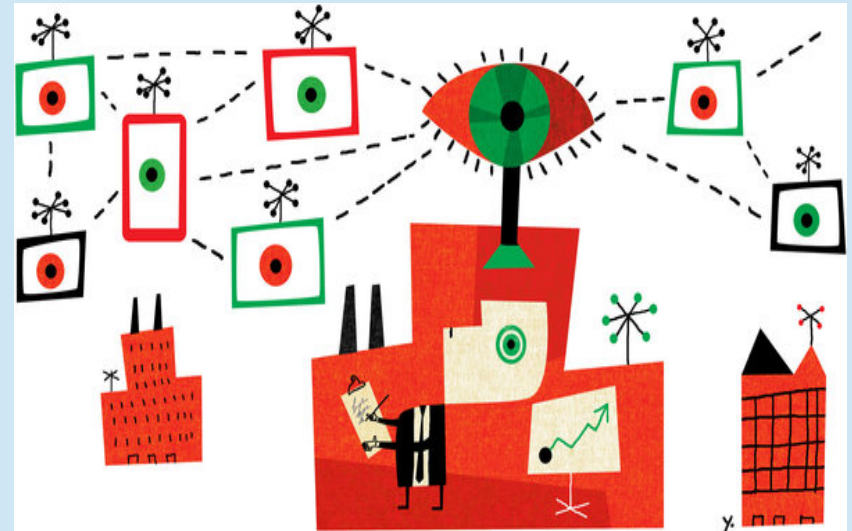
- This weather-forecasting toaster, connected to a phone point, was designed in 2001 by a PhD student
- It accessed the MetOffice website via a modem inside the toaster and translated the information into a 1, 2 or 3 for rain, cloud or sun
- The relevant symbol was then branded into the toast in the last few seconds of toasting

However, toast isn't actually a very good medium for getting your information...



Information overload or filter failure?

- We all know that there's masses of information available online these days, and it's constantly growing
- You often hear people talk about “information overload”
- But the real problem is not the amount of information, but our inability to filter it correctly
- Clay Shirky has an excellent talk on this topic
<http://bit.ly/oWJTNZ>



Big Data is not new!



Staff sorting 4M used tickets from #London Underground to analyse line use in 1939

NLP gives us a way to understand data

- **sort** the data to remove the rubbish from the interesting parts
- **extract** the relevant pieces of information
- **link** the extracted information to other sources of information (e.g. from DBpedia)
- **aggregate** the information according to potential new categories
- **query** the (aggregated) information
- **visualise** the results of the query

What is information extraction?

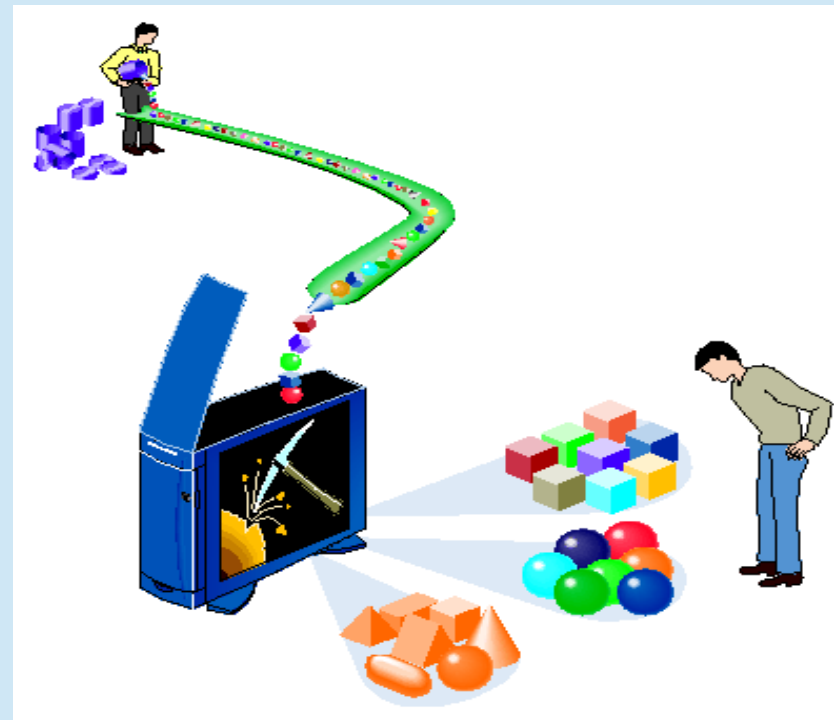
- The automatic discovery of new, previously unknown information, by automatically extracting information from different textual resources.
- A key element is the linking together of the extracted information to form new facts or new hypotheses to be explored further
- In IE, the goal is to discover previously unknown information, i.e. something that no one yet knows

IE is not Data Mining

Data mining is about using analytical techniques to find interesting patterns from large structured databases

Examples:

- using consumer purchasing patterns to predict which products to place close together on shelves in supermarkets
- analysing spending patterns on credit cards to detect fraudulent card use.



IE is not Web Search

- IE is also different from traditional web search or IR.
- In search, the user is typically looking for something that is already known and has been written by someone else.
- The problem lies in sifting through all the material that currently isn't relevant to your needs, in order to find the information that is.
- The solution often lies in better ways to ask the right question
- You can't ask Google to tell you about
 - all the speeches made by Tony Blair about foot and mouth disease while he was Prime Minister
 - all the documents in which a politician born in Sheffield is quoted as saying something about hospitals.

More about how we can do this will be revealed later...

**GATE Mimir:
Answering Questions
Google Can't**



Information Extraction Basics

- Entity recognition

is required for

- Relation extraction

which is required for

- Event recognition

which is required for

- Summarisation, answering questions, and other things

What is Entity Recognition?

- Entity Recognition is about recognising and classifying key Named Entities and terms in the text
- A **Named Entity** is a Person, Location, Organisation, Date etc.
- A **term** is a key concept or phrase that is representative of the text

Mitt Romney, the favorite to win the Republican nomination for president in 2012

Person Term Date

- Entities and terms may be described in different ways but refer to the same thing. We call this **co-reference**.

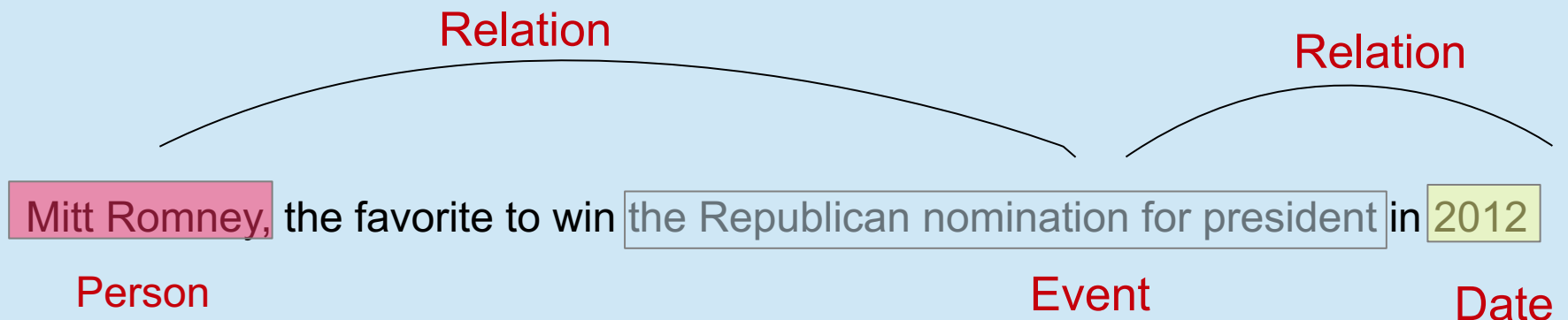
co-reference

The GOP tweeted that they had knocked on 75,000 doors in Ohio the day prior.

Organisation Location

What is Event Recognition?

- An event is an action or situation relevant to the domain expressed by some relation between entities or terms.
- It is always grounded in time, e.g. the performance of a band, an election, the death of a person



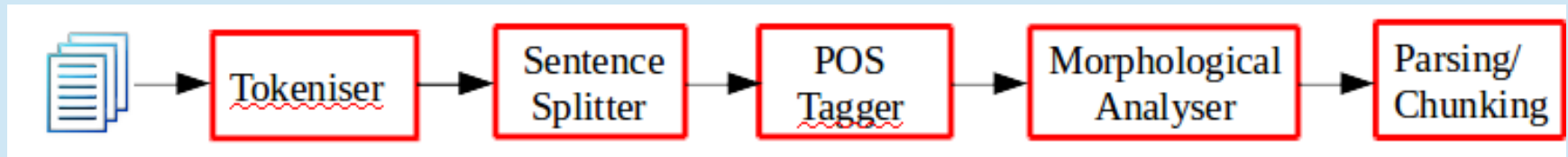
Why are Entities and Events Useful?

- They can help answer the “Big 5” journalism questions (who, what, when, where, why)
- They can be used to categorise the texts in different ways
 - look at all texts about Trump.
- They can be used as targets for opinion mining
 - find out what people think about Trump
 - When linked to an ontology and/or combined with other information, they can be used for reasoning about things not explicit in the text
 - seeing how opinions about different American presidents have changed over the years

NLP components for text mining

- A text mining system is usually built up from a number of different NLP components
- First, you need some Information Extraction tools to do the donkey work of getting all the relevant pieces of information and facts.
- Then you need some tools to apply the reasoning to the facts, e.g. opinion mining, information aggregation, semantic technologies, dynamics analysis
- GATE is an example of a tool for text mining which allows you to combine all the necessary NLP components

Low-level linguistic processing components



- These are needed to pre-process the text ready for the more complex IE tasks (NER, relations etc.)

Approaches to Information Extraction

Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- easier to understand results
- development could be very time consuming
- some changes may be hard to accommodate

Learning Systems

- use statistics or other machine learning
- developers do not need LE expertise
- requires large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus
- Can be unsupervised, but results are less good, and hard to adapt to a domain

GATE

The screenshot displays the GATE 2.1-alpha1 build 856 interface. The window title is "Gate 2.1-alpha1 build 856". The menu bar includes "File", "Options", "Tools", and "Help". The left sidebar shows a tree view with "Gate" at the top, followed by "Applications" (containing "ANNIE_0001E"), "Language Resources" (containing "corpus" and "newspaper text"), and "Processing Resources" (containing several ANNIE modules like "ANNIE Coreferencer_0", "ANNIE OrthoMatcher_", "ANNIE NE Transducer", "ANNIE POS Tagger_0", "ANNIE Sentence Split", "ANNIE Gazetteer_000", and "ANNIE English Token").

The main window shows a document titled "ANNIE_0001E" with the following text and annotations:

Threats to the resumption of the Northern Ireland peace talks receded today after a British cabinet minister entered the huge Maze prison near Belfast and pressed Protestant guerrillas held there to support continuing the discussions.

Northern Ireland Secretary Marjorie Mowlam sat down with members of two outlawed Protestant paramilitary groups and delivered a 14-point statement on why they should reverse a vote they took last weekend to condemn the talks. That vote had thrown the talks' future into question.

After she left, the prisoners did what she asked. The political party that speaks for them at the negotiating table, the Ulster Democratic Party, announced it was no longer considering boycotting the talks, which are set to resume Monday. Another party affiliated with imprisoned Protestant guerrillas, the Progressive Unionist Party, said it would decide on Sunday whether to attend.

The all-party talks, chaired by former U.S. senator George J. Mitchell (D-Maine), seek a political solution in Northern Ireland between Protestants, most of whom want to remain part of Britain, and Catholics, who want greater political rights, including, for some, political union with the Republic of Ireland to the south.

Throughout the conflict, the British government has held to the line that it talks to people who renounce violence, not to killers. To many people in Britain, it seemed today that Mowlam was being summoned by men convicted of crimes that include murder and arson.

"We are very unhappy about it," said Glyn Roberts, development officer for a Northern Ireland peace group called Families Against Intimidation and Terror. Mowlam spoke directly with terrorists, he said, "which many victims felt was grossly insulting."

The right sidebar shows a list of "Default annotations" and "Original markups annotations". The "Default annotations" list includes: Date, FirstPerson, JobTitle, Location (checked), Lookup, Organization (checked), Person (checked), Sentence, SpaceToken, Split, Title, Token, and Unknown. The "Original markups annotations" list includes: DOC, DOCNO, DOCTYPE, HEADER, and TEXT.

At the bottom of the window, there are buttons for "Annotations Editor" and "Features Editor".

What is GATE?

GATE is an NLP toolkit developed at the University of Sheffield over the last 20 years

It includes:

- **components** for language processing, e.g. parsers, machine learning tools, stemmers, IR tools, IE components for various languages...
- tools for **visualising** and **manipulating** text, annotations, ontologies, parse trees, etc.
- **various information extraction** tools
- **evaluation** and **benchmarking** tools

GATE components

- **Language Resources (LRs)**, e.g. lexicons, corpora, ontologies
- **Processing Resources (PRs)**, e.g. parsers, generators, taggers
- **Visual Resources (VRs)**, i.e. visualisation and editing components
- Algorithms are separated from the data, which means:
 - the two can be developed independently by users with different expertise.
 - alternative resources of one type can be used without affecting the other, e.g. a different visual resource can be used with the same language resource

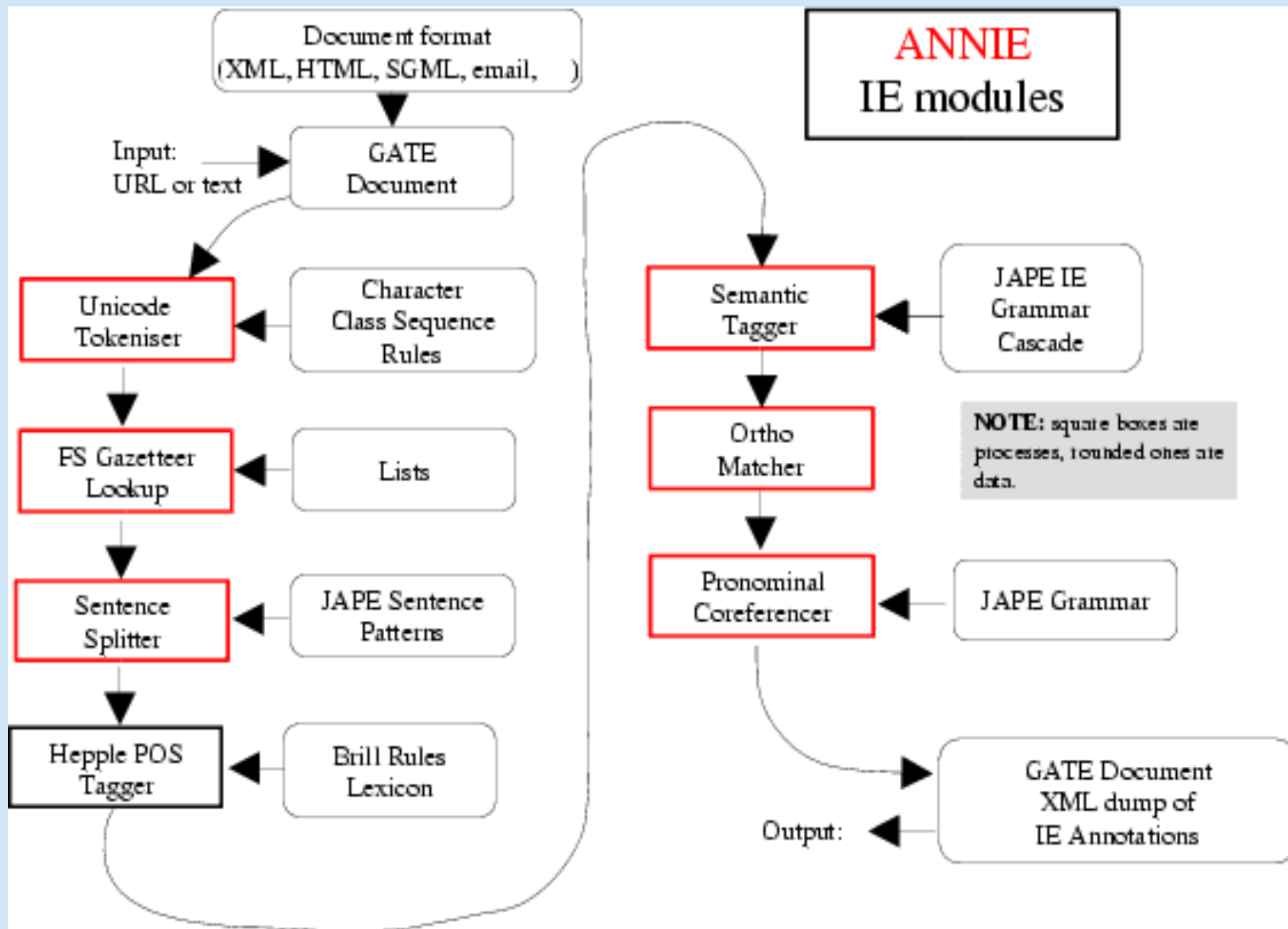
ANNIE

- **ANNIE** is GATE's rule-based IE system
- It uses the language engineering approach (though we also have tools in GATE for ML)
- Distributed as part of GATE
- Uses a finite-state pattern-action rule language, JAPE
- ANNIE contains a reusable and easily extendable set of components:
 - generic pre-processing components for tokenisation, sentence splitting etc
 - components for performing NE on general open domain text

What's in ANNIE?

- The ANNIE application contains a set of core PRs:
 - Tokeniser
 - Sentence Splitter
 - POS tagger
 - Gazetteers
 - Named entity tagger (JAPE transducer)
 - Orthomatcher (orthographic coreference)
- There are also other PRs available in the ANNIE plugin, which are not used in the default application, but can be added if necessary
 - NP and VP chunker, morphological analysis

ANNIE Modules



ANNIE English Tokeniser

- Tokenisation is based on Unicode classes
- It chops a piece of text into words and spaces, and also has features for numbers, punctuation, symbols, capitalisation etc.
- It converts constructs involving apostrophes into more sensible combinations
 - don't → do + n't
 - you've → you + 've
- Other tokenisers might have different definitions of Token
- Other languages might need different tokenisers, e.g. Chinese

Document with Tokens

The screenshot displays a text analysis tool interface. At the top, there are several tabs: "Annotation Sets", "Annotations List", "Annotations Stack", "Class", "Co-reference Editor", "Instance", and "Text". The "Text" tab is active, showing a document with several lines of text. The first line, "Union Appeals For Talks To End BA Strike", is highlighted in green. Below it, there are navigation links like "Skip to navigation", "Skip to content", "Home", "Contact Us", and "News Search". Further down, there are more lines of text, including "Union Appeals For Talks To End BA Strike" and "March 22, 2010". The main body of text is: "Union leaders on Sunday called for talks with British Airways bosses to end strike action by cabin crew that has led to the cancellation of hundreds of flights and disrupted travel plans for thousands of passengers." Below the text, there is a table with two columns: "Type" and "Features". The table lists five tokens with their respective features. On the right side of the interface, there is a vertical list of classification categories, each with a checkbox and a colored background. The "Token" category is checked and highlighted in green. Other categories include Date, FirstPerson, JobTitle, Location, Lookup, Money, Organization, Percent, Person, Sentence, SpaceToken, Split, Title, and Unknown. At the bottom of this list, there is a section for "Original markups".

Annotation Sets Annotations List Annotations Stack Class Co-reference Editor Instance Text

Union Appeals For Talks To End BA Strike

Skip to navigation | Skip to content |
Home | Contact Us | News Search;
HubPage
Airwise News
Airport Guide
Airwise Travel
Search
Union Appeals For Talks To End BA Strike
March 22, 2010

Union leaders on Sunday called for talks with British Airways bosses to end strike action by cabin crew that has led to the cancellation of hundreds of flights and disrupted travel plans for thousands of passengers.

