Information Extraction and Ethics

Natural Language Processing Module 2019 (Dr N Aletras)

Prof Jochen L Leidner, MA MPhil PhD FRGS
⟨leidner@acm.org⟩

Objectives

In this session, we aim to:

- get introduced to information extraction, its concepts and history
- learn about rule based and learning approaches to IE

Outline

- Motivation
- Background: Information Extraction
- Key Concepts
- Architecture
- Methods
  - Rule-based (Example: GATE/JAPE)
  - Machine learning based (Example: Nymble)
- Annotation
- Feature extraction
- Applications
- From named entity tagging to event extraction

"Information Extraction" Defined

Information Extraction (IE) is

- the extraction of structured (relational) data from unstructured (= textual) sources
- a practically-motivated engineering discipline (models not necessarily inspired by nature)
- the use of natural language processing techniques to populate the slots of structured templates with appropriate fillers

IE was conceived as a shortcut to built useful systems when full text understanding based on syntactic parsing was beyond the state of the art.

IE Example

"Concepcion, 23 Aug 88 (Santiago Domestic Service) – Police sources have reported that unidentified individuals planted a bomb in front of a Mormon Church in Talcahuano District. The bomb, which exploded and caused property damage worth 50,000 pesos, was placed at a chapel of the Church of Jesus Christ of Latter-Day Saints located at No 3856 Gomez Carreno Street.
The shock wave destroyed a wall, the roof, and the windows of the church, but did not cause any injuries.
Carabineros bomb squad personnel immediately went to the location and discovered that the bomb was made of 50 grams of an-fo ammonium nitrate-fuel oil blasting agents and a slow fuse."

IE Example: Terrorist Attack Event

Concepcion, **23 Aug 88** "(Santiago Domestic Service) – Police sources have reported that **unidentified individuals** planted a **bomb** in front of **a Mormon Church** in **Talcahuano District**.
The bomb, which exploded and caused **property damage worth 50,000 pesos**, was placed at a chapel of the Church of Jesus Christ of Latter-Day Saints located at No 3856 Gomez Carreno Street.
The shock wave destroyed a wall, the roof, and the windows of the church, but **did not cause any injuries**.
Carabineros bomb squad personnel immediately went to the location and discovered that the bomb was made of 50 grams of an-fo ammonium nitrate-fuel oil blasting agents and a slow fuse."
— MUC4 story TST4-MUC4-0001

The "Terrorist Attack" Template

Terrorist Attack	
Type of Attack:	________________
Perpetrator:	________________
Target:	________________
Location:	________________
Time:	________________
Casualties:	________________
Injured:	________________
Material Damage:	________________

The "Terrorist Attack" Template

Terrorist Attack	
Type of Attack:	`bomb` (ATTACK>BOMBING)
Perpetrator:	`unidentified individuals` (unknown)
Target:	`a Mormon church`
Location:	`Talcahuano District` (Chile>Talcahuano)
Time:	`23 Aug 88` (1988-08-23 00:00:00)
Casualties:	______ (0)
Injured:	`did not cause any injuries` (0)
Material Damage:	`property worth 50,000 pesos` (50000)

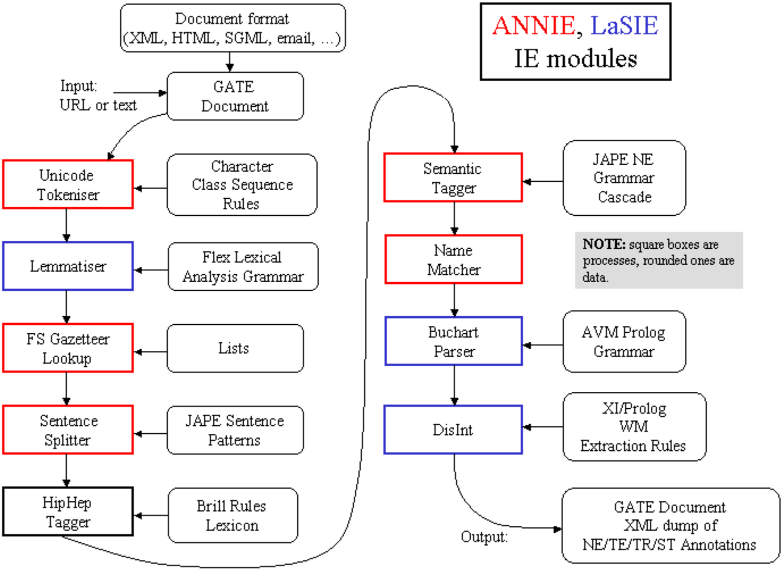
A Short History of Information Extraction

- 1981: NYU "'Linguistic String" project (N. Sager)
- 1982: DeJong's FRUMP system: 'sketchy scripts'
- Carnegie Group build JASPER IE system for Reuters (Andersen et al., 1986)
- 1987/1989: MUCK I+II: Naval operations messages
- 1991-1998 MUC 3-7: Message Understanding Contest
- 2000-2004: ACE: Automatic Content Extraction
  - from text spans to abstract entities
  - English, Chinese, Arabic
- 2010s: First neural approaches to IE