Type	Features
Token	{ category=NNP, kind=word, length=5, orth=upperInitial, string=Union}
Token	{ category=NNPS, kind=word, length=7, orth=upperInitial, string=Appeals}
Token	{ category=IN, kind=word, length=3, orth=upperInitial, string=For}
Token	{ category=NNS, kind=word, length=5, orth=upperInitial, string=Talks}
Token	{ category=TO, kind=word, length=2, orth=upperInitial, string=To}

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Percent
- Person
- Sentence
- SpaceToken
- Split
- Title
- Token
- Unknown
- ▶ Original markups

Sentence Splitter

- The default splitter finds sentences based on Tokens
- Main problem is to find whether a full stop or line break denotes the end of a sentence or something else
- Also what to do with things like bullet points
- Uses a gazetteer of abbreviations etc. and a set of rules to find sentence delimiters
- There are a few sentence splitter variants available which work in slightly different ways

Document with Sentences

Annotation Sets Annotations List Annotations Stack Class Co-reference Editor Instance Text

the opposition conservatives, ahead in opinion polls, have been turning up the pressure on Labour over its links to Unite, saying the government had failed to take action quickly enough because it did not want to alienate its financial backers.

"We deplore the strike, and the prime minister and the transport secretary have said that absolutely clearly," Foreign Secretary David Miliband told Sky News.

"The way to resolve these disputes is through negotiation, it is damaging for the company, it is damaging for the crews and it is damaging for the country."

The dispute arose because BA, which has 12,000 cabin crew, wants to save an annual GBP£62.5 million pounds (USD\$95 million) to help cope with a fall in demand, volatile fuel prices and increased competition from low-cost carriers.

A spokesman said there was no estimate yet as to how much the industrial action would cost the company.

Type	Features
Sentence {}	
Sentence {}	
Sentence {}	
Sentence {}	
Sentence {}	

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Percent
- Person
- Sentence
- SpaceToken
- Split
- Title
- Token
- Unknown
- Original markups

POS tagger

- Finds the part of speech for every word, e.g. noun, verb etc.
- Trained on WSJ, uses Penn Treebank tagset
- Other taggers may be trained on different corpora and have different tags
- Default ruleset and lexicon can be modified manually (with a little deciphering)
- Adds category feature to Token annotations
- Requires Tokeniser and Sentence Splitter to be run first

Morphological analyser

- Finds the root form of a word (singular noun, base form of verb)
- This is not the same as stemming (the root is a real word, a stem might not be)
- e.g. educating: root = educate; stem = educat
- Does not perform derivational analysis (will not conflate verbs and nouns)
- Not an integral part of ANNIE, but can be found in the Tools plugin as an “added extra”
- Flex based rules: can be modified by the user
- Generates “root” feature on existing Token annotations
- Requires tokenisation and POS tagging

Gazetteers

- Gazetteers are plain text files containing lists of names (e.g. rivers, cities, people, ...)
- The lists are compiled into Finite State Machines
- Each gazetteer has an index file listing all the lists, plus features of each list that help to categorise the lists
- Each entry can also have specific features and values, e.g. a DBpedia link
- Lists can be modified either internally using the Gazetteer Editor, or externally in your favourite editor
- Gazetteers generate Lookup annotations with relevant features corresponding to the list matched
- ANNIE gazetteer has about 60,000 entries arranged in 80 lists

Gazetteer editor

File Options Tools Help

Messages 1269258352.html... ANNIE ANNIE Gazetteer

airport.lst New List

List name	Major	Minor	Language
charities.lst	organization		
city.lst	location	city	
city_cap.lst	location	city	
company.lst	organization	company	
company_cap.lst	organization	company	
country.lst	location	country	
country_abbrev.lst	location	country_abbrev	
country_adj.lst	country_adj		
country_cap.lst	location	country	
currency_prefix.lst	currency_unit	pre_amount	
currency_unit.lst	currency_unit	post_amount	
date_key.lst	date_key		
date_unit.lst	date_unit		
day.lst	date	day	
day_cap.lst	date	day	
department.lst	organization	departmen	

New Entry Add Cols

Value
Aaccra
Aalborg
Aarhus
Ababa
Abadan
Abakan
Aberdeen
Abha
Abi Dhabi
Abidjan
Abilene
Abu
Abu Dhabi
Abuja
Acapulco

Filter: 1993 entries

Gaze Initialisation Parameters Gazetteer Editor

definition file
entries

entries for selected list

Named Entity Recognition

- Gazetteers can be used to find terms that suggest entities
- However, the entries can often be ambiguous
 - “May Jones” vs “May 2010” vs “May I help you?”
 - “Mr Parkinson” vs “Parkinson's Disease”
 - “General Motors” vs. “General Smith”
- Handcrafted grammars are used to define patterns over the Lookups and other annotations
- These patterns can help disambiguate, and they can combine different annotations, e.g. dates can be comprised of day + number + month

Named Entity Grammars

- Hand-coded rules written in JAPE applied to annotations to identify NEs
- Phases run sequentially and constitute a cascade of FSTs over annotations
- Because phases are sequential, annotations can be built up over a period of phases, as new information is gleaned
- Standard named entities: persons, locations, organisations, dates, addresses, money
- Basic NE grammars can be adapted for new applications, domains and languages

Example of a JAPE pattern-action rule

Rule: PersonName

```
(  
  {Lookup.majorType == firstname}  
  {Token.category == NNP}  
):tag
```

-->

```
:tag.Person = {kind = fullName}
```

Look for an entry in a
gazetter list of first
names

followed by a proper noun

create an annotation of type "Person"

give the annotation the feature
kind and value "fullName"

JAPE rules are usually a bit more complex

- They can make use of any annotations already created, e.g.
 - strings of text
 - POS categories
 - gazetteer lookup
 - root forms of words
 - document metadata, e.g. HTML tags
 - annotations “in context”
- They can use regular expression operators, including negative constraints, “contains”, “within” etc.
- And the RHS of a rule can contain any Java code you like

Document with NEs

The screenshot displays a document editor window with a toolbar at the top containing tabs for 'Annotation Sets', 'Annotations List', 'Annotations Stack', 'Co-reference Editor', 'Text', and a search icon. The main text area contains several paragraphs with various words highlighted in different colors, indicating named entities. A search tool is overlaid on the text, showing a dropdown menu with 'Date' selected, and a table with columns for 'kind' and 'year'. The 'kind' dropdown is set to 'kind' and the 'year' dropdown is set to 'year'. A button labeled 'Open Search & Annotate tool' is at the bottom of the search tool. On the right side, a sidebar lists various entity types under the heading 'Entities'. The 'Date' entity type is checked, and other types like 'Location', 'Organization', and 'Person' are also checked. The 'New' button is at the bottom right of the sidebar.

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

In so far as a political party in the **United States** can "decide" anything, the party decided not to have the fight it needed to have between reality-based Republicans and the other kind. And so but in disguised form. The

Given the state of our political a theory like this. There is key phrases.

1.) When I say "reality-based Republicans" I mean those who recognize the danger in trying to make descriptions of the world conform to their wishes. By the "other kind" I mean those who don't. Or: members of the **Republican** coalition who exhibit certain behaviors **F.A. Hayek** wrote about in **1960**. This quotation was dug up by Chis Mooney , author of The **Republican** War on Science . It is from Hayek's essay, "Why I am Not a Conservative."

Personally, I find that the most objectionable feature of the conservative attitude is its

English
Entities
 Date
 EmbeddedHead1
 EmbeddedHead2
 EnglishContent
 FirstPerson
 GermanContent
 Head
 JobTitle
 Location
 Lookup
 MOD
 MultiWord
 Organization
 Person
 Sentence
 SingleWord
 SpaceToken

Date
kind year
Open Search & Annotate tool

New

Document Editor Initialisation Parameters

Using co-reference

- Different expressions may refer to the same entity
- Orthographic co-reference module (orthomatcher) matches proper names and their variants in a document
- [Mr Smith] and [John Smith] will be matched as the same person
- [International Business Machines Ltd.] will match [IBM]

Co-reference in GATE

The screenshot displays the GATE Co-reference Editor interface. The main window shows a document with several paragraphs of text. The first paragraph is: "Completion of the **National Air Traffic Services** deal comes at a critical time for the government as it tries to push through the PPP for the **London** Underground." The second paragraph is: "The sale to a strategic investor of a 46 per cent stake in **Nats** is the first time in Europe that management control of en route air traffic services has passed into private hands." The third paragraph is: "It has been carried out despite a pledge by Labour before the 1997 general election that **UK** air was 'not for sale.'" The fourth paragraph is: "Under the terms of the deal, which was approved by the European competition authorities in May, the government has retained a 49 per cent stake and a golden share, while a 5 per cent stake is to be allocated to **Nats** 5,700 staff." The right-hand pane shows the "Co-reference Editor" settings. The "Sets" dropdown is set to "Default". The "Types" dropdown is set to "Organization", and a "Show" button is visible. Below this, the "Co-reference Data" section shows a list of entities under the "Default" set, each with a checked checkbox and a colored highlight: "National Air Traffic Services" (red), "Airline Group" (green), "UK" (yellow), "London" (purple), and "March" (pink). Red arrows point from the highlighted text in the document to the corresponding entries in the list.

Annotation Sets Annotations List Annotations Stack Class **Co-reference Editor** Instance T

Sets : Default

Types : Organization Show

Co-reference Data
♀ Default

- National Air Traffic Services
- Airline Group
- UK
- London
- March

Document Editor Initialisation Parameters

Other NLP Toolkits

- UIMA
- OpenCalais
- Lingpipe
- OpenNLP
- Stanford Tools

- All integrated into GATE as plugins

3. Analysing Social Media

I AM HIGHLY INTELLIGENT.

I KNOW WORDS. I KNOW THE BEST WORDS.

People don't write "properly" on social media

- Grundman:politics makes #climatechange scientific issue,people don't like knowitall rational voice tellin em wat 2do
- Want to solve the problem of #ClimateChange? Just #vote for a #politician! Poof! Problem gone! #sarcasm #TVP #99%
- Human Caused #ClimateChange is a Monumental Scam!
<http://www.youtube.com/watch?v=LiX792kNQeE> ... F**k yes!!
Lying to us like MOFO's Tax The Air We Breath! F**k Them!
- The last people I will listen2 about guns r those that know nothing about them&politicians who live in states w/strictest gun laws [#cali](#) [#ny](#)

Linguistic challenges of social media

- **Language**
 - Problem: typically exhibits very different language style
 - Solution: train specific language processing components
- **Relevance**
 - Problem: topics and comments can rapidly diverge.
 - Solution: train a classifier or use clustering techniques
- **Lack of context**
 - Problem: hard to disambiguate entities
 - Solution: data aggregation, metadata, entity linking

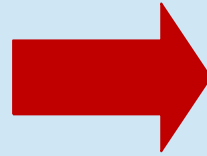
Challenges for NLP

- Noisy language: unusual punctuation, capitalisation, spelling, use of slang, sarcasm etc.
- Terse nature of microposts such as tweets
- Use of hashtags, @mentions etc causes problems for tokenisation #thisistricky
- Lack of context gives rise to ambiguities
- NER performs poorly on microposts, mainly because of linguistic pre-processing failure
 - Performance of standard IE tools decreases from ~90% to ~40% when run on tweets rather than news articles

NLP Pre-Processing Pipeline



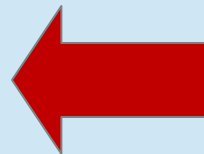
Text



Language ID

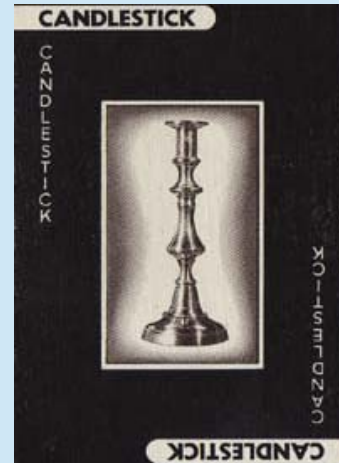
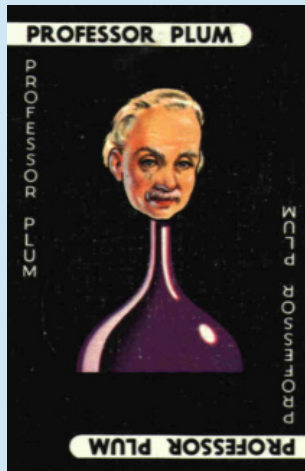
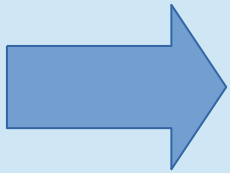


Tokenisation

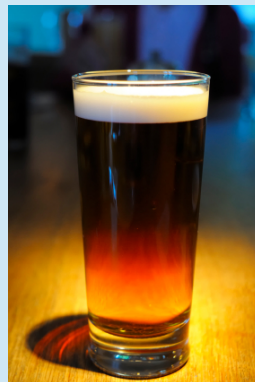
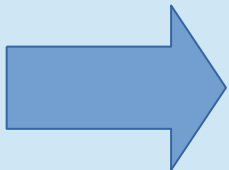


POS tagging

Named Entity Recognition and Linking



NER (Professor Plum)

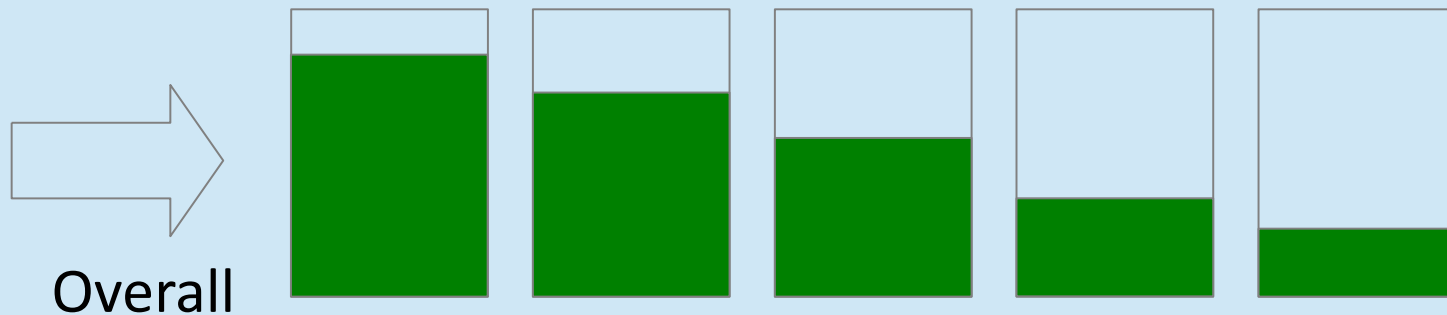
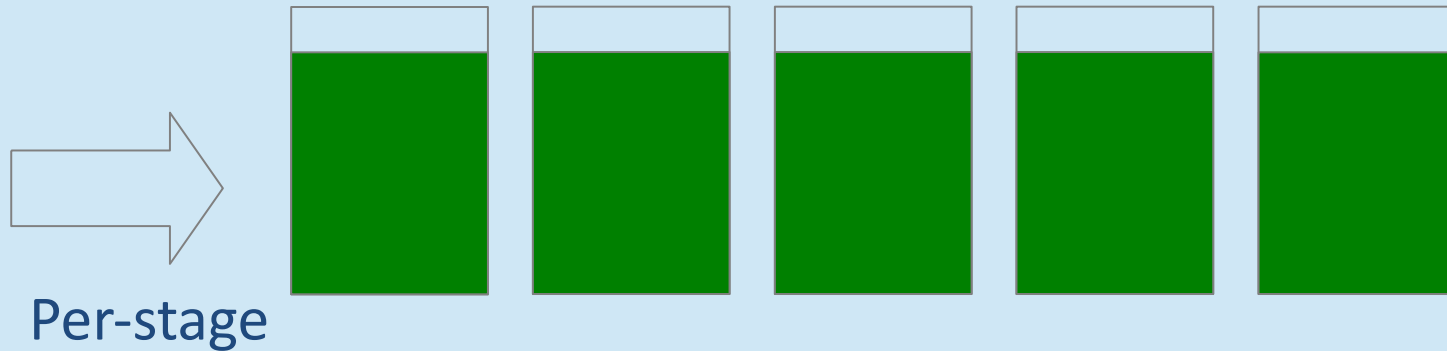


dbpedia.org/resource/.....
Michael_Jackson
Michael_Jackson_(writer)

Entity Linking

Pipelines for tweets

Errors have a cumulative effect



Good performance is important at each stage

Language Identification

The Jan. 21 show started with the unveiling of an impressive three-story castle from which Gaga emerges. The band members were in various portals, separated from each other for most of the show. For the next 2 hours and 15 minutes, Lady Gaga repeatedly stormed the moveable castle, turning it into her own gothic Barbie Dreamhouse .

News wire

LADY GAGA IS BETTER THE 5th TIME OH BABY(:

je bent Jacques Cousteau niet die een nieuwe soort heeft ontdekt, het is duidelijk, ze bedekken hun gezicht. Get over it

Twitter

I'm at 地铁望京站 Subway Wangjing (Beijing)

<http://t.co/KxHzYm00>

Tokenisation is only 80% accurate on tweets

Improper grammar, e.g. apostrophe usage:

doesn't → does n't

doesnt → doesn't

Smileys and emoticons: loss of information (e.g. sentiment)

I <3 you

This piece ;,,(so emotional

Punctuation for emphasis

*HUGS YOU**KISSES YOU* → * HUGS YOU **KISSES YOU *

Words run together / skip

I wonde rif Tsubasa is okay..

We need tools for hashtag analysis

- Hashtags need unravelling and disambiguating:
 - #nowthatcherisdead
 - #powergenitalia
 - #lesbocages
 - #molestationnursery
 - #teacherstalking
 - #therapist



Test your social media skills

What do these hashtags mean?

- #kktny
- #fomo
- #jomo
- #ootd
- #wcw

Hashtag Hijacking

Hashtags are not always used in the same way, or for their original purpose



Mark Tyrrell UKIP @MarkTyrrellUKIP · Jun 20

I'm very happy that Farage called out Barroso on his climate change lies in the parliament. **#WhyImVotingUKIP**



7



5



Why I'm Voting UKIP @WhyImVotingUKIP · May 22

#WhyImVotingUkip because I heard they were going to take Britain out of Europe, so I'm hoping they move us somewhere hot like the Caribbean.

♥ **SEX**
IS BETTER
— THAN —
TECH 📱

MARCH 29 8.30–9.30PM
#TURN**OFF** TOTURN**ON**

love sex
durex

Tweet Normalisation

- “RT @Bthompson WRITEZ: @libbyabrego honored?! Everybody knows the libster is nice with it...lol...(thankkkks a bunch;)”
- OMG! I’m so guilty!!! Sprained biibii’s leg! ARGHHHHH!!!!!!
- For some components to work well (POS tagger, parser), we need normalisation
- BUT uppercasing, and repetition often convey strong sentiment
- Other forms of “misspelling” might indicate information about the author
- First challenge: separate out-of-vocabulary and in-vocabulary
- Second challenge: fix mis-spelled IV words (e.g. Levenshtein edit distance)
- The ZOMG phenomenon

Part-of-speech tagging

- Similar issues as for normalisation – we don't have big datasets to train on
- Label unlabelled data with multiple taggers and accept tweets where tagger votes never conflict
- Model words using Brown clustering and word representations (Turian 2010)

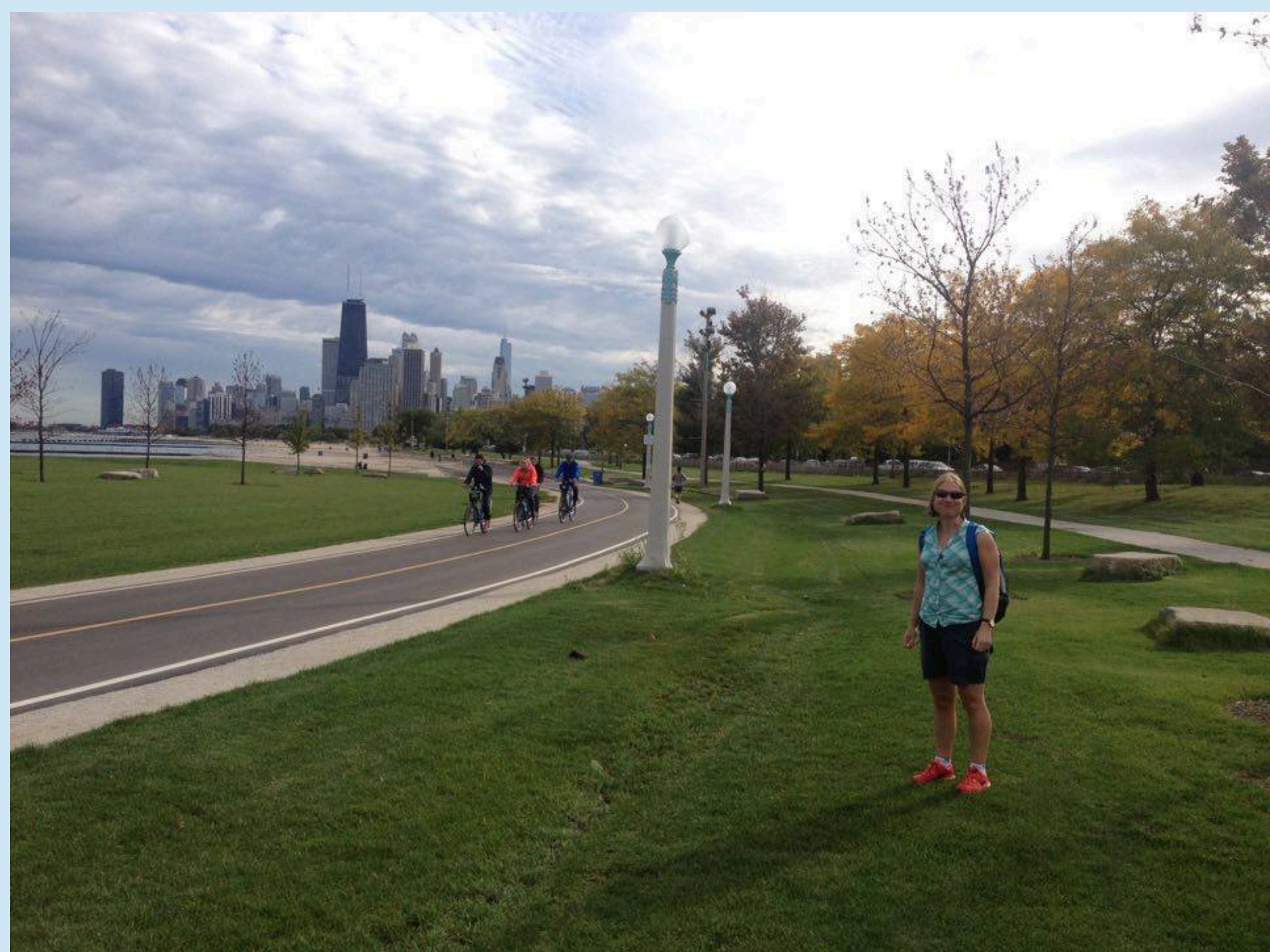
2m, 2ma, 2mar, 2mara, 2maro, 2marrow, 2mor, 2mora, 2moro, 2morow, 2morr, 2morro, 2morrow, 2moz, 2mr, 2mro, 2mrrw, 2mrw, 2mw, tmmrw, tmo, tmoro, tmorrow, tmoz, tmr, tmro, tmrow, tmrrow, tmrrw, tmrw, tmrww, tmw, tomaro, tomarow, tomarro, tomarrow, tomm, tommarow, tommarrow, tommoro, tommorrow, tommorrow, tommorw, tommrow, tomo, tomolo, tomoro, tomorow, tomorro, tomorrw, tomoz, tomrw, tomz

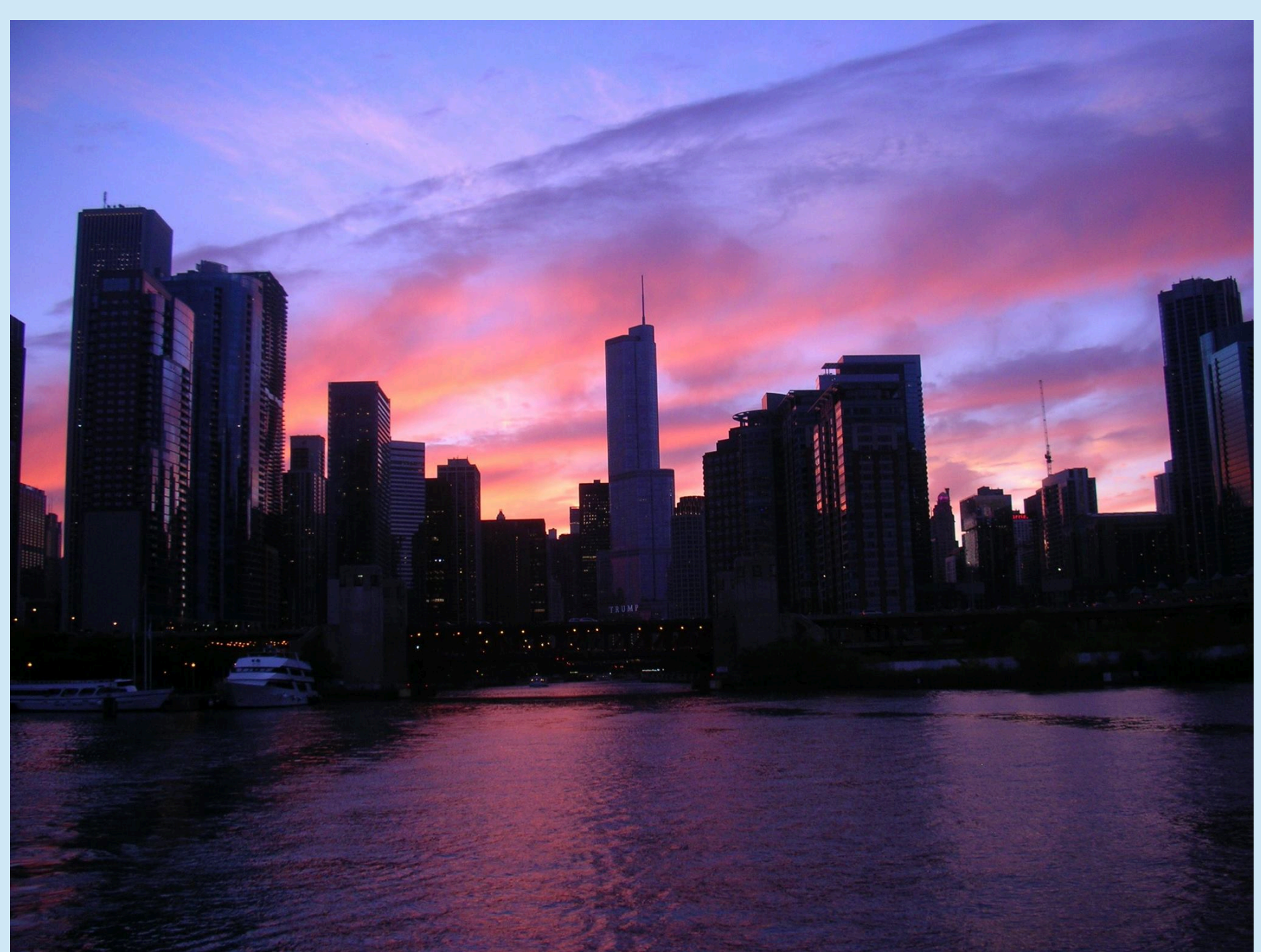
Named entities: lack of context causes ambiguity

Branching out from Lincoln park after dark ... Hello Russian Navy, it's like the same thing but with glitter!



??





Getting the NEs right is crucial

Branching out from Lincoln park after dark ... Hello Russian Navy
Navy, it's like the same thing but with glitter!



How do we deal with all these problematic tweets?

- Typical NLP pipeline means that degraded performance has a knock-on effect along the chain
- Short sentences confuse language identification tools
- Linguistic processing tools have to be adapted to the domain



- Retraining individual components on large volumes of data
- Adaptation of techniques from e.g. SMS analysis
- Development of new Twitter-specific tools (e.g. GATE's TwitIE)
- But....lack of standards, easily accessible data, common evaluation etc. are holding back development

4. Text Analytics for Semantic Search

Semantic Queries in Google

[Paris convention and visitors office - Official website - Paris tourism](#)

en.parisinfo.com/

Paris convention and visitors office diffuses all information to organise your stay or your trip in **Paris**: hotels and loadings, museums, monuments, going out, ...

[Our welcome centres](#) - [Paris Map](#) - [Transports and ...](#) - [Getting around](#) - [Book online](#)

[Paris - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Paris

Coordinates: 48°51′24″N 2°21′03″E﻿ / ﻿48.8567°N 2.3508°E﻿ / 48.8567; 2.3508. **Paris** is the capital and largest city of France. It is situated on the river ...

[List of tourist attractions in Paris](#) - [History of Paris](#) - [Demographics of Paris](#) - [Portal](#)

[Paris.com - Paris Travel Guide and hotel accommodation](#)

www.paris.com/

Paris.com : **Paris**, France tourist services offering hotel accommodation, holiday apartments. We guide you to the best **Paris** city tours and things to do!

[News for paris](#)



[Paris women finally allowed to wear trousers](#)

[BBC News](#) - 21 minutes ago

The French government overturns a 200-year-old ban on women wearing trousers in the capital, **Paris**, dating from November 1800.

[Skirts rule lifted: Centuries-old ban on women wearing trousers in Paris is finally axed](#)

[Mirror.co.uk](#) - 3 hours ago

[Women in Paris finally allowed to wear trousers](#)

[Telegraph.co.uk](#) - 1 day ago

[Paris | Travel | The Guardian](#)

www.guardian.co.uk/travel/paris

Latest news and comment on **Paris** from guardian.co.uk.



Paris

Paris is the capital and largest city of France. It is situated on the river Seine, in northern France, at the heart of the Île-de-France region. The city of Paris, within its administrative limits, has a population of about 2,230,000. [Wikipedia](#)

Population: 2,234,105 (2009)

Area: 105.4 km²

Weather: 8°C, Wind SW at 10 mph (16 km/h), 71% Humidity

Local time: Monday 23:12

Points of interest



Eiffel Tower



Louvre




Disneyland Resort Paris


Searching for Things, Not Strings

- 500 million entities that Google “knows” about
- Used to provide more accurate search results

See results about



[University of Cambridge](#)
The University of Cambridge is a public research university ...



[Cambridge](#)
The city of Cambridge is a university town and the administrative ...

- Summaries of information about the entity being searched



Anthony Blair

Anthony Charles Lynton Blair is a British Labour Party politician who served as the Prime Minister of the United Kingdom from 1997 to 2007. [Wikipedia](#)

Born: May 6, 1953 (age 59), [Edinburgh](#)

Full name: Anthony Charles Lynton Blair

Parents: [Hazel Corscadden](#), [Leo Blair](#)

Siblings: [William J. L. Blair](#)

Children: [Euan Blair](#), [Kathryn Blair](#), [Nicky Blair](#), [Leo Blair](#)

Education: [St John's College, Oxford \(1976\)](#), [Fettes College](#), [Chorister School](#), [University of Oxford](#)

People also search for



Gordon Brown

David Cameron

Margaret Thatcher

John Major

Facebook Graph Search

Current **Tesco** employees who like **Horses**

Home [Icons]

[Profile Picture]

Customer Service Assistant at Tesco

Likes Horses and Dogs

Studied [Name] at [Name]

Lives in Liverpool

Listens to [Name]

[Add Friend](#) [Message](#) [Search](#)

[Profile Picture]

Works at TESCO

Likes Horses

Studied [Name] at Uni. Wolverhampton

Lives in [Name]

Listens to [Name]

[Add Friend](#) [Message](#) [Search](#)

[Profile Picture]

Works at TESCO

Likes Horses

Studied at [Name]

Lives in [Name]

Listens to [Name]

[Add Friend](#) [Message](#) [Search](#)

[Profile Picture]

Works at Tesco

Likes Horses

Studied at [Name]

Lives in London, United Kingdom

4 followers

[Add Friend](#) [Follow](#) [Message](#) [Search](#)

More Than 100 People [View Grid](#)

REFINE THIS SEARCH [Dropdown]

Gender [Add...](#)

Relationship [Add...](#)

Current Employer [Tesco](#) [Add](#)

[Position...](#)

[Employer Location...](#)

[Time Period...](#)

Current City [Add...](#)

Hometown [Add...](#)






School [Add...](#)

Friendship [Add...](#)

Likes [Horses](#) [Add](#)

... [SEE MORE](#)

EXTEND THIS SEARCH [Dropdown]

[More pages they like](#)

[Photos of these people](#)

[These people's friends](#)

... [SEE MORE](#)

Semantic Enrichment

- Textual mentions aren't actually that useful in isolation
 - knowing that something is a “Person” isn't very helpful
 - knowing which real-life Person it refers to can be very useful
- Disambiguating mentions against an ontology provides extra context
- This is where **semantic enrichment** comes in
- The end product is a set of textual mentions linked to an ontology, otherwise known as **semantic annotations**
- Annotations on their own can be useful but they can also
 - be used to generate corpus level statistics
 - be used for further ontology population
 - form the basis of summaries
 - be indexed to provide semantic search

Automatic Semantic Enrichment

- Use Text Mining, e.g.
 - Information Extraction – recognise names of people, organisations, locations, dates, references, etc.
 - Term recognition – identify domain-specific terms
- Automatically extend article metadata to improve search quality

Mining medical records

- Medical records contain a large amount of unstructured text
 - letters between hospitals and GPs
 - discharge summaries
- These documents might contain information not recorded elsewhere
 - it turns out doctors don't like forms!
 - often information-specific fields are ignored, with everything put in the free text area

Medical Records at SLAM

- NIHR Biomedical Research Centre at the South London and Maudsley Hospital are using our text mining tools in a number of their studies
- Developed applications to extract:
 - the results of mental state tests, and the date the test was administered
 - education level (high school, university, etc.)
 - smoking status
 - medication history
- They have even had promising results predicting suicides

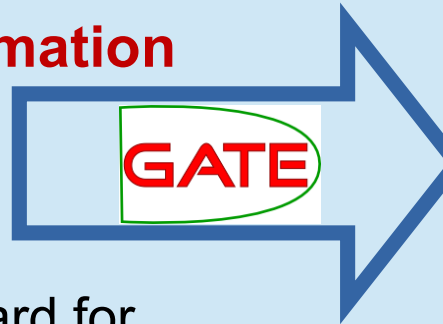
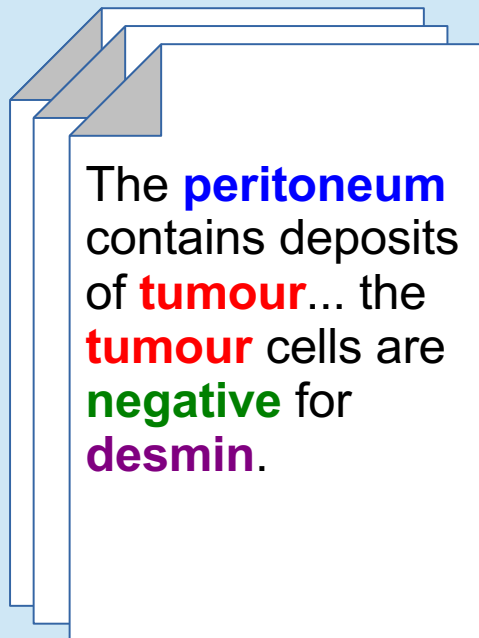
30% of medical information is structured

Hb	WBC	RB C	Plt
14.1	5.1	4.5	210
10.2	36.6	5.0	420
13.4	10.1	5.1	180
12.3	8.3	4.6	340

Easy to search and manipulate with computers

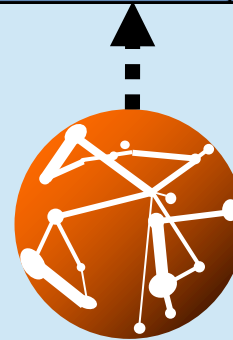
GATE structures the free text portion of the medical record, and makes it available for research, management, and clinical care

70% of medical information is in free text



Test		Result
desmin	has finding	negative
Condition		Locus
tumour	has location	peritoneum
...

Hard for computers to understand. Ambiguous, nuanced, and complexity



GATE can link the text to external knowledge, terminologies and coding schemes, for intelligent search and analysis

As used by several UK hospitals and world-leading medical record software vendors

Cancer Research: can GATE cure cancer?

- Genome Wide Association Studies (GWAS) aim to investigate genetic variants across the whole genome
 - With enough cases and controls, this allows them to state that a given SNP (Single Nucleotide Polymorphism) is related to a given disease.
 - A single study can be very expensive in both time and money to collect the required samples.
- Can we reduce the costs by analysing published articles to generate prior probabilities for each SNP?

COMBATING CANCER WITH BAKING SODA

Did you know...that baking soda has been shown to fight cancer, stave off colds and flu, and even treat radiation poisoning...all for just pennies a day?

Dr. Simonchini an oncologist in Rome originally made the connection between fungal infections and cancer proliferation.

He realized that when a tumor was flushed with baking soda (which is anti-fungal), it shrank and completely disappeared within days.



www.undergroundhealthreporter.com

Is GATE better than baking soda?

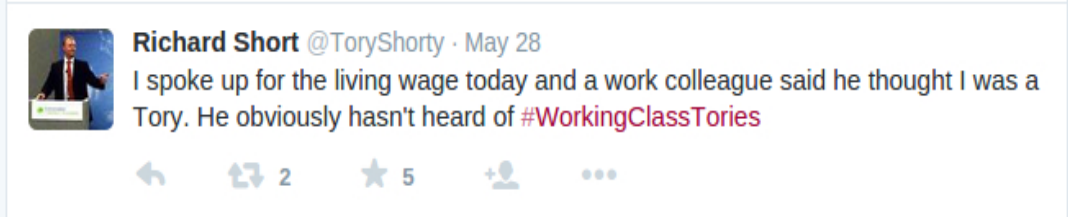
- In conjunction with IARC (International Agency for Research on Cancer, part of the WHO) we developed a text analysis approach to mine PubMed
- We showed retrospectively that our approach would have saved over a year's worth of work and more than 1.5 million Euros
- We completed a new study which found a new cause for oral cancer
 - Oral cancer is rare enough that traditional methods would have failed to find enough cases to make the study plausible

Government Web Archive

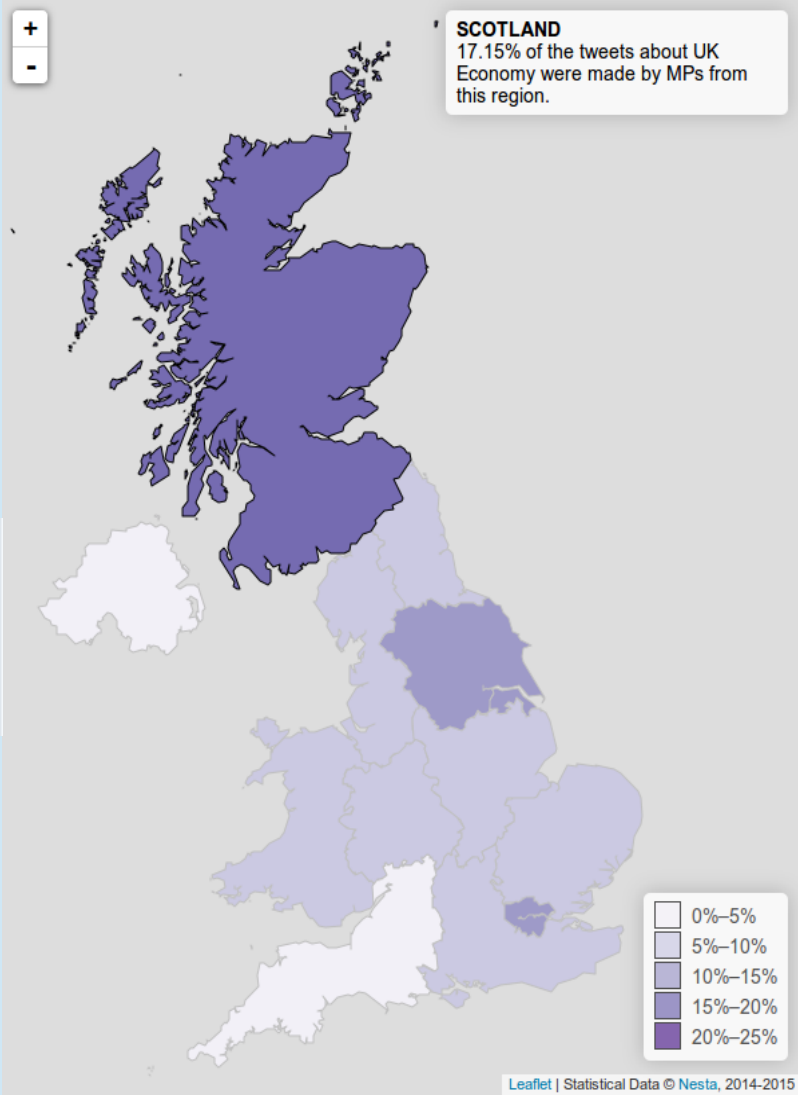
- We developed a semantic annotation application to process every crawled page in the archive.
- Entities annotated included; people, companies, locations, government departments, ministerial positions, social documents, dates, money....
- Where possible, annotations were linked to an ontology which
 - was based on DBpedia
 - was extended with UK government-specific concepts
 - included the modelling of the evolution of government
- Annotations were indexed to allow for complex semantic querying of the collection

Political Futures Tracker

Where in the UK did Conservative MPs tweet more about the economy?