Information Extraction System Architecture

- **Preprocessor/Tokenizer**: split story into units and ultimately word tokens
- **Gazetteer**: lexical look-up of important (to your task) words/phrases
- **POS tagger**: tag/disambiguate words w.r.t. parts of speech
- **Chunk parser**: find basic noun and verb phrases
- **Named entity tagger**: identify and classify proper names
- **Relationship tagger**: find relations between entities
- **Event Template Analyzer**: populate fact/event templates
- The result is a **structured fact/event database**

IE System Architecture – Example



Fundamental Methods Used for IE

- **Rule-based**: human experts (computational linguists) manually write general linguistic rules and task-specific extraction rules.
  Trigger keywords, regular expressions, pattern/action, (cascaded) Finite State Transducers (FSTs)
- Supervised **Machine learning-based**: humans (domain experts) manually annotate text spans indicating entities, relations, facts etc. in a training corpus, and an expert (computational linguist) formulates a set of features; these get used to extract information if statistically correlated with classes of entities, relations etc. sought.
- Insight: shallow processing works well: SRI Tacitus → SRI FASTUS (Hobbs et al. 1992)

Example Patterns (FASTUS, Appelt et al., 1993)

```
killing of <HumanTarget>
<GovtOfficial> accused <PerpOrg>
bomb was placed by <Perp>
on <PhysicalTarget>
<Perp> attacked <HumanTarget>
<PhysicalTarget> with <Device>
<HumanTarget> was injured
<HumanTarget>'s body
```

From Entity over Relation to Scenario Template

- **Template Element Recognition (TE)**: extract information pertaining to organizations, persons and artifacts (NE tagging, parsing)
- **Relationship Extraction (RE)**: extract information about how individual entities stand in relationship to each other, drawing on a pre-defined inventory of relation types
- **Scenario Template Recognition (ST)**: extract pre-specified event information and relate the event to particular organizations, persons and/or artifacts (slot fillers)
- Information for filling a template often often spread across several sentences

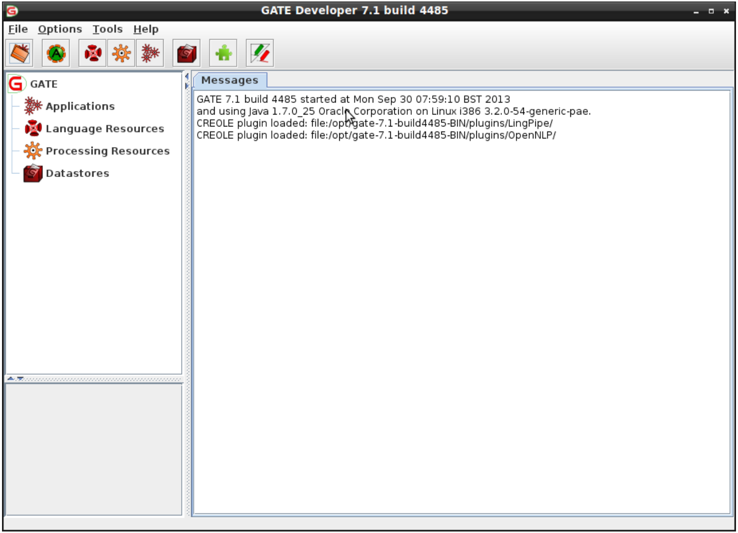
Experimental Methodology

- Create a **gold data** set reference corpus with **ground truth**)
- Split gold data into three parts:
  - **development/training set**: used to study the data, train machine learning processes; can be inspected
  - **development test** ("devtest") set: cannot be inspected, cannot be used for training; repeatedly used to measure improvements of system quality by comparing system output with ground truth.
  - **test set**: cannot be inspected; only used once for final evaluation run at project end. Completely **unseen data** (to the system and developers).
- Gold data split:
  could be e.g. 80% train : 10% dev-test : 10% test

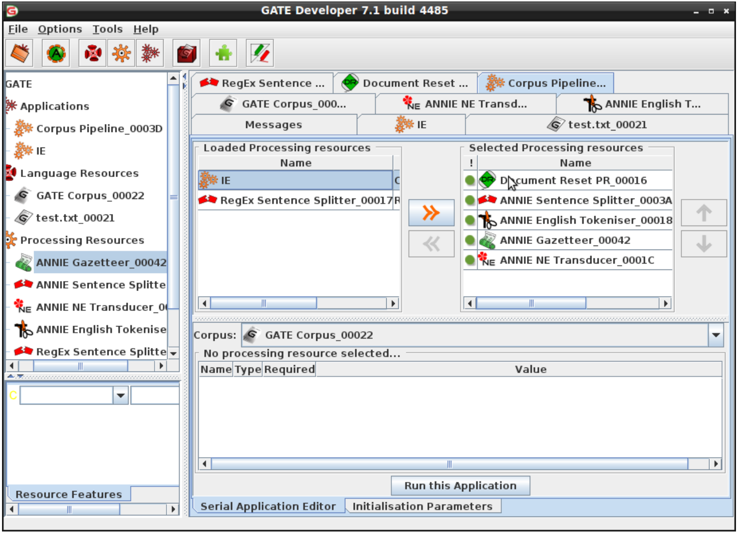
The GATE framework: From GUI to Guts

- GATE: General Architecture for Text Engineering - an open-source framework for *language engineering*
- Under development at the University of Sheffield for a long time (Cunningham et al., 2013)
- Current version: 8.4.2 (as of today) available from `http://gate.ac.uk`
- Java-based platform comprising GUI, APIs, and a workflow system
- GATE comes with pre-existing, contributed data (*Language Resources*) and components (*Processing Resources*)

Constructing IE Pipelines with GATE (1/2)



Constructing IE Pipelines with GATE (2/2)



JAPE Introduction

- Part of the GATE platform
- Language to specify FSTs following the "patterns & actions" paradigm
- Sets of rules, broken down into processing phases
- Each rule tests matching conditions and typically adds annotations in the affirmative case

JAPE Rules – Regular Expressions over Annotations

Jape rules are organized into phases (cascades).
Each JAPE rule has three parts:

- **Header**: name of the rule
- **Left-Hand Side (LHS)**: pattern
- **Right-Hand Side (RHS)**: action – e.g. add annotations
- Structure:

```
Rule: TagUnknownName
(
 {Token.category == NNP}
):x
-->
 :x.Unknown = { kind = "PN", rule = TagUnknownName }
```

- See also: `http://gate.ac.uk/sale/tao/splitch8.html#chap:jape`

JAPE Rule Format – Example 1: URL Prefix

```
Phase: UrlPre
Input:  Token SpaceToken
/* important: specify input! */
Options: control = appelt

Rule: Urlpre
( ((({Token.string == "http"} |
    {Token.string == "ftp"})
   {Token.string == ":"}
   {Token.string == "/"}
   {Token.string == "/"}
  ) |
  ({Token.string == "www"}
   {Token.string == "."}
  )
):urlpre
-->
 :urlpre.UrlPre = {rule = "UrlPre"}
```

Gazetteers in GATE

How to use lists of keywords/phrases:

- Create a `*.def` file listing all gazetteers (with their major/minor categories):

```
cities.lst:location:city
organizations.lst:organization
surnames.lst:name:surname
forenames.lst:name:forename
```

- List one key phrase, name or keyword per lin in each of these files, e.g. in `cities.lst`:

```
Abu Dhabi
Berlin
Chicago
Frankfurt
...
```

- Create a gazetteer Processing Resource in your application pipeline that references your `*.def` file.

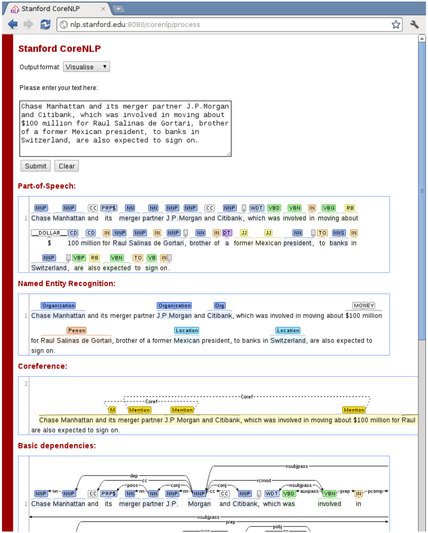
Common Machine Learning Methods for Information Extraction

- **Hidden Markov Models (HMMs)**
- Conditional Random Fields (CRF)
- Support Vector Machine (SVM)
- Artificial Neural Networks (NNs), in particular "deep" neural nets for sequence tagging (RNN, LSTM)

BIO Encoding for Labels Classifying Text Spans (CoNLL)

- Annotate word tokens with inside/outside of class information:
  `The_O Oracle_I-ORG CEO_O 's_O name_O`
  `is_O Larry_I-PER Ellison_I-PER ._O`
  (horizonal)
  or:
  ```
  The O
  Oracle I-ORG
  CEO O
  ...
  ```
  (vertical)
- B tag used to demarcate adjacent `I` tags (otherwise, two adjacent "I" tokens part of the same text span could not be distinguished from the case of two adjacent, separate text spans)

Annotation of Gold Data with BRAT (http://nlp.brat.org)



Feature Extraction – Spam Classification Example (1/2)

- **Feature**: a piece of evidence intended to help the classifier map the input to the right target class
- **Feature vector**: a vector $\vec{F}$, the components $F_j = \phi_j(d_i)$, of which are results applying a feature function to the data point d_i
- Example: "Spam versus Ham" email?
  - number of "!"s included in email body
  - length of the email in characters
  - does the word "cash" occur in the title or body?
- Example feature vectors:
  (2, 2392, no) $\mapsto$ `HAM` (genuine e-mail)
  (4, 520, yes) $\mapsto$ `SPAM`
  (1, 2392, no) $\mapsto$ `HAM`
  (0, 16337, no) $\mapsto$ `HAM`
  (1, 6i320, yes) $\mapsto$ `SPAM`

Feature Extraction – Features for Person NERC (2/2)

In English:

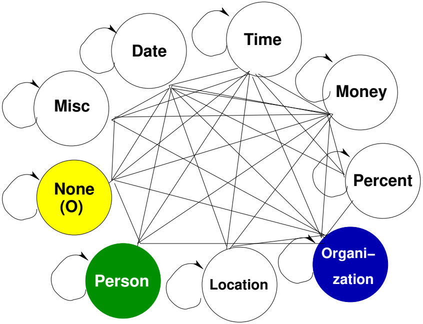
- Token begins with capital letter: `[A-Z][a-z]+`
- Token ends in characters `-man`
- Token to the left is from list Mr., Mrs., Miss, Dr., Prof., Sir, Lord, CEO, ...
- Token to the right is academic title, affiliation: BSc, Ph.D., M.A., FRS, M.P.
- Tokens to the right are `[,]? who`
- Token is from a list (gazetteer) of names

Training a Model

- Determine the model's parameters from data: induction
- Training corpus has counts, from which model probabilities are computed
- Smoothing: re-distribution of probability mass from seen to unseen events (to avoid zero probabilities, which make any product go zero)

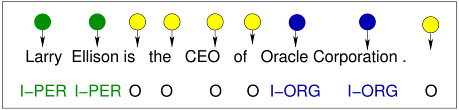
HMMs for IE e.g. in Nymble (Bikel et al., 1997)

HMM Transitions:



Not shown: initial tranistions/probabilities

Most likely state (label) sequence (Viterbi alg.):



Features in Nymble (Bikel et al., 1997)

Word Feature	Example Text	Intuition
`twoDigitNum`	90	Two-digit year
`fourDigitNum`	1990	Four digit year
`containsDigitAndAlpha`	A8956-67	Product code
`containsDigitAndDash`	09-96	Date
`containsDigitAndSlash`	11/9/89	Date
`containsDigitAndComma`	23,000.00	Monetary amount
`containsDigitAndPeriod`	1.00	Monetary amount, percentage
`otherNum`	456789	Other number
`allCaps`	BBN	Organization
`capPeriod`	M.	Person name initial
`firstWord`	*first word of sentence*	No useful capitalization information
`initCap`	Sally	Capitalized word
`lowerCase`	can	Uncapitalized word
`other`	,	Punctuation marks, all other words

Real Applications of IE – Some Use Cases

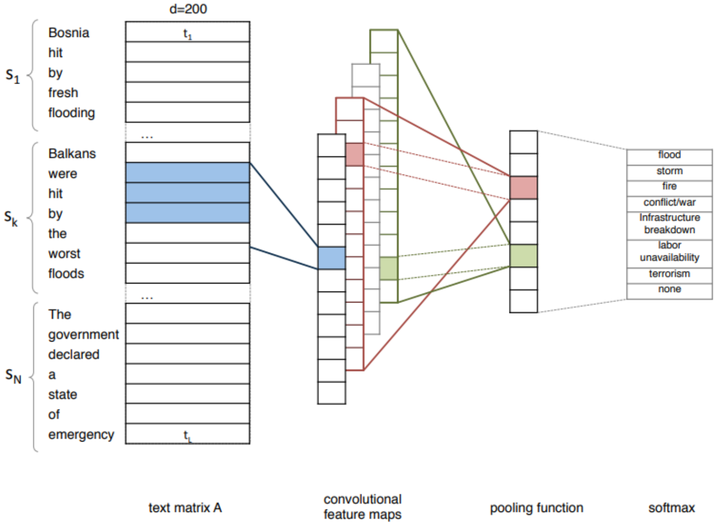
There are many application domains for IE systems:

- Bio-medical IE applications (genes & proteins)
- Financial IE applications (mergers & acquisitions, CEO changes))
- Legal IE applications (judges & attorneys)
- Intelligence & Police IE applications (terrorists & crimes)
- e-Commerce IE applications (brands & products)

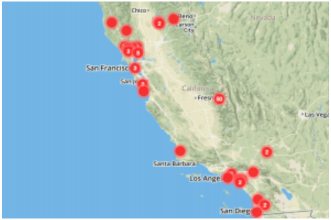
From NE tagging over Relation Extraction to Event Extraction

- (Named) Entity Tagging: extracting mentions of things (people, places, ...)
- Relation Extraction: extracting mentions of relationships between things (son-of, CEO-of, located-in, famous-for)
- Event Extraction: something happens (event of type T at time t and in location ℓ

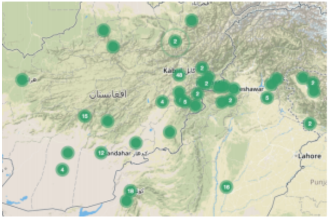
Event Extraction from News with Neural Models (Nugent et al., 2017) (1/2)



Event Extraction from News with Neural Models (Nugent et al., 2017) (2/2)



Wildfires in CA



Conflicts in Afghanistan

Current Developments & Future Directions in IE

- Utilizing robust syntactic and semantic components where available
- From classic IE to *open IE*
- From small data sets to Web-scale data sets (bigger data and a simpler algorithm usually beats smaller data and a more sophisticaed algorithm!)
- Knowledge-rich methods are going to be combined with machine learning methods
- Use of word embeddings and deep (= multiple hidden layers) neural networks,

Summary

In this session, we learned:

- what information extraction is, and a bit of its history
- the difference between rule-based and machine learning approaches
- how named entities can be extracted from text

References (1/2)