Richard Short @ToryShorty · May 28
I spoke up for the living wage today and a work colleague said he thought I was a Tory. He obviously hasn't heard of [#WorkingClassTories](#)



5.1 Semantic Annotation

Why ontologies for semantic search?

- **Semantic annotation:** rather than just annotating the word “Cambridge” as a location, link it to an ontology instance
 - Differentiate between *Cambridge, UK* and *Cambridge, Mass.*
- **Semantic search via reasoning**
 - So we can infer that this document mentions a city in Europe.
 - Ontologies tell us that this particular Cambridge is part of the country called the UK, which is part of the continent Europe.
- **Knowledge source**
 - If I want to annotate *strikes* in baseball reports, the ontology will tell me that a *strike* involves a *batter* who is a *person*
 - In the text “BA went on strike”, using the knowledge that BA is a company and not a person, the IE system can conclude that this is not the kind of strike it is interested in

More semantic search examples

Q: {ScalarValue}{MeasurementUnit} ->

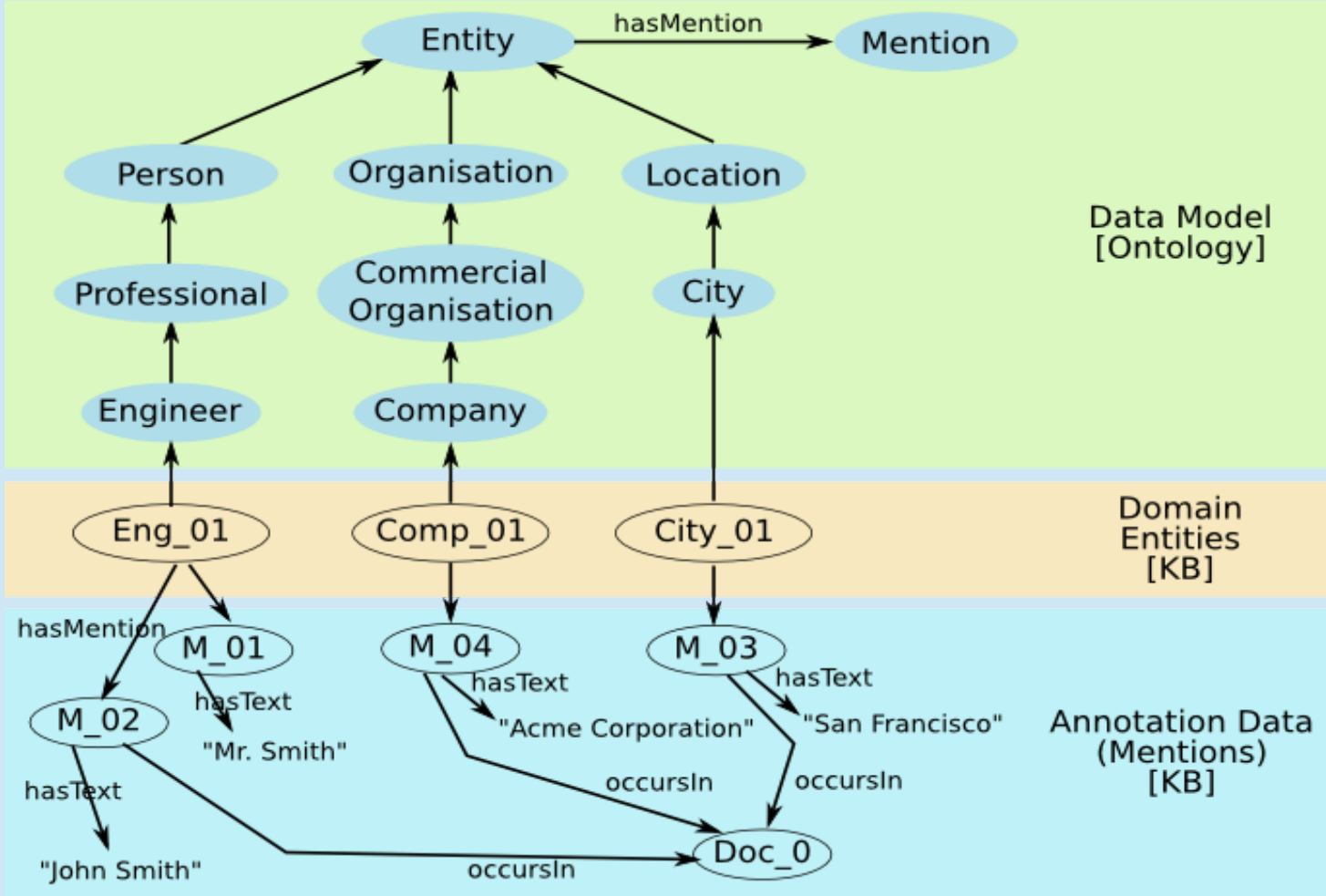
A: "12 cm", "190 g", "two hours"

Q: {Reference} ->

A: JP-A-60-180889

A: Kalderon et al. (1984) Cell 39:499-509

Semantic Annotation



M_02 M_04 M_03
John Smith works at Acme Corporation in San Francisco.

What is an Ontology?

- Set of concepts (instances and classes)
- Relationships between them (is-a, part-of, located-in)
- Multiple inheritance
- Classes can have more than one parent
- Instances can have more than one class



DBpedia

- Machine readable knowledge on various entities and topics, including:
 - 410,000 places/locations,
 - 310,000 persons
 - 140,000 organisations
- For each entity we have:
 - entity name variants (e.g. IBM, Int. Business Machines)
 - a textual abstract
 - reference(s) to corresponding Wikipedia page(s)
 - entity-specific properties (e.g. latitude and longitude for places)

Example from DBpedia

D About: Thames Barrier

dbpedia.org/page/Thames_Barrier

About: Thames Barrier

An Entity of Type : Feature, from Named Graph : <http://dbpedia.org>, within Data Space : <dbpedia.org>

The Thames Barrier is the world's second-largest movable flood barrier and is located downstream of central London, United Kingdom. Its purpose is to prevent London from being flooded by exceptionally high tides and storm surges moving up from the sea. It needs to be raised (closed) only during high tide; at ebb tide it can be lowered to release the water that backs up behind it.

■ ■ ■

owl:sameAs	<ul style="list-style-type: none">▪ http://cs.dbpedia.org/resource/Bariéry_na_Temži▪ http://de.dbpedia.org/resource/Thames_Barrier▪ http://fr.dbpedia.org/resource/Barrière_de_la_Tamise▪ http://it.dbpedia.org/resource/Thames_Barrier▪ http://sws.geonames.org/2636058/▪ freebase:Thames Barrier
geo:geometry	▪ POINT(0.0367 51.4977)
geo:lat	▪ 51.497700 (xsd:float)
geo:long	▪ 0.036700 (xsd:float)

Links to GeoNames
And Freebase

Latitude & Longitude

GeoNames

- 2.8 million populated places
 - 5.5 million alternate names
- Knowledge about NUTS country sub-divisions
 - use for enrichment of recognised locations with the implied higher-level country sub-divisions
- However, the sheer size of GeoNames creates a lot of ambiguity during semantic enrichment
- We use it as an additional knowledge source, but not as a primary source (DBpedia)

5.2 Semantic Annotation In GATE

Information Extraction for the Semantic Web

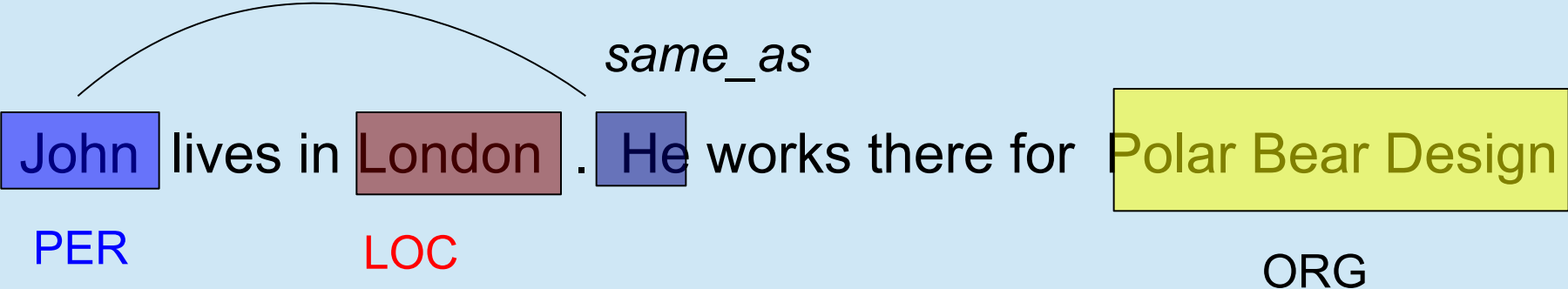
- Traditional IE is based on a flat structure, e.g. recognising Person, Location, Organisation, Date, Time etc.
- For the Semantic Web, we need information in a hierarchical structure
- Idea is that we attach semantic metadata to the documents, pointing to concepts in an ontology
- Information can be exported as an ontology annotated with instances, or as text annotated with links to the ontology

Traditional NE Recognition

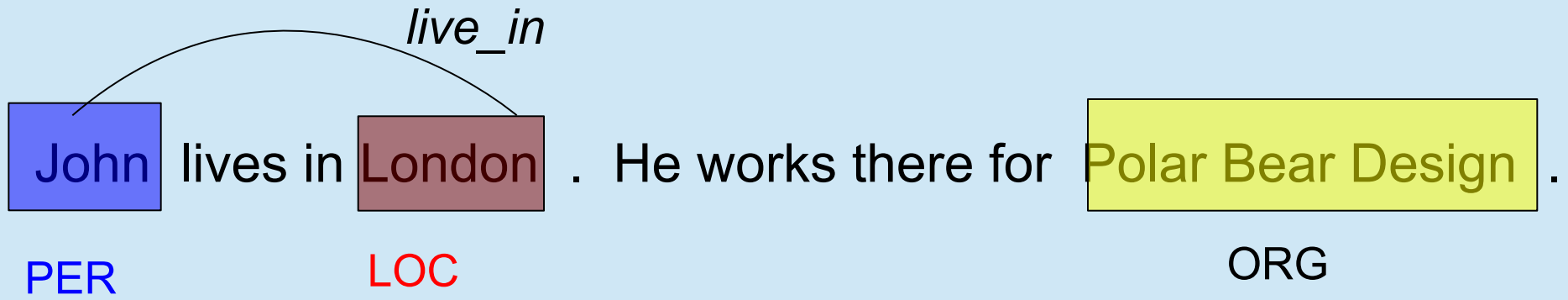
John lives in **London** . He works there for **Polar Bear Design** .

PERSON LOCATION ORGANISATION

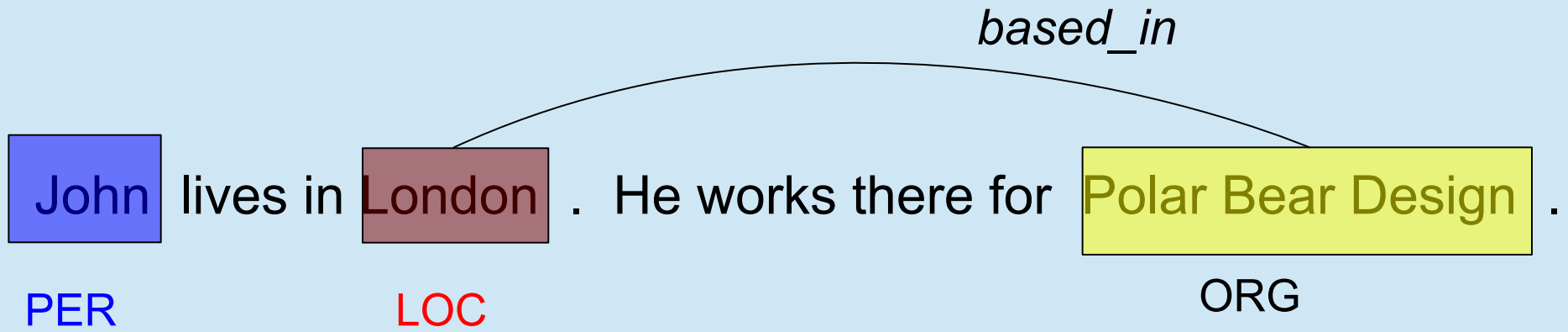
Co-reference



Relations

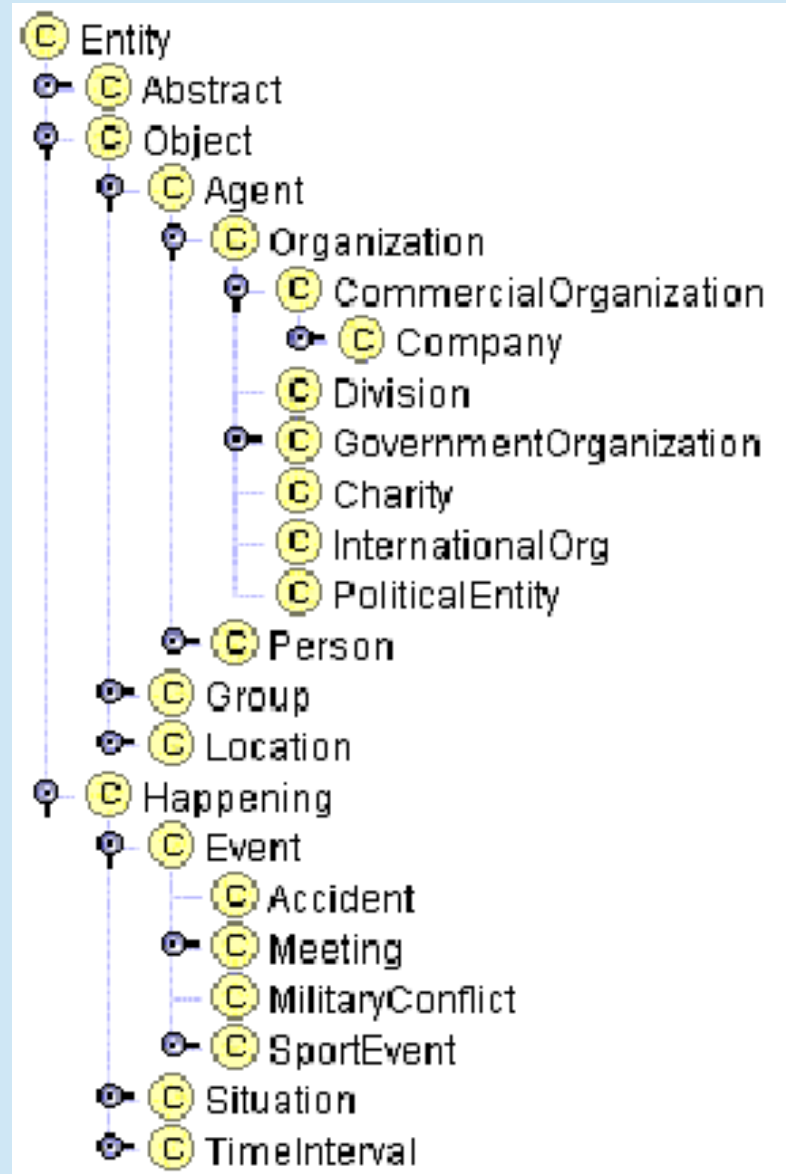


Relations (3)



Richer NE Tagging

- Attachment of instances in the text to concepts in the domain ontology
- Disambiguation of instances, e.g. Cambridge, MA vs Cambridge, UK



Ontology-based IE

The screenshot shows an ontology editor with two main panes. The left pane, titled 'Classes & Instances', displays a hierarchical tree of classes. The right pane, titled 'Properties', lists various properties with their corresponding data types.

Classes and Instances:

- University
- GovernmentOrganization
 - Government
 - Ministry
- InternationalOrganization
- PoliticalEntity
 - Parliament
 - PoliticalParty
- ReligiousOrganization
- ResearchOrganization
 - Institute
 - University
- SportOrganization
 - SportClub
 - SoccerClub
 - SportsFederation
- StockExchange
- Team
- Person
 - Man
 - Woman
- Brand

Properties:

- controls [Object]
- description <http://www.w3.org/2001/X>
- generatedBy [EntitySource]
- hasAddress [Address]
- hasAlias [Alias]
- hasBrother [Man]
- hasChild [Person]
- hasContactInfo [ContactInformation]
- hasDaughter [Woman]
- hasEMail [EMail]
- hasFather [Man]
- hasFax [PhoneNumber]
- hasInternetAddress [InternetAddress]
- hasMainAlias [Alias]
- hasMobilePhone [PhoneNumber]
- hasMother [Woman]
- hasOldName [Alias]
- hasParent [Person]
- hasPhone [PhoneNumber]
- hasPosition [JobPosition]
- hasProfession [Profession]

John lives in London. He works there for Polar Bear Design.

Ontology-based IE (2)

The image shows a screenshot of an ontology browser. On the left is a tree view of classes, and on the right is a list of properties. A red arrow points from the 'City' class in the tree to the 'hasAlias' property in the list.

Tree View (Left):

- RadioStation
- TVChannel
- Currency
- Location
 - AstronomicalObject
 - Facility
 - GlobalRegion
 - LandRegion
 - NonGeographicLocation
 - PoliticalRegion
 - Country
 - County
 - MilitaryAreas
 - Province
 - UrbanDistrict
 - PopulatedPlace
 - City (highlighted)
 - Capital
 - CountryCapital
 - LocalCapital
 - WaterRegion
 - Archipelago
 - Bay
 - Fjord
 - Channel
 - Canal
 - Creek
 - Gulf
 - Harbor
 - Lake

Property List (Right):

- hasAddress [Address]
- hasAirport [Airport]
- hasAlias [Alias] (highlighted)
- hasContactInfo [ContactInformation]
- hasEMail [EMail]
- hasFax [PhoneNumber]
- hasInternetAddress [InternetAddress]
- hasMainAlias [Alias]
- hasMobilePhone [PhoneNumber]
- hasOldName [Alias]
- hasPhone [PhoneNumber]
- hasStationaryPhone [PhoneNumber]
- hasUniversity [University]
- hasWebPage [WebPage]
- isDefinedBy [ALL RESOURCES]
- isOwnedBy [Agent]
- label [ALL RESOURCES]
- laconicDescription <http://www.w3.org/2001/XMLSchema#string>
- latitude <http://www.w3.org/2001/XMLSchema#float>
- locatedIn [Location]
- longitude <http://www.w3.org/2001/XMLSchema#float>
- mainLabel <http://www.w3.org/2001/XMLSchema#string>
- partOf [Entity]
- populationCount <http://www.w3.org/2001/XMLSchema#integer>
- seeAlso [ALL RESOURCES]
- subRegionOf [Location]

John lives in **London**. He works there for Polar Bear Design.

Automatic Semantic Annotation in ENVIA

- Locations (linked to DBpedia and GeoNames)
 - Annotate the place name itself (e.g. Norwich) with the corresponding DBpedia and GeoNames URIs
 - Also use knowledge of the implied reference to the levels 1, 2, and 3 sub-divisions from the Nomenclature of Territorial Units for Statistics (NUTS).
 - For Norwich, these are East of England (UKH – level 1), East Anglia (UKH1 – level 2), and Norfolk (UKH13 – level 3).
 - Similarly use knowledge to retrieve nearby places

“South Gloucestershire” Example

Messages lucene

Annotation Sets

Managing flood risk on the S
January 2011
South Gloucestershire to Hi
Managing flood risk in the S
We are the Environment Ag
better place _ for you, and f
breathe, the water you drin
Government and society as
healthier. The Environment
Please click on the bookma
brochure to specific points
Managing flood risk in the S
Somerset 1

Type	Set	Start
Sem_Location		57
Sem_Location		97
Sem_Location		97
Sem_Location		97
Sem_Location		149
Sem_Location		160
Sem_Location		160

211 Annotations (1 selected)

Sem_Location

alternateName	South Gloucestershire
caption	South Gloucestershire
count	2
countryCode	GB
geonamesURI	http://sws.geonames.org/3333198/
inst	http://dbpedia.org/resource/South_Gloucestershire
latitude	51.5
longitude	-2.41667
lookupRule	fullString
matched	South Gloucestershire
name	South Gloucestershire
parentAdminURI	http://sws.geonames.org/6269131/ , http://sws.geonames.org/3333198/
parentCountryInst	http://sws.geonames.org/2635167/
popularitySimilarity	1.0
randomIndexing	0.0
specificitySimilarity	0.0
string	South Gloucestershire
stringSimilarity	0.2688679
structuralSimilarity	0.0

Semantic Annotation (2)

- Organisations (linked to DBpedia)
 - Names of companies, government organisations, committees, agencies, universities, and other organisations
- Dates
 - Absolute (e.g. 31/03/2012) and relative (yesterday)
- Measurements and Percentages
 - e.g. 8,596 km² , 1 km, one fifth, 10%

Semantic Search: An Overview

GATE Mimir

- can be used to index and search over text, annotations, semantic metadata (concepts and instances)
- allows queries that arbitrarily mix full-text, structural, linguistic and semantic annotations
- is open source



What can GATE Mimir do that Google can't?

Show me:

- all documents mentioning a temperature between 30 and 90 degrees F (expressed in any unit)
- all abstracts written in French on patent documents from the last 6 months which mention any form of the word “transistor” in the English abstract
- the names of the patent inventors of those abstracts
- all documents mentioning steel industries in the UK, along with their location

Search News Articles for Politicians born in Sheffield

GUS - GATE Unified Search

services.gate.ac.uk/mimir/gpd/search/gus#page=1

Search powered by **Mimir**

Searching index: News Demo

```
{Person sparql="SELECT DISTINCT ?inst WHERE { ?inst :birthPlace <http://dbpedia.org/resource/Sheffield> . ?inst a :Politician }"}}
```

Search

Results 1 - 10 of 41

Oona King's knife crime pledge in mayoral candidate bid (cached)
BBC News - **Oona King's** knife crime pledge in

Oona King's knife crime pledge in mayoral candidate bid (cached)
reddit StumbleUpon Twitter Email Print **Oona King's** knife crime pledge in

Oona King's knife crime pledge in mayoral candidate bid (cached)
pledge in mayoral candidate bid **Ms King** lost her parliamentary seat to

Oona King's knife crime pledge in mayoral candidate bid (cached)
to George Galloway in 2005 **Oona King** promised to improve the lives

<http://demos.gate.ac.uk/mimir/gpd/search/gus>

Easily Create Your Own Custom GATE Mimir Interfaces

People in the News

demos.gate.ac.uk/pin/?name=&bornIn=Sheffield&famousAs=Politician|OfficeHolder&after=01%2F04%2F2011&before=30%2F04%2F2011

PEOPLE IN THE NEWS

Looking For...

Name:

Fuzzy Name Matching

Born In:

Famous As:

In Articles...

Published Between and

Classified As:

Ignore Boilerplate Text

Search

Results 1 to 2 of 2 [Show Underlying Mimir Query](#)

[Scottish election: Respect Coalition Against Cuts profile](http://www.bbc.co.uk/news/uk-scotland-13048761)
<http://www.bbc.co.uk/news/uk-scotland-13048761> - [Cached](#)
... Bow - whose sitting MP **Oona King** had voted for the war ...
... success came when Galloway overturned **Oona King's** 10,000- ...

Powered by GATE Mimir
© The University of Sheffield, 2011

<http://demos.gate.ac.uk/pin/>

MIMIR: Searching Text Mining Results

- Searching and managing text annotations, semantic information, and full text documents in one search engine
- Queries over annotation graphs
- Regular expressions, Kleene operators
- Designed to be integrated as a web service in custom end-user systems with bespoke interfaces
- Demos at <http://services.gate.ac.uk/mimir/>

Scaling Up

- We annotated 1.08 million web pages using a GATE language analysis pipeline.
 - Documents crawled using Heritrix10 , with total content size of 57 GiB or 6.6 billions plain text characters.
 - The indexing server has 2 Intel Xeon 2.8GHz CPUs 11 GB of RAM, and runs 64 bit Ubuntu Linux. Indexing process was 94 hours.
- We also indexed 150 million similar web pages, using two hundred Amazon EC2 Large Instances running for a week to produce a federated index
- Mimir runs on GateCloud.net, so easy to scale up

Search for string **Harriet Harman**

Harriet Harman

Search

Documents 1 to 20 of 81:

[UK childcare needs to be more affordable - CentreForum \(cached\)](#)

quality' MP **Harriet Harman** was the architect

[Birth weight among social mobility checks - Nick Clegg \(cached\)](#)

's deputy leader **Harriet Harman** said Mr Clegg

[Ed Miliband's shadow cabinet and ministerial teams \(cached\)](#)

Miliband Opposition leader **Harriet Harman** Deputy Leader & in 2011. **HARRIET HARMAN** - DEPUTY LEADER

[PM's response to Skinner Commons question 'shameful' \(cached\)](#)

. Deputy leader **Harriet Harman** wrote on Twitter

[Daily Politics and Sunday Politics highlights of 2012 \(cached\)](#)

Sunday April 29 **Harriet Harman** on Hunt, Cameron and Clegg **Harriet Harman** struggles with bank on health by **Harriet Harman**
PMQs: Harriet ...

[Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance' \(cached\)](#)

Shadow Culture Secretary **Harriet Harman** says Jeremy Hunt culture secretary, **Harriet Harman**, told the

[No Rupert Murdoch deal, says Alastair Campbell \(cached\)](#)

Search for string **Harriet Harman says**

Harriet Harman says

Search

Documents 1 to 20 of 29:

Leveson Inquiry: Jeremy Hunt 'sought News Corp guidance' (cached)

Shadow Culture Secretary **Harriet Harman says** Jeremy Hunt was

Ed Miliband defends Iraq war condemnation (cached)

, says Harman **Harriet Harman says** Labour will be

Ed Miliband tells Labour: We're the optimists now (cached)

, says Harman **Harriet Harman says** Labour will be

Labour must have credible deficit plan, says Darling (cached)

, says Harman **Harriet Harman says** Labour will be

David Miliband says he won't join brother Ed's team (cached)

, says Harman **Harriet Harman says** Labour will be

Balls: Labour must fight cuts 'every inch of the way' (cached)

, says Harman **Harriet Harman says** Labour will be

Search with morphological variants: Harriet Harman root:say

Harriet Harman root:say

Search

Documents 21 to 18 of 38:

[David Cameron criticised for 'calm down dear' jibe](#) (cached)

former equality minister **Harriet Harman** said Mr Cameron's

[Queen's Speech: Biggest change to voter registration](#) (cached)

, Labour's **Harriet Harman** said the government was

[Harriet Harman struggles with bank bonus and job figures](#) (cached)

in Coventry, **Harriet Harman** said: "I

[PMQs: Harriet Harman and Nick Clegg on unemployment](#) (cached)

Labour, but **Harriet Harman** said unemployment was falling

[Leveson Inquiry: Jeremy Hunt fair on BSkyB, says top civil servant](#) (cached)

Shadow culture secretary **Harriet Harman** said: "David

[Jeremy Hunt: I followed due process over BSkyB](#) (cached)

But Labour's **Harriet Harman** said Mr Hunt had

[Ed Miliband 'will marry' but politics 'got in the way'](#) (cached)

Replace strings with NEs: {Person} root:say

{Person} root:say

Search

Documents 1 to 20 of 3980:

Apple's Sir Jonathan Ive reaffirms desire to stay at company (cached)

Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

Diamond Jubilee Tube train was faulty (cached)

be happening' **Ms Siggs said**: "It

Warning over deep-ocean stowaways (cached)

embarrassment. But **Dr Voight says** the experience is it," **Dr Voight said**. "We

School building system not fit for purpose, review says (cached)

shadow education secretary **Andy Burnham said** Mr Gove had . General secretary **Chris Keates said** the capital budget BCSE) director **Ty Goddard said** there was "

EU wants Greece to stay in eurozone, says Van Rompuy (cached)

's Europe editor **Gavin Hewitt says** the crisis gives UK Prime Minister **David Cameron said** "there was . German Chancellor **Angela Merkel said** the bonds,

Huhne partner loses privacy case (cached)

000 costs. **Ms Trimmingham said** this could become Daily Mail's **Andrew Pierce said** it was a Speaking outside court **Ms Trimmingham said** she was disappointed

{Person} AND root:say – 11803 hits

[Stone Roses reunion gig hailed by fans](#) (cached)

by fans By Ian Youngs Entertainment reporter, BBC News The gi ... re. "They've never played so well together," said 43-year-old together," said 43-year-old Andrew Rudder, from Ashton 43-year-old Andrew Rudder, from Ashton under Lyne. But opinion ... s voice. "He can't sing but he never could," said Tom Six, ...

[000011_http://www.bbc.co.uk/news/northern_ireland/](#) (cached)

02:07 Michaela McAreavey trial - the first day Watch 02:51 Minister says she has ' :51 Minister says she has 'heard enough' Watch 01:12 Flat bombs find - man arrested Watch 01:39 Michaela's husband braves

[Warning over deep-ocean stowaways](#) (cached)

using the famous Alvin sub say the vehicle picked famous Alvin sub say the vehicle picked up limpets from a depth of ... s had to cope with huge pressure changes as Alvin conducted its dives pressure changes as Alvin conducted its dives. The researchers report ... matter of some embarrassment. But Dr Voight says the experience is ...

[In pictures: Royal arts gathering](#) (cached)

Queen. Sir Paul McCartney, who was among the musicians the event, said he was " the event, said he was "a big fan" of the monarch. Artist David Hockney shared a few

[School building system not fit for purpose, review says](#) (cached)

purpose, review says Accessibility links Skip to content Skip to I ... ucation & Family Home World UK England N. Ireland Scotland Wales Business World UK England N. Ireland Scotland Wales Business Politics Health ... building system not fit for purpose, review says Some schools awaiting purpose, review says Some schools awaiting rebuilds rely on tempor ... government-commissioned review by Sebastian James of Dixons Group ...

[Hewlett-Packard to cut 27,000 jobs by the end of 2014](#) (cached)

World UK England N. Ireland Scotland Wales Business Politics Health ... cut 27,000 jobs by end of 2014. The company said the cuts - . The company said the cuts - about 8% of its workforce - will r ... riod fell 3% on a year ago to \$30.7bn. Meg Whitman, HP's year. HP said in a statement that the money would be reinve ... riod fell 3% on a year ago to \$30.7bn. Meg Whitman, HP's ...

{Person} [0..5] root: say – 5495 hits

Documents 1 to 20 of 5495:

Apple's Sir Jonathan Ive reaffirms desire to stay at company (cached)

Today programme, **Sir Jonathan said** he would stay partner". **Sir Jonathan said** that Apple's

Diamond Jubilee Tube train was faulty (cached)

be happening' **Ms Siggs said**: "It

Warning over deep-ocean stowaways (cached)

using the famous **Alvin sub say** the vehicle picked embarrassment. But **Dr Voight says** the experience is it," **Dr Voight said**. "We

School building system not fit for purpose, review says (cached)

government-commissioned review by **Sebastian James of Dixons Group said** value for money by Education Secretary **Michael Gove**, **Mr James said**. Schools with shadow education secretary **Andy Burnham said** Mr Gove had ...

EU wants Greece to stay in eurozone, says Van Rompuy (cached)

, European Council **President Herman Van Rompuy has said**. He was European Council President **Herman Van Rompuy has said**. He was 's Europe editor **Gavin Hewitt says** the crisis gives ...

Huhne partner loses privacy case (cached)

000 costs. **Ms Trimingham said** this could become Daily Mail's **Andrew Pierce said** it was a Speaking outside court **Ms Trimingham said** she was disappointed ...

Migration to UK more than double government target (cached)

. Immigration Minister **Damian Green said**: "Our figures. Chairman **Sir Andrew Green said**: "You

No 'inappropriate' government contact, News Corp lobbyist tells Leveson (cached)

adviser. Fred **Michel said** he did not 's team. **Adam Smith stood down after saying** his e-mails with inquiry, Mr **Michel says** he did not ...

Patent Annotation: Data Model

- DocumentObject
 - Claim
 - DocumentPage
 - DocumentSection
 - Equation
 - Figure
 - Formula
 - Table

- DocumentSection
 - Abstract
 - BackgroundArt
 - BestMode
 - BibliographicData
 - ClaimSection
 - DetailedDescription
 - DisclosureOfInvention
 - DrawingDescription
 - DrawingsSection
 - Effects
 - OtherInformation
 - PreferredEmbodiement
 - PriorArt
 - RepresentativeDrawing
 - SearchReport
 - SummaryOfInvention
 - TechnicalProblem
 - TechnicalSolution
 - TechnologicalField
 - UsageOfInvention

- Reference
 - RefClaim
 - RefCompany
 - RefEquation
 - RefFig
 - RefFormula
 - RefLiterature
 - RefPatent
 - RefProduct
 - RefTable

- MeasurementUnit
 - AccelerationUnit
 - AmountUnit
 - AngleUnit
 - AreaUnit
 - ChargeUnit
 - CurrentUnit
 - EnergyUnit
 - ForceUnit
 - FrequencyUnit
 - InformationUnit
 - LengthUnit
 - LuminousIntensityUnit
 - MassUnit
 - MusicalNoteLengthUnit
 - PowerUnit
 - PressureUnit
 - SpeedUnit
 - TemperatureUnit
 - TimeUnit
 - UnknownUnit
 - VoltageUnit
 - VolumeUnit

An Example Text

Annotation Sets

Annotations List

Co-reference Editor

Text



[0039] Worthy of note, the aforementioned hydrolysis of BTSP to H₂O₂ is the simplest of many scenarios which require a certain Measurement interval for H₂O. In a more general way, the need for a protic solvent is stated in FIG. 4 b.

[0040] In accord with the previous observations, additives such as pyridines serve to prevent sensitive epoxide ring opening by buffering the highly acidic rhenium species. Notably, compared to the original system, the amount of ligand necessary to achieve the desired Measurement now decreased from 12 to 0.5-1 mol% in both MTO and Re₂O₇ systems (The use of 12 mol% of pyridine completely arrested the reaction, presumably due to base-mediated decomposition of MTO). In some instances MTO loadings can be lowered to 0.25 mol% without affecting conversions—a manifestation of prolonged catalyst lifetime under the present conditions.

[0041] The use of Re₂O₇, ReO₃(OH) and ReO₃ as catalyst precursors is a particularly important feature of the present protocol. Catalytic activities of these inorganic rhenium species for epoxidation with H₂O₂ were known to be very poor. For the epoxidation of C₂-20 olefins with stoichiometric Re₂O₇ in the presence of pyridine, see: Union Oil Co. of California (Fenton, D. M.) U.S. Pat. No. 3,316,279; (c) for early applications of Re₂O₇ in olefin/H₂O₂ oxidation catalysis see: duPont de Nemours and Co. (Parshall, G. W.) U.S. Pat. Nos. 3,657,292 and 3,646,130 (1972); (d) Warwel and co-workers found that Re₂O₇ is a more effective epoxidation catalyst if the right solvent is chosen. Their system employs 60% aqueous H₂O₂ in 1,4-dioxane at 90°C and 1,2-diols are isolated in good yields, the initially formed epoxides being unstable in this system: Warwel, S.; Ruschgen Klaas, M.; Sojka, M. Chem. Commun. 1991, 1578; (e) Herrmann, W. A.; Correia, J. D. G.; Kuhn, F. E.; Artus, G. R. J. Chemistry—A European Journal 1996, 2, 168.

[0042] Generally, the high acidity of these systems does not allow epoxides to be isolated except in special cases such as from cis-cyclooctene (which affords an epoxide which is particularly resistant to acid-catalyzed ring opening). In the present system,

safe.preprocessing

Measurement

Reference

Section

Figure reference

Measurement unit

Patent reference

Literature reference

Hands-On with patent data

<http://demos.gate.ac.uk/mimir/patents/search/index>

Text. Matches plain text.

Example: `nanomaterial`

Linguistic variations of text

Example: `(root:nanomaterial | root:nanoparticle)`

Annotation. Matches semantic annotations.

Syntax: `{Type feature1=value1 feature2=value2...}`

Example: `{Abstract lang="DE"}`

Sequence Query. Sequence of other queries.

Syntax: `Query1 [n..m] Query2...`

Example: `from {Measurement} [1..5] {Measurement}`

Inclusion Queries

IN Query. Hits of one query only if inside another.

Syntax: Query1 IN Query2

Example: (root:nanomaterial | root:nanoparticle) IN {Abstract}

Number of times these words are mentioned in patent abstracts (as well as links to the actual documents)

OVER Query. Hits of a query, only if overlapping hits of another.