- Bird/Klein/Loper (2009), *Natural Language Processing with Python*, Sebastopol, CA: O'Reilly
- Costantino and Coletti (2008), *Information Extraction in Finance*, Southampton: WIT Press
- Cunningham et al. (2013), Developing Language Processing Components with GATE Version 7 (a User Guide), University of Sheffield, [online] `http://gate.ac.uk/sale/tao/`, Chapter 2-3, 5-6, and 10.
- Hastie, Tibshirani and Friedman (2009), *The Elements of Statistical Learning (2nd ed.)*, New York, NY: Springer
- Jurafsky/Martin (2008), *Speech and Language Procesing (2nd ed.)*, Upper Saddle River, NJ: Prentice Hall
- Manning/Schütze (1999) *Statistical Natural Language Processing* MIT Press.
- Moens (2006), *Information Extraction: Algorithms and Prospects in a Retrieval Context* Dordrecht: Springer

References (2/2)

- Nugent & Leidner (2016), "Risk Mining: Company-Risk Identification from Unstructured Sources" *Proc. ICDM*
- Plachouras, Leidner & Garrow (2016) "Quantifying Self-Reported Adverse Drug Events on Twitter: Signal and Topic Analysis" *Ann. Meeting of the Social Media Soc.*
- Pustejovsky/Stubbs (2012), Natural Language Annotation for Machine Learning, Sebastopol, CA: O'Reilly, Chapters 5-6 and 8
- Zanasi (2005), *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, Southampton: WITPress

Backup Slides

Some Well-Known IE Systems

SAM	PAM	FRUMP	ATRANS	Proteus
PLUM	Tabula Rasa	LOLITA	Scrabble	NameTag
Alembic	Hasten	NLToolset	IE^2	TurboTag
SIFT	TASC	FACILE	AutoSlog	ANNIE
LasIE	Elie	NetOwl	Oki	BALIE
Diderot	JASPER	Tacitus	FASTUS	SCISOR
Circus	REES	Identifinder	Nymble	InfoXtract

Homework: Constructing IE Pipelines with GATE (1/2)

- File→New Corpus Pipeline, give name e.g. TestIE
- Processing Resources→Add Annie
- Processing Resources→Add Document Reset, ANNIE Sentence Splitter, ANNIE English Tokeniser, ANNIE Gazetteer, and ANNIE NE Transducer.
- Double click on TestIE and move the components Document Reset, ANNIE Sentence Splitter, ANNIE English Tokeniser, ANNIE Gazetteer, and ANNIE NE Transducer from the left list to the right using ">>"; ensure the order is exactly as given here from top to bottom!
- Language Resources→Add Corpus Document, click on sourceURL and select a text file to process

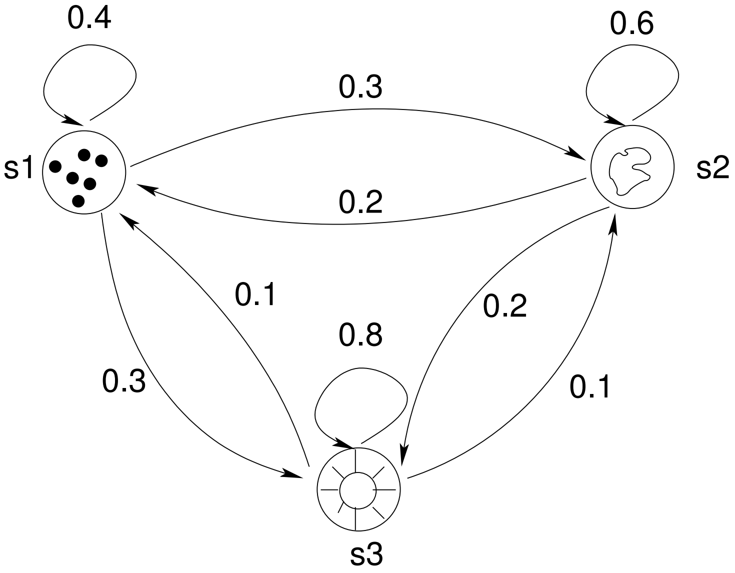
Homework: Constructing IE Pipelines with GATE (2/2)

- Language Resources→Add Corpus, then add document created above via drop-down menu to the empty corpus
- Double-click on TestIE Application, select corpus in drop-down and click Run this Application
- Double-click on your test document under Language Resources and select the annotations (PERSON, ORGANIZATION) you would like to view

Sequence Tagging

- It is one thing to classify a data point in isolation
  Example (1): "`dinosaur`" $\mapsto$ noun
- It is quite another thing to classify, taking into account *context*
  Example (2a): "`(let's) walk (there)`" $\mapsto$ `walk/VVB`
  Example (2b): "`(a) walk (in the rain)`" $\mapsto$ `walk/NN`
- If "walk" has two possible readings, how can we compute the right one? $\rightarrow$**Ambiguity**
- Example (3): `I can can the can .` $\mapsto$ `I/PRP can/AUX can/VVB the/AT can/NN ./.`

Markov Model – Introductory Example (1/2)



Markov Model – Introductory Example (2/2)

- Simple probabilistic finite-state model of the weather (**Markov Chain**)
- 3 States: s1: rainy, s2: cloudy, s3: sunny
- Transitions labelled with probabilities $a_{ij} \geq 0, \forall i, j \leq N$, $\sum_{j=1^N} a_{ij} = 1, \forall i$
- What is the probability of the observation sequence $O = (sunny, sunny, sunny, rainy, rainy, sunny, cloudy, sunny)$?