Syntax: Query1 OVER Query2

Example: {Abstract} OVER (root:nanomaterial | root:nanoparticle)

Finds all abstracts that contain nanomaterial(s) or nanoparticle(s)

Date restrictions

```
(  
  {Abstract lang="EN"} OVER  
    (root:nanomaterial | root:nanoparticle )  
)  
IN  
{PatentDocument date > 20050000}
```

YYYYMMDD

Find references to literature or patents in the prior art or background sections, which contain nanomaterial/nanoparticle

```
{Reference type="Literature"}
```

```
|
```

```
{Reference type="Patent"}
```

```
) IN
```

```
{Section type="PriorArt"}
```

```
|
```

```
{Section type="BackgroundArt"}
```

```
)
```

```
OVER
```

```
(root:nanomaterial | root:nanoparticle)
```

Queries Using External Knowledge

{Measurement spec="1 to 100 volts"}

- Uses GNU Units (<http://www.gnu.org/software/units/>) to convert measurements and normalise them to SI units

{Measurement spec="1 to 100 kg m² / A s³"}

- **Example hits:** 10 volts, 2V, +20 and -20 volts; ±10V; +/- 100V; +3.3 volts

{Measurement spec="1 to 100 m / s"}

- **Example hits:** 40 km/hr, 60m/min, 100cm/sec, 60 fps; 10 to 2000 cm/sec

Searching LOD with SPARQL

- SQL-like query language for RDF data
- Simple protocol for querying remote databases over HTTP
- Query types
 - *select*: projections of variables and expressions
 - *construct*: create triples (or graphs) based on query results
 - *ask*: whether a query returns results (result is true/false)
 - *describe*: describes resources in the graph

SPARQL Example

Software companies founded in the US

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
```

```
PREFIX dbp-ont: <http://dbpedia.org/ontology/>
```

```
PREFIX geo-ont: <http://www.geonames.org/ontology#>
```

```
PREFIX umbel-sc: <http://umbel.org/umbel/sc/>
```

```
SELECT DISTINCT ?Company ?Location
```

```
WHERE {
```

```
    ?Company rdf:type dbp-ont:Company ;
```

```
        dbp-ont:industry dbpedia:Computer_software ;
```

```
        dbp-ont:foundationPlace ?Location .
```

```
    ?Location geo-ont:parentFeature dbpedia:United_States .
```

```
}
```

SPARQL Results

SPARQL Query

Results for your query (81) - [Edit query](#)

View as [Exhibit](#)

Company	Location
dbpedia:Redxpress	dbpedia:Glen_Allen,_Missouri
dbpedia:Borland	dbpedia:California
dbpedia:Lawson_Software	dbpedia:Minneapolis
dbpedia:Exterro,_Inc.	dbpedia:Oregon
dbpedia:Tableau_Software	dbpedia:Seattle
dbpedia:NeuroDimension	dbpedia:Florida
dbpedia:Computer_Usage_Company	dbpedia:New_York_City
dbpedia:Macromedia	dbpedia:San_Francisco
dbpedia:Core_International,_Inc	dbpedia:Florida
dbpedia:Cerulean_Studios	dbpedia:Connecticut
dbpedia:Cornerstone_OnDemand	dbpedia:California

Try at: <http://factforge.net/sparql>

Documents mentioning Persons born in Sheffield

```
{Person sparql = "SELECT ?inst WHERE { ?inst :birthPlace <http://dbpedia.org/resource/Sheffield>}"}
```

Ed Miliband's shadow cabinet and ministerial teams (cached)
and employment minister **David Blunkett** before becoming an

The BBC web document does not mention Sheffield at all:
<http://www.bbc.co.uk/news/uk-politics-11494915>

The relevant text snippet is:

HILARY BENN - SHADOW COMMUNITIES AND LOCAL GOVERNMENT SECRETARY

As the son of former Labour cabinet minister Tony Benn, the MP for Leeds Central is part of a political dynasty. Regarded as more pragmatic than his father, he was a union official and special adviser to then education and employment minister David Blunkett before becoming an MP in 1999. Well-regarded as international development and environment secretary under Gordon Brown despite having a generally low profile. At the **age of 53** stood for the deputy leadership in 2007, coming fourth. One of Ed Miliband's primary supporters in the leadership contest.

Try these on the BBC News demo:

<http://services.gate.ac.uk/mimir/gpd/search/index>

Gordon Brown [0..3] root:say

{Person} [0..3] root:say

{Person inst ="http://dbpedia.org/resource/Gordon_Brown"} [0..3] root:say

{Person sparql="SELECT ?inst WHERE { ?inst a :Politician }"} [0..3] root:say

{Person sparql = "SELECT ?inst WHERE { ?inst :party
<http://dbpedia.org/resource/Labour_Party_%28UK%29> }" } [0..3] root:say

{Person sparql = "SELECT ?inst WHERE {
?inst :party <http://dbpedia.org/resource/Labour_Party_%28UK%29> .
?inst :almaMater <http://dbpedia.org/resource/University_of_Edinburgh> }"
} [0..3] root:say

User Interfaces for SPARQL-based Semantic Search

- SPARQL-based semantic searches, tapping into LOD resources are extremely powerful
- However, impossible to write by the vast majority of users
- User interfaces for SPARQL-based semantic search:
 - Faceted searches (see ExoPatent next)
 - Form-based searches (see EnviLOD, PIN)
 - Text-based searches (natural language interfaces for querying ontologies), e.g. FREyA

Faceted Search: ExoPatent Example

- Use semantic information to expose linkages between documents based on the intersecting relationships between various sets of data from
 - FDA Orange Book (23,000 patented drugs)
 - Unified Medical Language System (UMLS) – database of medical terms (370,000)
 - Patent bibliographic information
- Search for diseases, drug names, body parts, references to literature and other patents, numeric values, ranges
- Demo uses a small set of patents (40,000)

ExoPatent: Faceted Search

Facets >

Selected Items

- AMOXICILLIN
- GENTAMICIN

Recent Items

- Human herpesvirus 1
- MERCK & CO INC

Terms from FDA Orange Book

FDA Drug Name

25 of 1456 shown below.

- ALBUTEROL SULFATE
- CARBAMAZEPINE
- CLOTRIMAZOLE
- ETHOSUXIMIDE
- GENTAMICIN
- GUANFACINE HYDROCHLORIDE
- INDOCIN
- MERCAPTOPURINE
- MISOPROSTOL
- NALBUPHINE
- NEOSAR
- NIZATIDINE
- OCTREOTIDE ACETATE
- PERPHENAZINE
- PODOFILOX
- POTASSIUM CITRATE

Active Ingredients

25 of 1460 shown below.

- ALBUTEROL SULFATE
- AMOXICILLIN
- AMPHETAMINE SULFATE
- CEFTAZIDIME
- CETRORELIX
- CLAVULANATE POTASSIUM
- DEXTROAMPHETAMINE SULFAT...
- DOXAZOSIN MESYLATE
- ETHOSUXIMIDE
- GENTAMICIN SULFATE
- HEPARIN SODIUM
- HYDRALAZINE HYDROCHLORID...
- HYDROCHLOROTHIAZIDE
- METHOTREXATE SODIUM
- METRIZOIC ACID
- NALBUPHINE HYDROCHLORIDE

Applicant

25 of 273 shown below.

- ABBOTT GMBH & CO KG
- ALLERGAN INC
- ALZA CORP
- ALZA CORPORATION
- BOARD OF REGENTS, THE UN...
- BOEHRINGER INGELHEIM PHA...
- FUI SZ TECHNOLOGIES LTD
- FUI SZ TECHNOLOGIES LTD.
- HOFFMANN LA ROCHE
- HOFFMANN-LA ROCHE INC.
- MERCK & CO INC
- MERCK & CO., INC.
- NOVEN PHARMA
- NOVEN PHARMACEUTICALS, I...
- PROCTER & GAMBLE
- SCHERING CORP

UMLS Concept

25 of 4130 shown below.

- Ankylosing spondyl
- Autoimmune hemol
- Brucella melitensis
- Bullous pemphigoi
- Cytomegalovirus
- Enterobacter aerog
- Enterovirus
- Genus: Coronavir
- Heart failure
- Human herpesvirus
- Human Herpesviru
- Myotonic Dystrophy
- Postpericardiotomy
- Respiratory syncyti
- Rhinovirus
- Scleroderma

Document Keyword Filter

Patent Documents Containing FDA-related Terms

1-10 of 362 documents matching the search criteria.

Matching documents: **362**

- [DOCUMENTS](#)
- [TIMELINES](#)

Publication Date	Patent Number	Assignee(s)	Title
10-11-2005	US-20050250705-A1	BOEHRINGER INGELHEIM PHARMA GM...	Spray-dried powder comprising at least one 1,... ... pefloxacin, amifloxacin, fleroxacin, tosufloxacin, pruliflo... irioxacin, pazufloxacin, clinafloxacin and sitafloxacin; amin... such as, for example, gentamicin , netilmicin, paramecin, t... amibacin, bezafloxacin, racemycin, streptomycin, vancomycin

Find all applicants who filed patents related to mitochondria, as well as drug names and active ingredients

Selected Items	Terms from FDA Orange Book			
<input type="checkbox"/> Mitochondria	FDA Drug Name	Active Ingredients	Applicant	UMLS Concept
Recent Items (No recent items)	<input type="text"/> 25 of 1451 shown below.	<input type="text"/> 25 of 1141 shown below.	<input type="text"/> 25 of 828 shown below.	<input type="text"/> 25 of 7998 shown below.
	ALBUTEROL SULFATE CARBAMAZEPINE CLOTRIMAZOLE ETHOSUXIMIDE GENOTROPIN GENTAMICIN HYZAAR MERCAPTOPYRINE MISOPROSTOL NALBUPHINE NEOSAR NIZATIDINE OCTREOTIDE ACETATE PERPHENAZINE PODOFILOX POTASSIUM CITRATE	ALBUTEROL SULFATE AMOXICILLIN AMPHETAMINE SULFATE CEFTAZIDIME CETRORELIX CINOXACIN CLAVULANATE POTASSIUM DEXTROAMPHETAMINE SULFAT... DOXAZOSIN MESYLATE ETHOSUXIMIDE ETOPOSIDE GENTAMICIN SULFATE HEPARIN SODIUM HYDROCHLOROTHIAZIDE METHOTREXATE SODIUM NALBUPHINE HYDROCHLORIDE	AJINOMOTO CO., INC. ALLERGAN INC BOARD OF REGENTS, THE UN... BOEHRINGER MANNHEIM GMBH BRITISH TECH GROUP CHIRON CORP CHIRON CORPORATION FUJIREBIO KABUSHIKI KAIS... FUJIREBIO KK HARVARD COLLEGE HESKA CORPORATION MANDEL ARKADY MEDICAL DISCOVERIES INC MERCK & CO INC MERCK & CO., INC. MINEMURA TSUYOSHI	Acetylation Adenoviruses Aedes Alphavirus Animals Anopheles Genus Ants Anus Bacteria Bacteriophage la Bacteriophages Blood Blood capillaries Cell Wall Cells Chagas Disease

Semantic Search over Content and Annotations



The University Of Sheffield.



Semantic Enrichment with Linked Open Data: A Case Study on Environmental Science Literature

Search [Help](#)

Keywords

Narrow down your search:

Location

Restrict your search to paragraphs sentences

- none
- population
- longitude
- latitude
- name
- country code
- population density
- with nearby

EnviLOD Semantic Search UI



The University Of Sheffield.



HR Wallingford



Semantic Enrichment with Linked Open Data: A Case Study on Environmental Science Literature

Search [Help](#)

Keywords

Narrow down your search:

Search to document paragraphs sentences

- none
- Document
- Location
- Date
- Organization
- River

EnviLOD Semantic Search UI



Semantic Enrichment with Linked Open Data: A Case Study on Environmental Science Literature

Search [Help](#)

Keywords

Narrow down your search:

Location

Restrict your search to paragraphs sentences

- none
- population
- longitude
- latitude
- name
- country code
- population density
- nearby

EnviLOD Semantic Search UI



Semantic Enrichment with Linked Open Data: A Case Study on Environmental Science Literature

Search [Help](#)

Keywords

Narrow down your search:

Location


Restrict your search to document paragraphs sentences

Example Results

Development and flood risk : : practice guide

Example hits:

..." flood defences, or to flood alleviation schemes which provide benefit to the wider community. An example is provided below. Case study The Avenue Site, Chesterfield - example of organisations working" " working together to help reduce flood risk and create wetland habitats This ongoing project is involving the restoration and de-contamination of a former major coking works to the south of Chesterfield by the East Midlands Development" " of new wetland, a flood storage area and a restored section of the River Rother. The project will result in reductions in flood risk downstream in Chesterfield. A steering group comprising" ...

Keywords: Flood control--Great Britain, Flood damage prevention--Great Britain, Floodplain management--Great Britain, Other social problems and services ([other metadata](#) )

Lower Derwent flood risk management strategy : non-technical summary

Example hits:

..." the Environment Agency. Managing Flood Risk in Derby and the Lower Derwent The" " . We cannot prevent floods. There

...and the underlying SPARQL Query