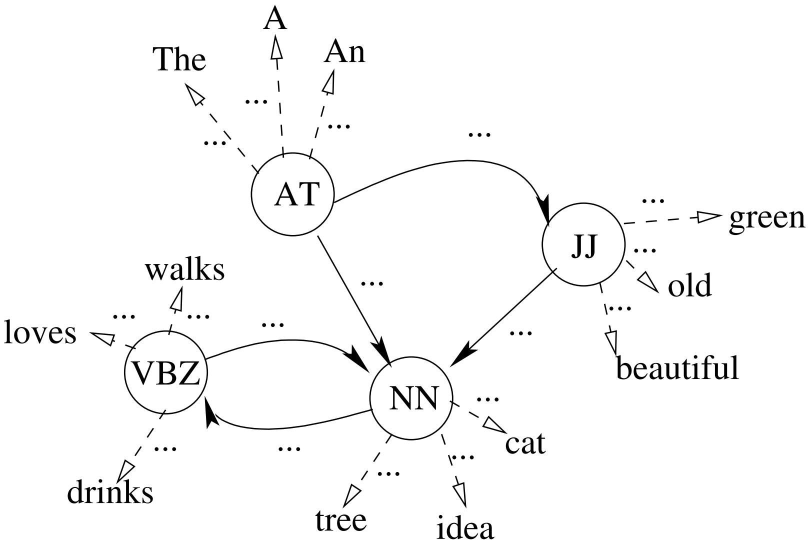
$$\begin{aligned}
P(O|Model) &= P[s3, s3, s3, s1, s1, s3, s2, s3|Model] \\
&= P[3]P[3|3]^2 P[1|3]P[1|1]P[3|1]P[2|3]P[3|2] \\
&= \pi_3 {a_{33}}^2 a_{31} q_{11} a_{11} a_{13} a_{32} a_{23} \\
&= 1.0 \cdot 0.8^2 \cdot 0.1 \cdot 0.4 \cdot 0.3 \cdot 0.1 \cdot 0.2 \\
&= 1.536 \times 10^{-4}
\end{aligned}$$

Hidden Markov Models (HMM)

A **Hidden Markov Model** $\lambda = (Q, \Sigma, A, B, \Pi)$ is formally defined as

- a set of N *hidden states* $Q = (q1, \ldots, q_N)$, $N = |Q|$
- a set Σ of M *observation symbols*, $M = |\Sigma|$
- a state-transition distribution (*transition probabilities*)
  $A = \{a_{ij} = P(q_{t+1} = j | q_t = i)\}, 1 \leq i, j \leq N$
- an observation symbol probability distribution (*emission probabilities*)
  $B = \{b_{ik} = b_o(o_k) = P(o_k | q_i)\}, 1 \leq k \leq M, o_k \in \Sigma$ (the probability that the output is o_k, given that the current state is q_i)
- an *initial state distribution*
  $\Pi = \{\pi_i = P(q_i | t = 1)\}, 1 \leq i \leq N$

Hidden Markov Models (HMM) – POS Tagging Example



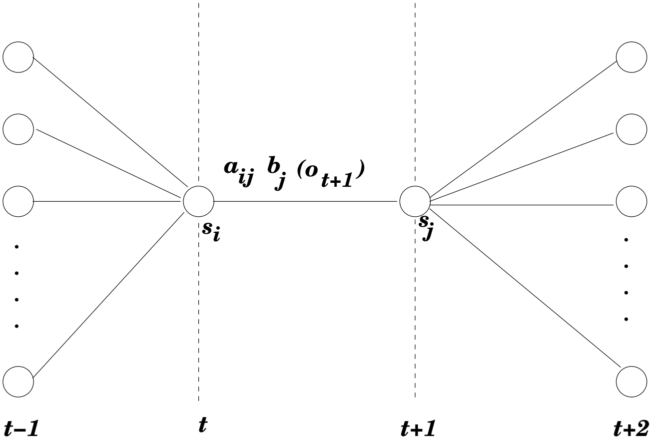
Three Questions to an HMM

- **Problem 1 (evaluation).** Given the observation sequence $O = (o_1, o_2, \ldots, o_T)$ and a model $\lambda = (Q, \Sigma, A, B, \Pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of a observation sequence, given a model?
- **Problem 2 (decoding).** Given the observation sequence observation sequence $O = (o_1, o_2, \ldots, o_T)$ and the model λ, how can we find a corresponding state sequence $q^* = q_1, q_2, \ldots, q_T$ that most likely generated O ("optimally explains the observations")?
- **Problem 3 (learning).** How do we estimate the model parameters A, B and Π so as to maximize $P(O|\lambda)$?

The Viterbi algortihm (1/2) (Viterbi, 1967)

- Finding the most likely hidden state sequence, given an observation, requires at first glance exponenial runtime complexity ($O(N^n)$, as all possible states are valid hypotheses for each observation, so candidate states/readings multiply)
- Luckily, we can find a linear-runtime ($O(n)$) solution using dynamic programming (remembering partial solutions at each step) in a data structure called **trellis**.
- The Viterbi algorithm uses one pass from left to right looking at initial, transition and emission probabilities while storing a history of the locally "best" (most likely) state (back pointer).