```
{Sem_Location dbpediaSparql="select distinct ?inst
where {
  {{ ?inst <http://dbpedia.org/property/north> ?loc} UNION
  { ?inst <http://dbpedia.org/property/east> ?loc } UNION
  { ?inst <http://dbpedia.org/property/west> ?loc } UNION
  { ?inst <http://dbpedia.org/property/south> ?loc } UNION
  { ?inst <http://dbpedia.org/property/northeast> ?loc } UNION
  { ?inst <http://dbpedia.org/property/northwest> ?loc } UNION
  { ?inst <http://dbpedia.org/property/southeast> ?loc } UNION
  { ?inst <http://dbpedia.org/property/southwest> ?loc }
FILTER(REGEX(STR(?loc), \"Sheffield\", \"i\"))}"} AND (root:"flood")
```

Ongoing work: Use GeoSparql instead and be able to specify distances and reason with the richer information in GeoNames

Flooding in Oxford

Keywords

Narrow down your search:

Restrict your search to document paragraphs sentences

Mimir Query: [Show](#)

Showing 1 to 3 of 3 hits. Pages: [1](#)

[The governments response to Sir Michael Pitt?s review of the summer 2007 Floods : progress report](#)

Example hits:

..." fund early action to tackle flood risk. Applications were due by 30 November. Successful applic...olk 16. Northumberland 17. North Yorkshire 18. Nottinghamshire 19. Oxfordshire4 20. Somerset 21"" 4 Includes all authorities in Oxfordshire 5 Includes responses from all authorities in Suffolk...mation about this publication and copies are available from: Flood Management Division 2D Ergon" ...

Keywords: ([other metadata](#))

[Flooding in London : a London Assembly scrutiny report](#)

Example hits:

..." study in 2000 by the Flood Hazard Research Centre at Middlesex University of the longer-term health effects of the 1998 flooding in Banbury and Kidlington in the Thames"" 1998 flooding in Banbury and Kidlington in the Thames region. Studies in flood-affected areas r...ced by the engineering profession on the human distress caused by flooding - its social impact." ...

Keywords: Flood damage prevention--England--London, Floods--England--London, Flood control--England--London, Emergency management--England--London, Floodplain management, England, London, Other social problems and services ([other metadata](#))

We don't just have to look at politicians saying and measuring things

- If we first process the text with other NLP tools such as sentiment analysis, we can also search for positive or negative documents
- Or positive/negative comments about certain people/topics/things/companies
- In the Decarbonet project, we looked at people's opinions about topics relating to climate change, e.g. fracking
- We could index on the sentiment annotations too
- Other people are using the combination of opinion mining and MIMIR to look at e.g. customer feedback about products / companies on a huge corpus

Explicitly Choosing The Search Classes

GATE Prospector

Search ▾

Diseases Pathogens Pathogenesis Vaccine Animals and Models Custom Mimir Query

URI:

Disease ▾
Bacterial_disease ▾
Viral_disease ▾
Acquired_immunity
Acute_hepatitis
Argentinian_ha

Document Metadata

Dates: to

source:

Search

Documents Terms

[PubMed By PMIDAbstract1000.txt](#)
pertussis and Haemophilus influenzae.

[PubMed By PMIDAbstract10033.txt](#)
syncytial virus (RSV). The injection of purified RSV in Freund's inoculation of purified RSV with Bordetella pertussis ...

[PubMed By PMIDAbstract10058.txt](#)
cell infiltrates. Hepatitis and splenitis with

[PubMed By PMIDAbstract10085.txt](#)
coli, Haemophilus influenzae and Proteus mirabilis

[PubMed By PMIDAbstract1012.txt](#)
and one-half received hepatitis A vaccine (control

[PubMed By PMIDAbstract1013.txt](#)
features of fatal influenza virus infection in national surveillance for influenza-associated deaths

1-20 of 10,256

Environmental signals implicated in Dr fimbriae release by pathogenic Escherichia coli. Afa/Dr diffusely adhering Escherichia coli have been shown to cause urinary tract infections and enteric infections. Virulence of Dr-positive IH11128 bacteria is associated with the presence of Dr fimbriae. In this report, we show for the first time that the Dr fimbriae are released in the extracellular medium in response to multiple environmental signals. Production and secretion of Dr fimbriae are clearly thermoregulated. A comparison of the amounts of secreted fimbriae showed that the secretion is drastically increased during anaerobic growth in minimal medium. The effect of anaerobiosis on secretion seemed to depend on both the growth phase and the culture medium. The secretion was maximal during the logarithmic-phase growth and corresponded to 27 and 57% of total Dr fimbriae produced by bacteria grown in mineral medium+glucose and LB broth, respectively. Thus, the anaerobic environment of the colon would favour the secretion of Dr fimbriae during bacterial multiplication. The controlled release of the Dr fimbriae, which is carried out in the absence of cellular lysis, appears independent of the action of proteases or a process of maturation. The mechanism employed in the liberation of Dr fimbriae thus seems different from that described for the adhesive EHA and Hap of Bordetella pertussis.

Choosing A Specific Instance

GATE Prospector

Search

Diseases Pathogens Pathogenesis Vaccine Animals and Models Custom Mimir Query

URI:

Disease	Bacterial_disease	Burkitts_lymphoma
	Viral_disease	Cervical_cancer
		Chandipura_encephalitis

Document Metadata

Dates: to

source:

Search

Documents Terms

PubMed By PMIDAbstract1306.txt the development of cervical cancer . The HPV	<p>Eradication of established tumors by vaccination with recombinant Bordetella pertussis adenylate cyclase carrying the human papillomavirus 16 E7 oncoprotein. High-risk human papillomaviruses (HPV) such as HPV16 are associated with the development of cervical cancer. The HPV16-E6 and HPV16-E7 oncoproteins are expressed throughout the replicative cycle of the virus and are necessary for the onset and maintenance of malignant transformation. Both these tumor-specific antigens are considered as potential targets for specific CTL-mediated immunotherapy. The adenylate cyclase (CyaA) of Bordetella pertussis is able to target dendritic cells through specific interaction with the alpha(M)beta(2) integrin. It has been previously shown that this bacterial protein could be used to deliver CD4(+) and CD8(+) T cell epitopes to the MHC class II and class I presentation pathways to trigger specific Th and CTL responses in vivo, providing protection against subsequent viral or tumoral challenge. Here, we constructed recombinant CyaA containing either the full sequence or various subfragments from the HPV16-E7 protein. We show that, when injected to C57BL/6 mice in absence of any adjuvant, these HPV16-recombinant CyaAs are able to induce specific Th1 and CTL responses. Furthermore, when injected into mice grafted with HPV16-E7-expressing</p>
PubMed By PMIDAbstract3999.txt volume regulation of cervical cancer cells. On of RVD in cervical cancer cells, while	
PubMed By PMIDAbstract664.txt , and cervical cancer . Except for	
PubMed By PMIDAbstract10702.txt all women like cervical cancer , which might	
PubMed By PMIDAbstract11522.txt menstruating women with cervical cancer in situ showing	
PubMed By PMIDAbstract11796.txt management of invasive cervical cancer becomes even more	

1-20 of 77

What diseases are in these documents?

GATE Prospector

Search

Diseases Pathogens Pathogenesis Vaccine Animals and Models Custom Mimir Query

URI:

Disease Bacterial_disease Burkitts_lymphoma
 Viral_disease Cervical_cancer
 Chandipura_encephalitis

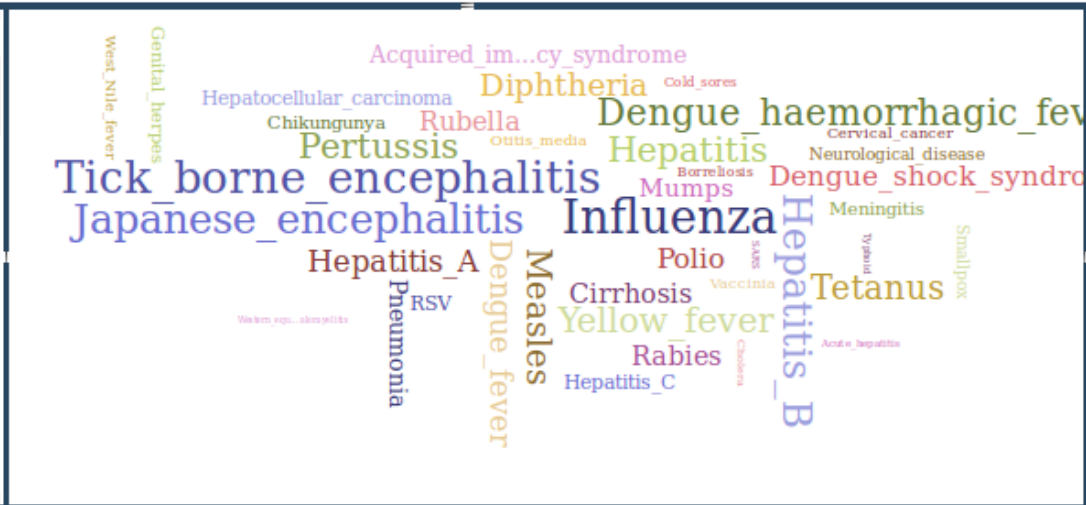
Dates: to
source:

Search

Documents **Terms**

Select top terms of type from the top retrieved documents.

Term	Count
Influenza	6,683
Tick_borne_encephaliti	5,501
Japanese_encephalitis	3,743
Hepatitis_B	3,201
Dengue_haemorrhagic_	1,944
Pertussis	1,852
Hepatitis	1,588
Yellow_fever	1,580
Measles	1,428
Tetanus	1,344



this as a term set named

Saved term sets

- {Disease} (40)

What Pathogens?

GATE Prospector

Search ▾

Diseases **Pathogens** Pathogenesis Vaccine Animals and Models Custom Mimir Query

URI:

- Pathogen ▾
 - Bacteria ▾
 - Virus ▾

Document Metadata

Dates: to

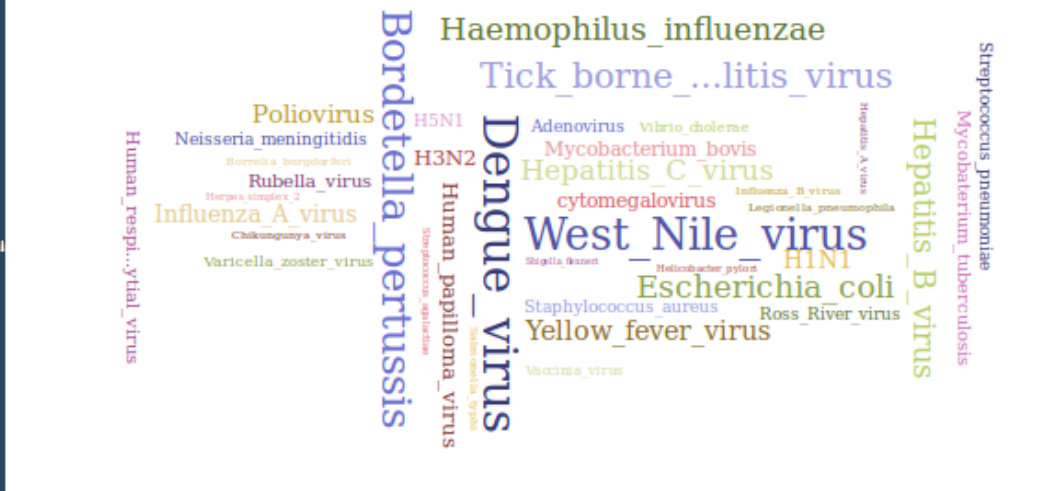
source: ▾

Search

Documents **Terms**

Select top 40 ▾ terms of type {Pathogen} ▾ from the top <All> ▾ retrieved documents.

Term	Count
Dengue_virus	17,867
West_Nile_virus	11,726
Bordetella_pertussis	9,909
Japanese_encephalitis_vi	5,093
Human_immunodeficien	4,939
Tick_borne_encephalitis	3,899
Haemophilus_influenzae	2,709
Escherichia_coli	2,342
Hepatitis_B_virus	2,043
Hepatitis_C_virus	1,567

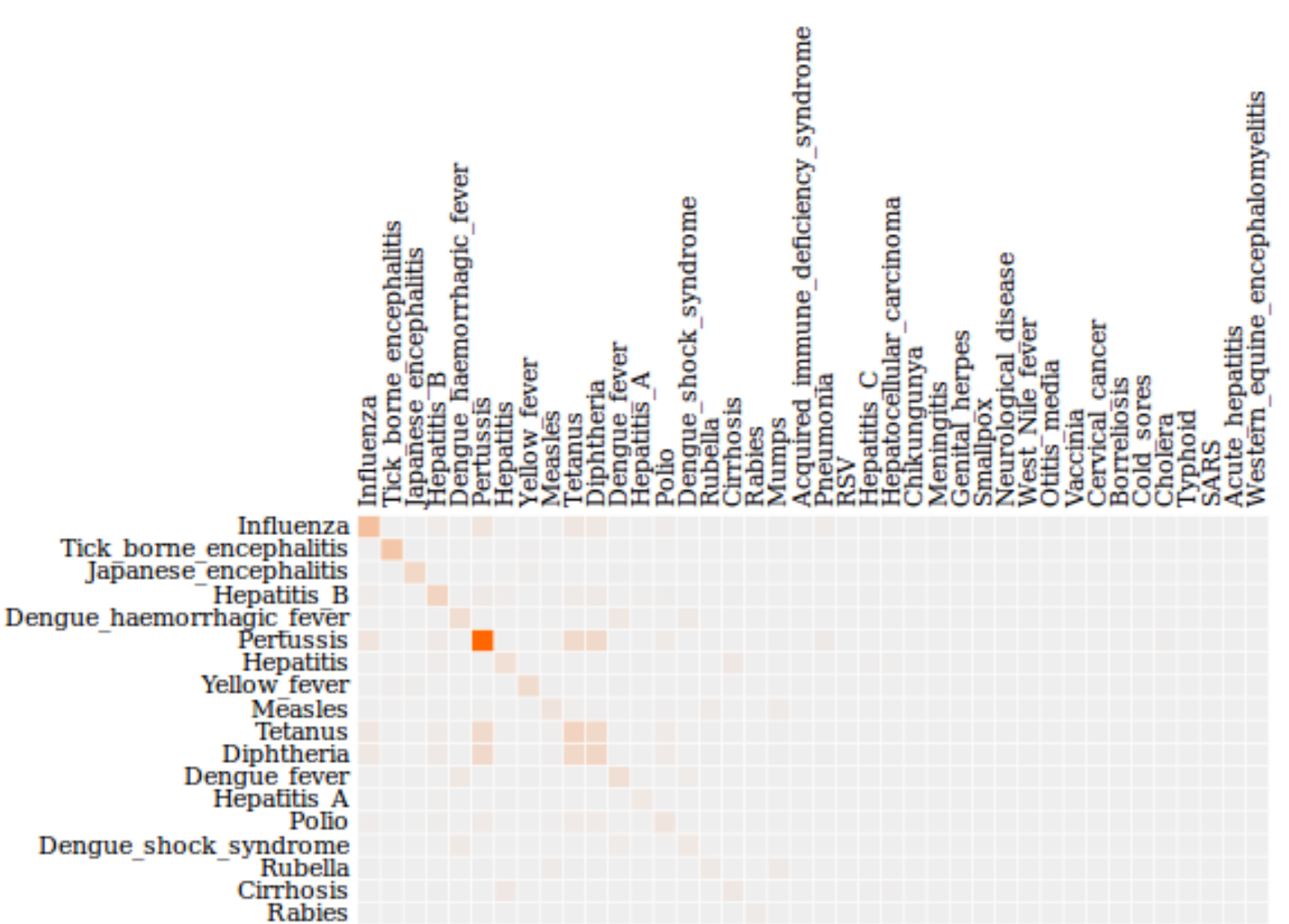


Save this as a term set named

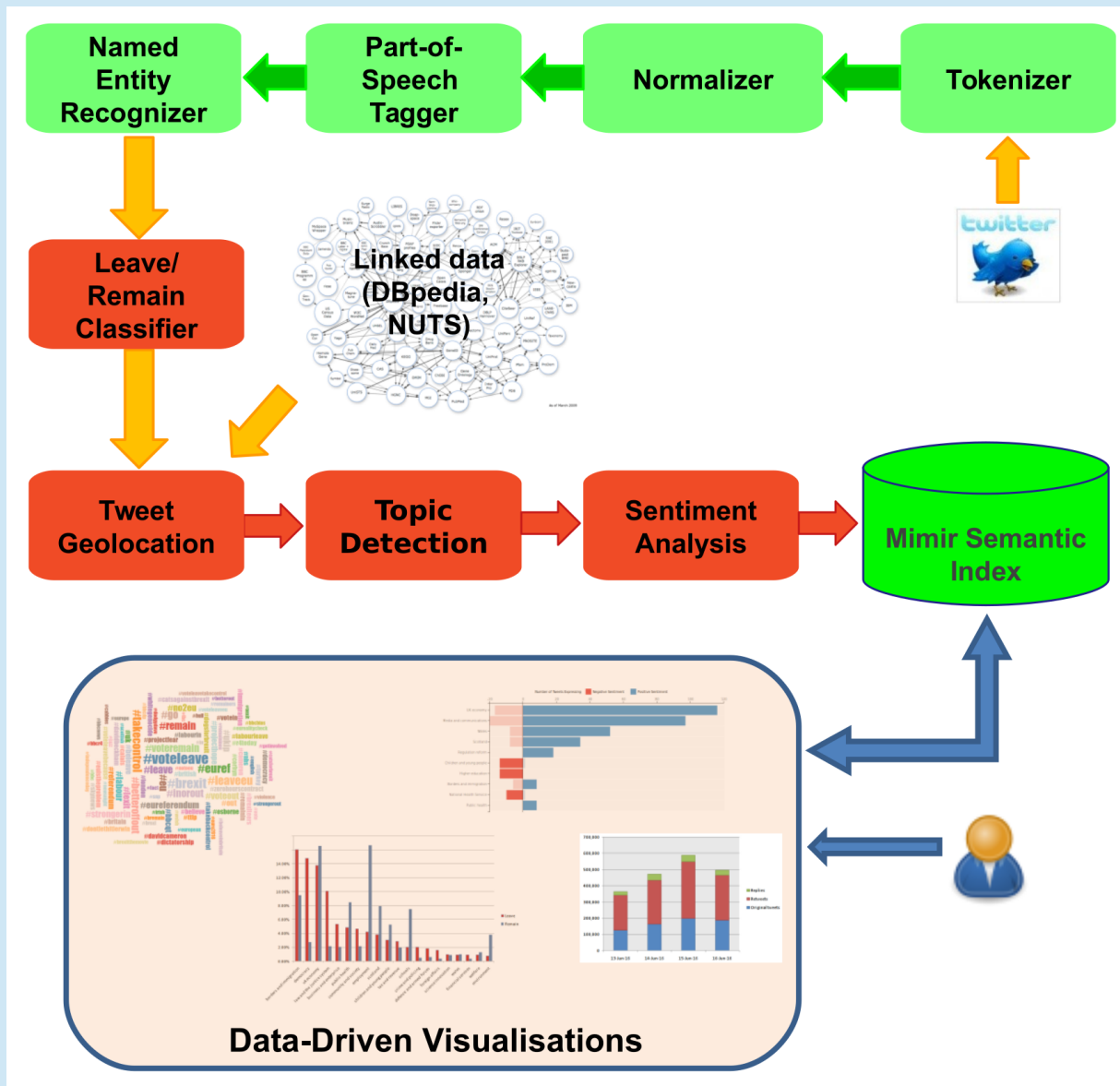
Saved term sets

- {Disease} (40) ✕

Disease vs Disease Co-occurrences



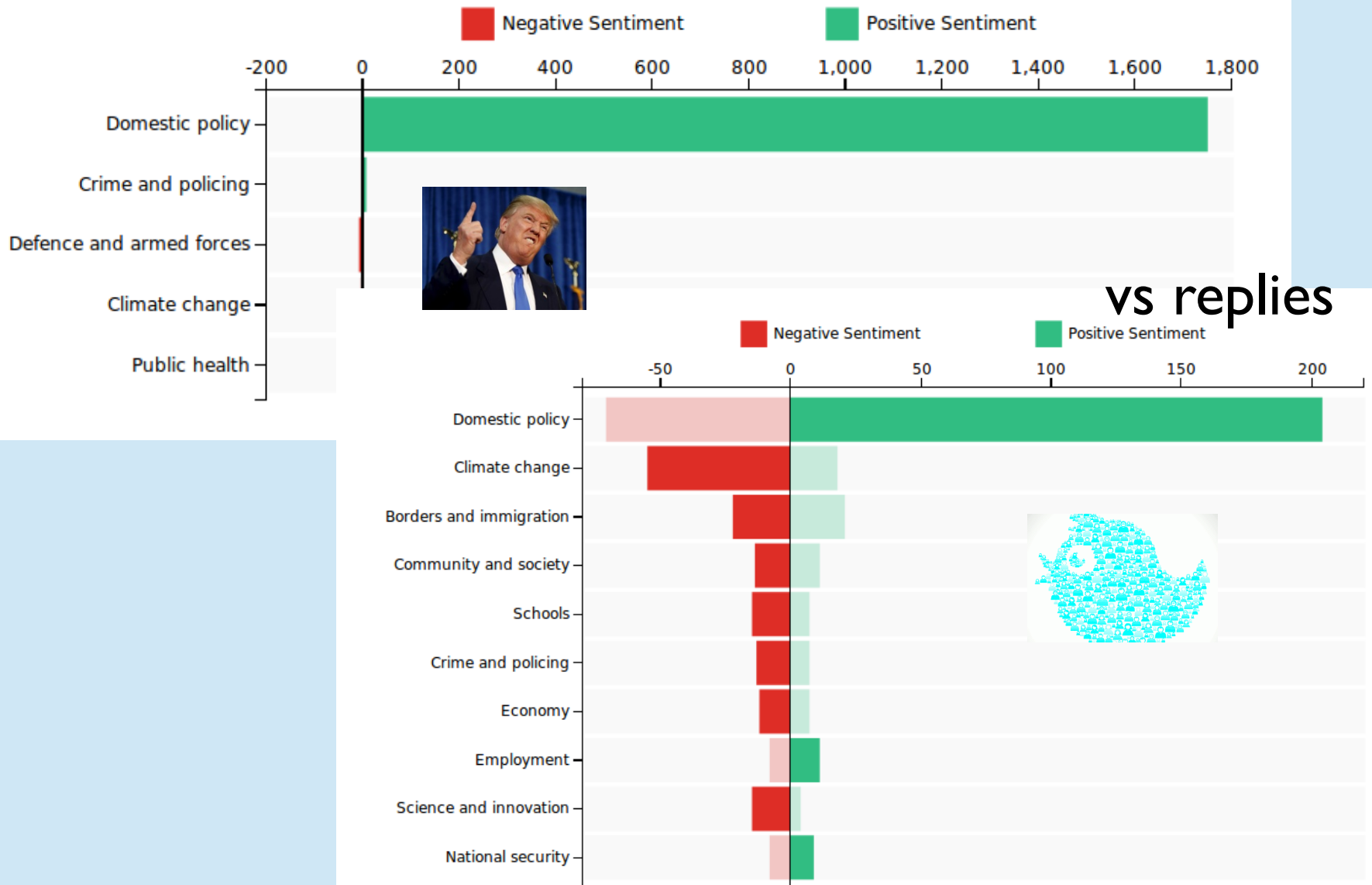
Some cool applications with GATE



Back to Trump (there's no escape)



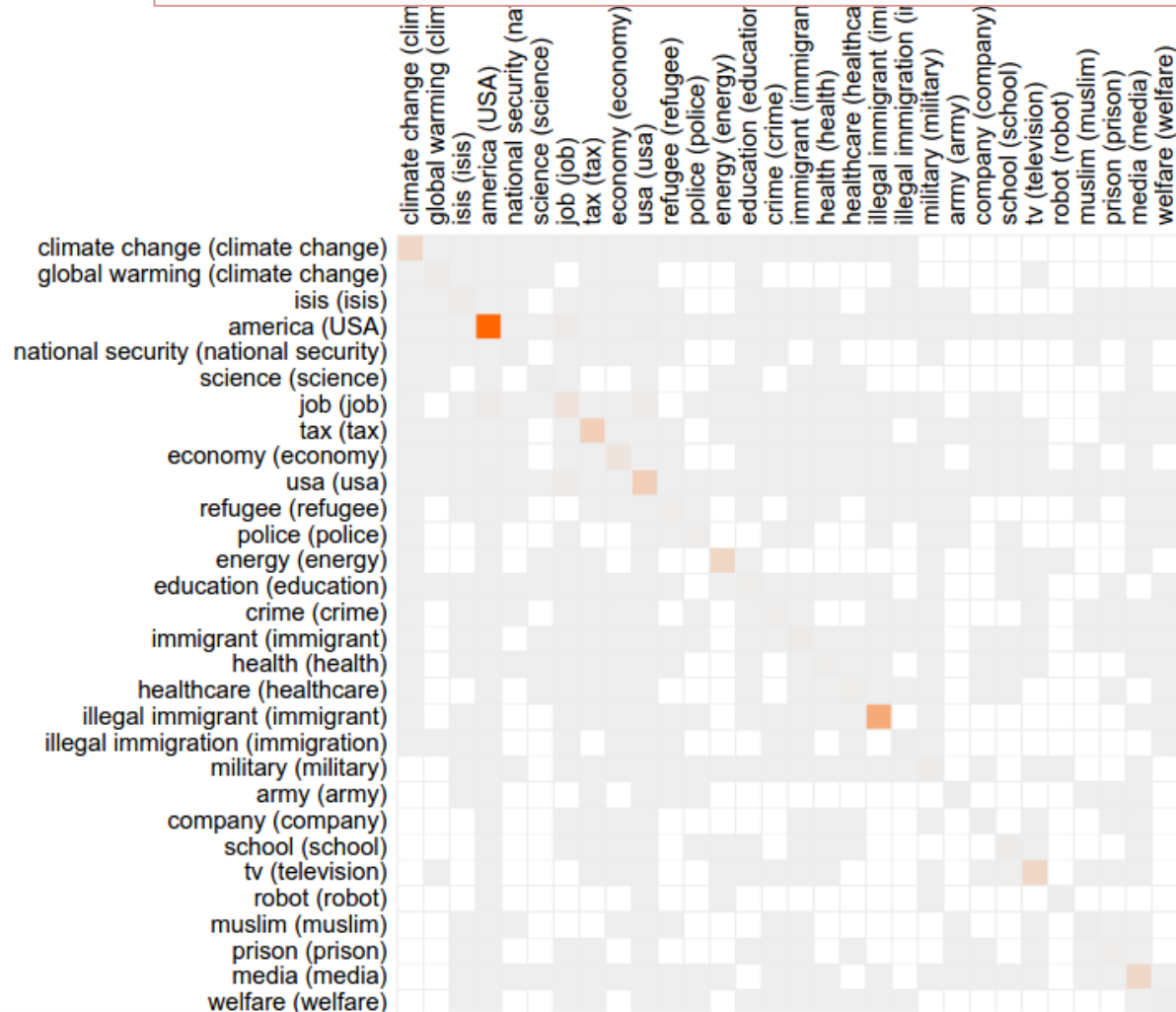
Real-time Opinion Monitoring



Climate change, ISIS and Trump

@realDonaldTrump Someone needs to tell Putin Isis and China to beware, the global warming is coming. That will stop them. Not.

@realDonaldTrump WHY IS EVERYONE IN THIS DEBATE BLAMING GLOBAL WARMING!?!?
WHAT DOES THAT HAVE TO DO WITH ISIS?!?!?



Querying election data with MIMIR

- Dataset: every tweet by MP / Candidate / Party, plus all replies/retweets
- Find all tweets where a Conservative MP talked about the economy

Searching Index "2015-03-09"