Trellis Data Structure



The Viterbi algortihm (2/2) (Viterbi, 1967)

Let $\delta_t(i) = \max_{q_1,\ldots q_{t-1}} P[q_1 q_2 \ldots q_{t-1,q_t=i}, o_1 o_2 \ldots o_t|\lambda]$.

1. Initialization. (Start with initial state probabilities)

$$\begin{aligned}\delta_1(i) &= \pi_i b_i(o_1), 1 \leq i \leq N \\ \phi_1(i) &= 0\end{aligned}$$

2. Recursion. (main part, see previous slide)

$$\begin{aligned}\delta_t(j) &= \max_{1\leq i\leq N}[\delta_{t-1}(i) a_{ij} b_j(o_t)], 2 \leq t \leq T, 1 \leq j \leq N \\ \phi_t(j) &= \arg\max_{1\leq i\leq N}[\delta_{t-1}(i) a_{ij}], 2 \leq t \leq T, 1 \leq j \leq N\end{aligned}$$

3. Termination.

$$\begin{aligned}P^* &= \max_{1\leq i\leq N}[\delta_T(i)] \\ q_T^* &= \arg\max_{1\leq i\leq N}[\delta_T(i)]\end{aligned}$$

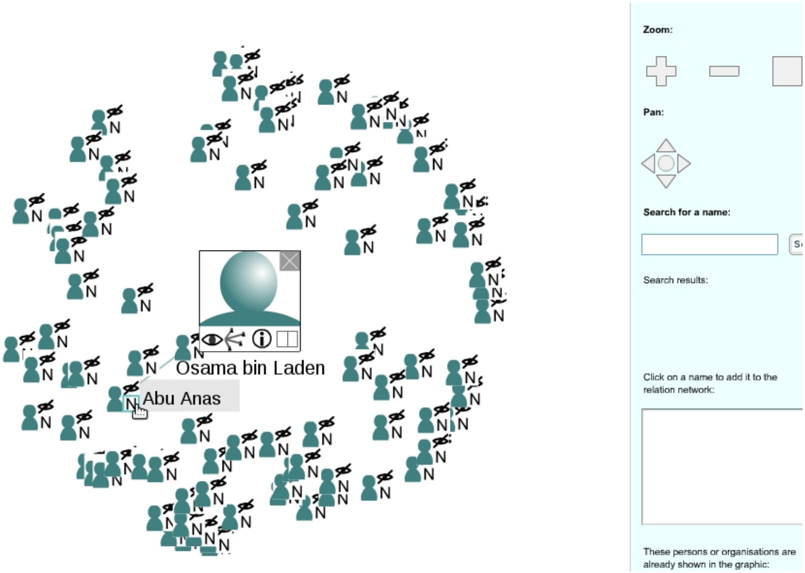
4. Path backtracking (finding the most likely hidden state sequence).

$$q_t^* = \phi_{t+1}(q_{t+1}^*), t = T-1, T-2, \ldots, 1$$

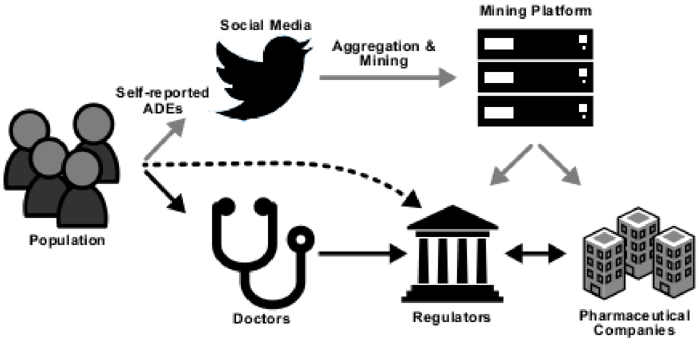
JRC European Media Monitor (EMM) (1/3) – Steinberger et al.

JRC European Media Monitor (EMM) (2/3)

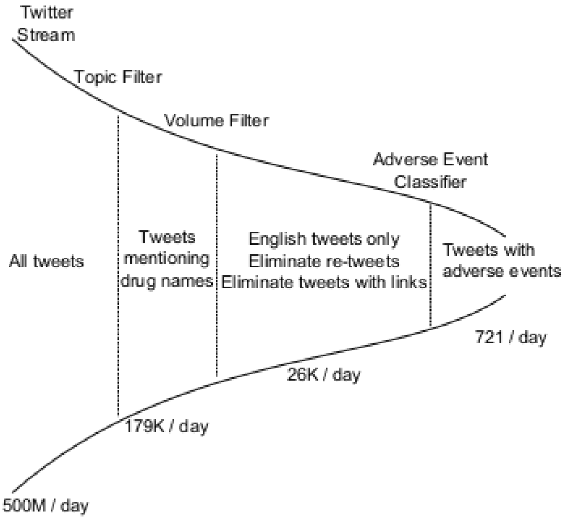
JRC European Media Monitor (EMM) (3/3)