```
{DocumentAuthor author_party="Conservative Party"} OVER  
{Topic theme="uk_economy"}
```

Search



Richard Short
@ToryShorty

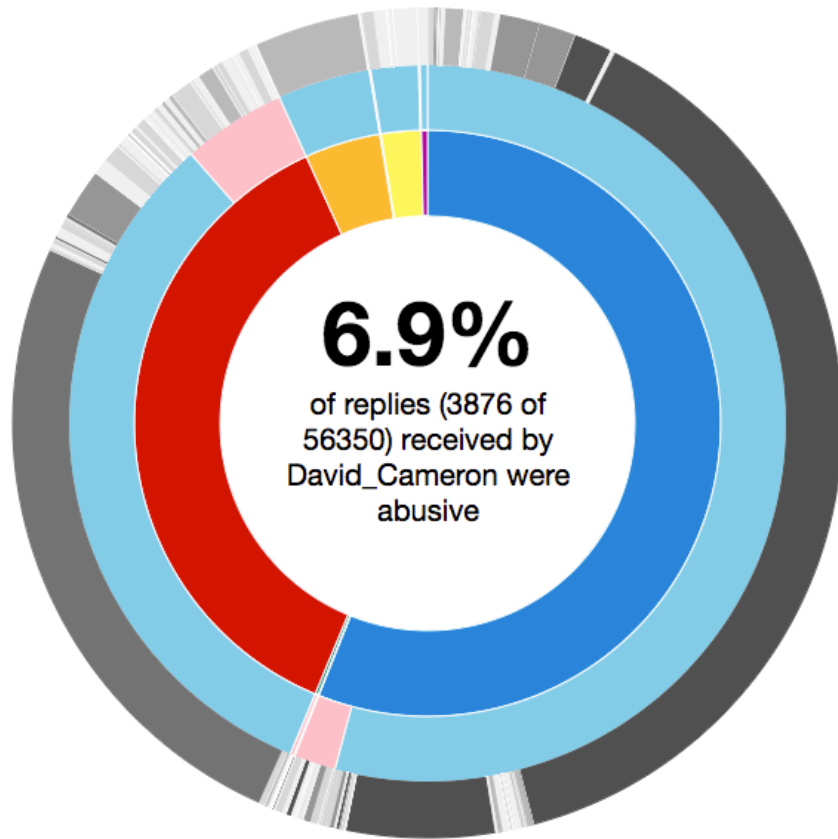
With so many protected from Labour's pension raid are they sure it will even generate £2.7bn [#bbcsp](#)

Parties / themes co-occurrence

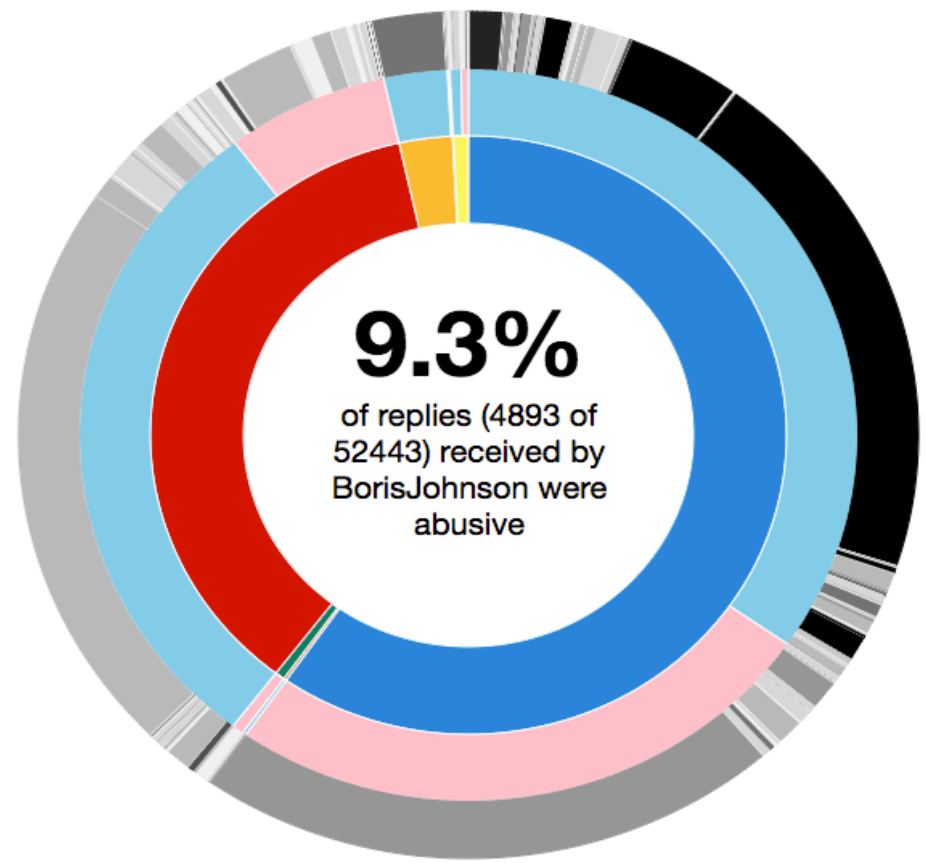


Hate speech towards MPs on Twitter

2015



2017



Environmental behaviour analysis

- Based on the assumption that users in different behavioural stages communicate differently (different emotions, directives, etc.)



Pajarito @lindopajarito . 2h

Our building needs 40% of all energy consumed in Switzerland! ☹️

Desirability: Negative sentiment (expressing personal frustration- anger/sadness)



DJPajarito @DJPajaritoGenial . 12h

I'm so proud when I remember to save energy and I know however small it's helping.

Buzz: Positive sentiment (happiness/joy). I/we + present tense



HotelPajarito @HotelPajarito . 18h

Join us today today to switch of a light for EH! 😊

Invitation: Positive sentiment (happy) + use of vocatives


Recognition of environmental terms in Decarbonet

what is global warming #global warming causes #global warming effects.





Instance	http://reegle.info/glossary/1062
majorType	climate
minorType	reegle-alt
prefLabel	anthropogenic climate change causes
rule	ReegleAlt
string	global warming causes

A positive tweet

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text 

Life flourishing in Antarctic fjords despite climate change <http://t.co/7TWq00>


SentenceSentiment

<input type="radio"/>	comment			<input type="checkbox"/>
<input checked="" type="radio"/>	entity_string	Antarctic		<input type="checkbox"/>
<input checked="" type="radio"/>	polarity	positive		<input type="checkbox"/>
<input checked="" type="radio"/>	rule	SentenceEntitySentiment		<input type="checkbox"/>
<input type="radio"/>	sarcasm	no		<input type="checkbox"/>
<input checked="" type="radio"/>	score	0.5		<input type="checkbox"/>
<input checked="" type="radio"/>	sentiment_string	flourishing		<input type="checkbox"/>
<input checked="" type="radio"/>				<input type="checkbox"/>




Type Open Search & Annotate tool

SentenceSen =positive, rule=Sentence

A negative tweet

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text 

Parody video exposes P&G's use of dirty palm oil <http://t.co/Pnb3eLmTHE> #Climate #Change #Solar


◀ ▶  ◀ ▶  

SentenceSentiment ▼






<input type="radio"/> comment	▼	▼	✗	
<input checked="" type="radio"/> polarity	▼	negative	▼	✗
<input checked="" type="radio"/> rule	▼	SentenceSentiment	▼	✗
<input type="radio"/> sarcasm	▼	no	▼	✗
<input checked="" type="radio"/> score	▼	-0.5	▼	✗
<input checked="" type="radio"/> sentiment_string	▼	dirty	▼	✗
<input checked="" type="radio"/>	▼	▼	▼	✗

▶ Open Search & Annotate tool

A sarcastic tweet

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text 

Nice to know the people in charge, who have so much power, are making smart decisions... #sarcasm #climatechange
[http://www.theguardian.com/environment/planet-oz/2014/feb/24/climate-change-dick-warburton-sceptic-australia-renewable-energy-target-review?CMP=tw_tfd ...](http://www.theguardian.com/environment/planet-oz/2014/feb/24/climate-change-dick-warburton-sceptic-australia-renewable-energy-target-review?CMP=tw_tfd)

SentenceSentiment

<input type="radio"/>	comment			<input type="checkbox"/>
<input checked="" type="radio"/>	polarity	negative		<input type="checkbox"/>
<input checked="" type="radio"/>	rule	SentenceEntitySentiment		<input type="checkbox"/>
<input type="radio"/>	sarcasm	yes		<input type="checkbox"/>
<input checked="" type="radio"/>	score	-0.5		<input type="checkbox"/>
<input checked="" type="radio"/>	sentiment_string	Nice		<input type="checkbox"/>
<input checked="" type="radio"/>				<input type="checkbox"/>

Features

rule-SentenceEntitySentiment sarcasm=yes score

Term recognition and sentiment analysis in Decarbonet

Demo

This is a simple demo of the main web service API, with the JSON output translated into HTML. Please type or paste some text into the box below.

Text to Process:

Cars pollute by emitting carbon dioxide #climatechange

↓ Process Text ↓

Google Chrome

Cars pollute by emitting carbon dioxide #climatechange

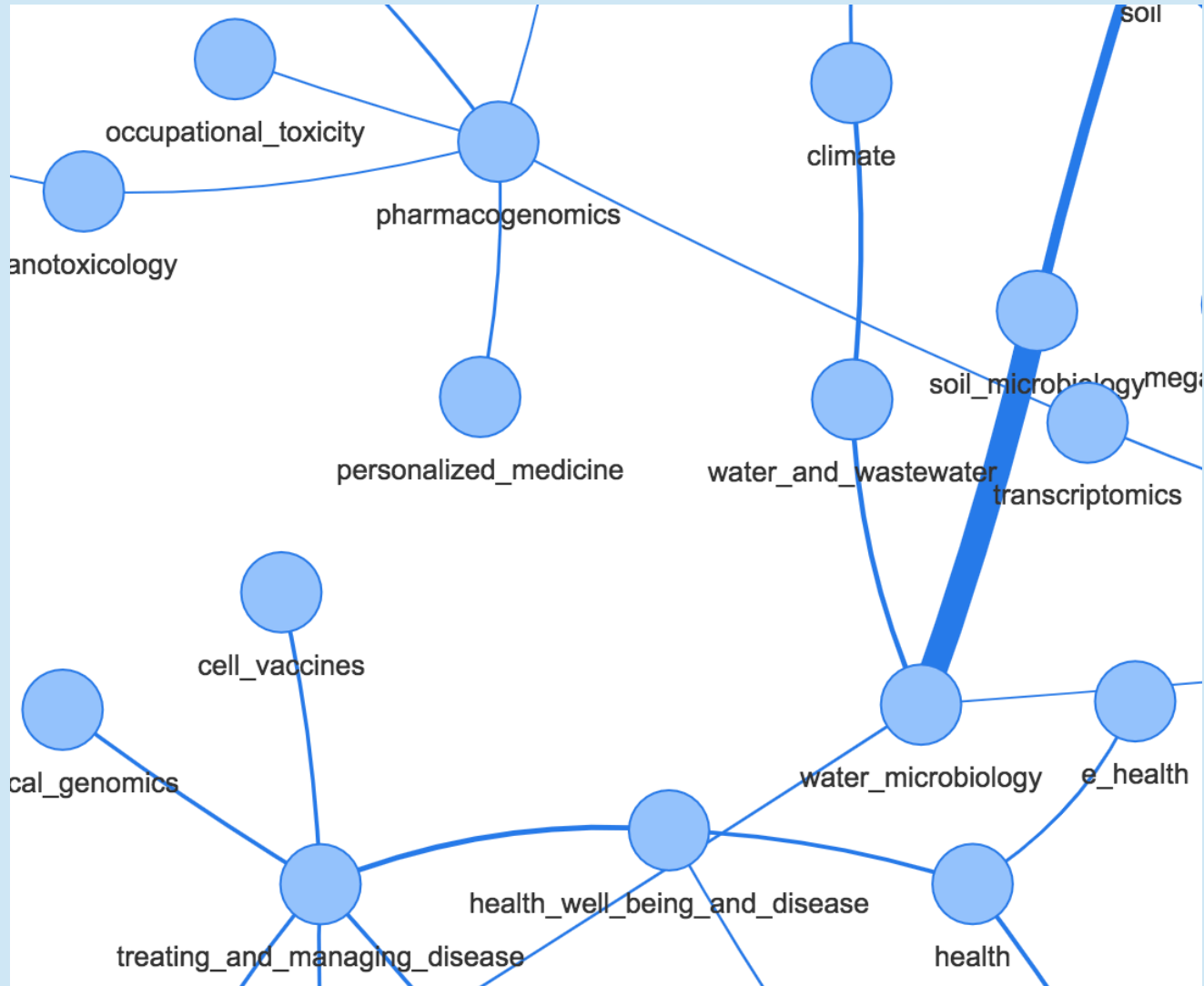
- <http://services.gate.ac.uk/decarbonet/sentiment>

KNOWMAK:

Mapping the state of European research

- Build an ontology to map between user queries (who's doing what where?) and databases of projects, patents, publications
- Build indicators on the data (how many patents by which actors in which country?)
- Build visualisation tools to show the results
- Searching and mapping are done via keywords mapped to topics (ontology classes) based around KETs and SGCs
- This deals with the problem that the terms used can vary widely between users and between document types
- [Ontology search demo](#)

Topic visualisation



<http://www.dcs.shef.ac.uk/~adam/stuff/knowmak/visualization/>

FILTER

Indicator

Select indicator ▾

Topic

Select topic ▾

Region

Select region ▾

Search for any of the above...

Years

2015 - 2015

clear all Selected

Austria x Germany x Italy x

Patent applications x Publications x

Pub top 10% cited x Cit. score NORM x



DATA

Download: selected data · visualised data · map

	Austria	Italy	Germany
↻ Patent applications – 2015	36	96	74
Publications – 2015	111	444	222
Publications in the top 10% cited – 2015	8	10	50
Citation score – 2015	200	400	100
● Citation score normalized – 2015	300	500	80

Summary

- Text mining is a very useful pre-requisite for doing all kinds of more interesting things, like semantic search
- Semantic web search allows you to do much more interesting kinds of search than a standard text-based search
- Text mining is hard, so it won't always be correct
- This is especially true on lower quality text such as social media
- We run an annual GATE training course in June in Sheffield, where you can spend a whole week learning all this and more!

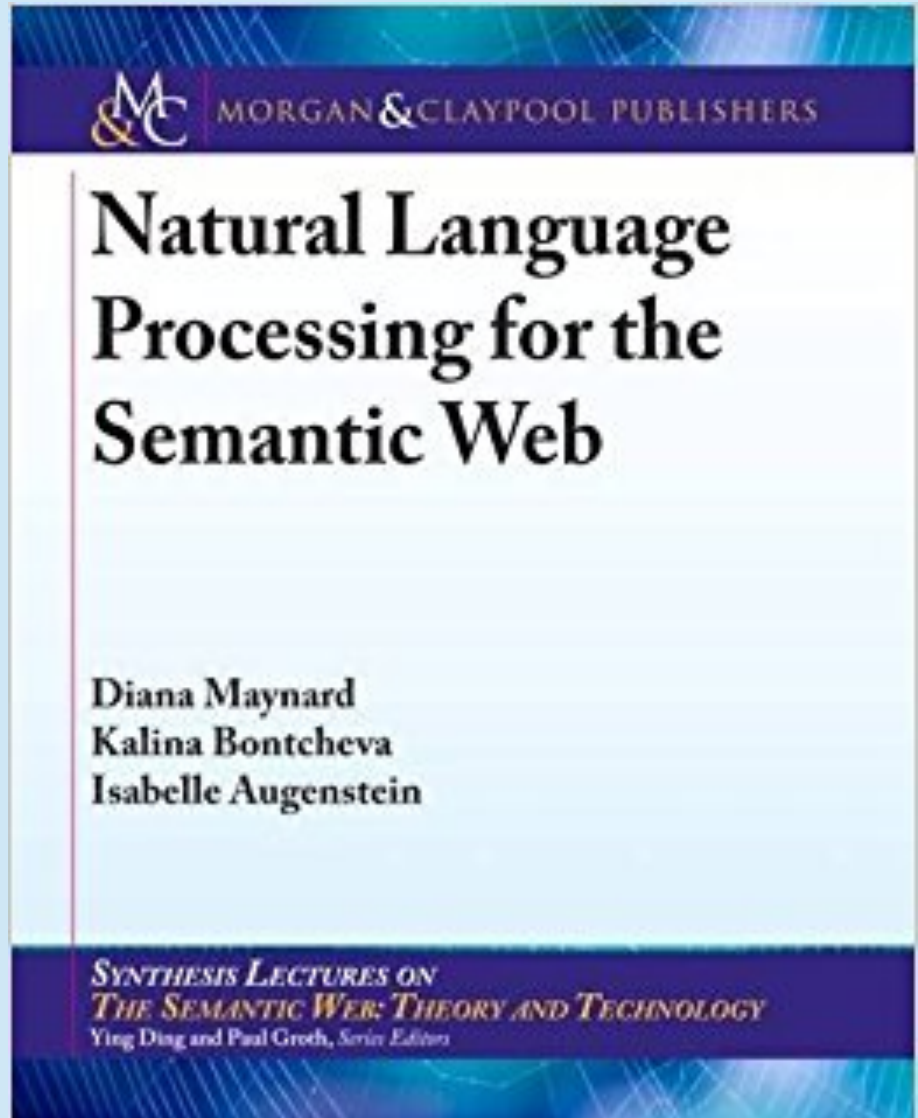
Acknowledgements

This work is supported by:

- the European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework and H2020 Programmes for R&D
 - DecarboNet (610829) <http://www.decarbonet.eu>
 - SoBigData (654024) <http://www.sobigdata.eu>
 - COMRADES (687847) <http://www.comrades-project.eu>
 - KNOWMAK (726992) <http://knowmak.eu>
- Nesta <http://nesta.org.uk>

What to ask Santa for this Christmas

- Great basic intro to NLP
- Uses GATE as examples
- Discusses other tools and the differences between them
- Chapters on semantic search, social media analysis, sentiment analysis, cool applications, and more



Key Publications

- Semantic Search over Documents and Ontologies (2014) K Bontcheva, V Tablan, H Cunningham. Bridging Between Information Retrieval and Databases, 31-53
- D. Maynard, I. Roberts, M. A. Greenwood, D. Rout and K. Bontcheva. A Framework for Real-time Semantic Social Media Analysis. Web Semantics: Science, Services and Agents on the World Wide Web, 2017
- V. Tablan, I. Roberts, H. Cunningham, and K. Bontcheva. GATECloud.net: a Platform for Large-Scale, Open-Source Text Processing on the Cloud. [Philosophical Transactions of the Royal Society A](#), 371(1983), 2013
- More papers on the GATE website: <http://gate.ac.uk/gate/doc/papers.html>

Some useful links

- GATE: <http://gate.ac.uk>
- GateCloud: <https://cloud.gate.ac.uk>
- Annual GATE training course in June in Sheffield: <https://gate.ac.uk/family/training.html>
- Download GATE: <http://gate.ac.uk/download>
- GATE blog posts on social media analysis: <http://gate4ugc.blogspot.co.uk/>
- UK elections monitor <http://gate.ac.uk/projects/pft>
- [Blog post on abuse of MPs:](#)
- COMRADES project on disasters: <http://gate.ac.uk/projects/comrades>
- KNOWMAK project and demos: <http://gate.ac.uk/projects/knowmak>
- SoBigData project: <http://sobigdata.eu>

Has your head exploded yet?



Questions?