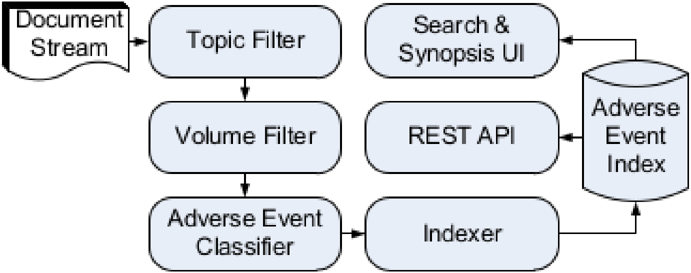
Pharmacology IE – Medical Drug Use Monitoring (Plachouras et al., 2016) (1/6)



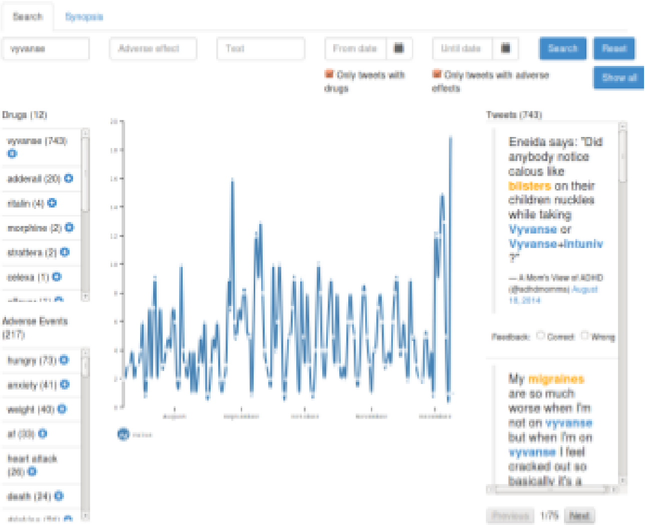
Pharmacology IE – Medical Drug Use Monitoring (Plachouras et al., 2016) (2/6)



Pharmacology IE – Medical Drug Use Monitoring (Plachouras et al., 2016) (3/6)



Pharmacology IE – Medical Drug Use Monitoring (Plachouras et al., 2016) (4/6)

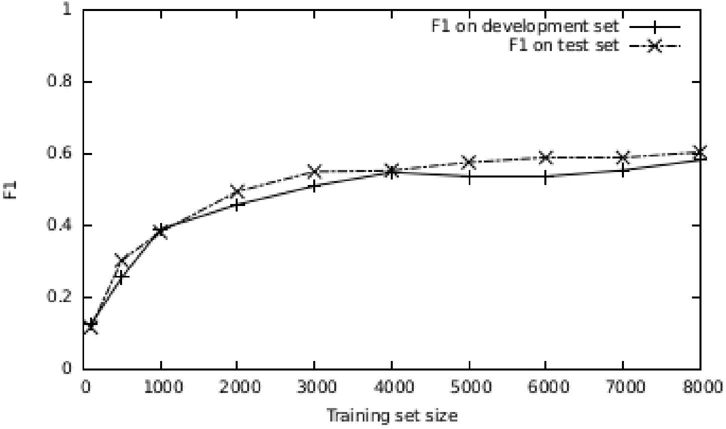


Pharmacology IE – Medical Drug Use Monitoring (Plachouras et al., 2016) (5/6)

	Tweet text
TP	wellbutrin doesnt even work for me it just makes me really anxious idk why im still taking it
TP	Took some ibuprofen that has me so drowsy
TP	Insomnia and heart palpitations due to prednisone. Takteng mga side effects 'to. /wrist
TP	This Vicodin is making me feel like a tweaker because I'm so itchy!
TP	guys i took an ibuprofen 2 days ago and i still have heart palpitations
TN	I once gave a treatment to some patients with glutathione which I knew later clinic use it. Most of them have nausea for 5 mins after inj.
FN	I seriously never have any energy thanks accutane lol @probs_accutane all I want to do is sleep
FP	My mouth taste like 1200 Mg's of ibuprofen yet my head still hurts and I'm still feeling dizzy. Wtf
FP	@ash_hein they gave me Tylenol 3s & yea kinda my mouth still hurts a little & I'm still swollen

Pharmacology IE – Medical Drug Use Monitoring (Plachouras et al., 2016) (6/6)

			Development			Test			
Model	C	w^+	P	R	F1	P	R	F1	
Baseline			0.269	0.683	0.386	0.267	0.705	0.387	
BIN_NGRAM1,2	0.050	8	0.636	0.504	0.562	0.560	0.540	0.549	*
ALL FEATURES	0.025	9	0.573	0.590	0.582	0.550	0.669	**0.604**	*†



Where to Get Linguistic Gold Data From

- Search the Linguistic Data Consortium (LDC) catalog (http://catalog.ldc.upenn.edu)
- Search ELRA catalog (http://catalog.elra.info, http://www.hlt-evaluation.org)
- Browse DFKI's LT World http://www.lt-world.org
- Browse MetaNet (http://www.meta-net.eu)
- Browse the ACL Anthology (http://acl.ldc.upenn.edu), ACM Digital Library (http://dl.acm.org) and IEEExplore (http://ieeexplore.ieee.org)
  - some papers on a topic may contain associated data
  - some conferences, notably the bi-annual LREC, specialize on presenting new linguistic resources
- Contact expert researchers
- Recruit linguists/domain experts
- Crowdsourcing