Applied Natural Language Processing

Dr Jeff Mitchell, University of Sussex

Autumn 2025

Sequence labelling

Previously

- Part-of-speech tagging
- Hidden Markov Models (HMMs)
- Viterbi Algorithm

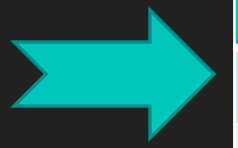
This time

- Information extraction
  - tasks
- Chunking
  - chunking vs parsing
  - IOB labelling
- Named Entity Recognition
  - what is it?
  - detection and classification
  - variation
  - ambiguity

Information extraction

- process of turning unstructured information in text into structured information (e.g., in database)
- e.g., given newswire text extract:
  - countries and their current prime ministers
  - companies and their CEOs
  - actors and their spouses

*Mrs May spoke to **Spanish Prime Minister Pedro Sánchez** on Wednesday evening*



country	role	person
Spain	prime minister	Pedro Sanchez

Tasks in Information Extraction

- Text Chunking
  - simplified form of parsing
- Named Entity Recognition (NER)
  - finds and classifies strings of tokens that mention named entities
- Coreference Resolution
  - links named entity mentions in a document that refer to the same entity
- Entity linking
  - associates named entity mentions with concepts in a knowledge base
- Relation recognition
  - finds and classifies relationships between entities
- Event recognition
  - finds and classifies events and the entities in roles associated with the event

Chunking

Syntactic Chunking

- Grouping tokens into syntactically correlated phrases (chunks)
  - phrases that play similar roles in sentences
  - we usually distinguish noun phrases (NPs) and verb phrases (VPs)
  - substitution of 1 NP for another NP or 1 VP for another VP results in grammatically plausible sentences (ignoring agreement issues)

[The current account deficit]$_{NP}$ [will narrow]$_{VP}$ to [only $ 1.8 billion]$_{NP}$

Example chunks

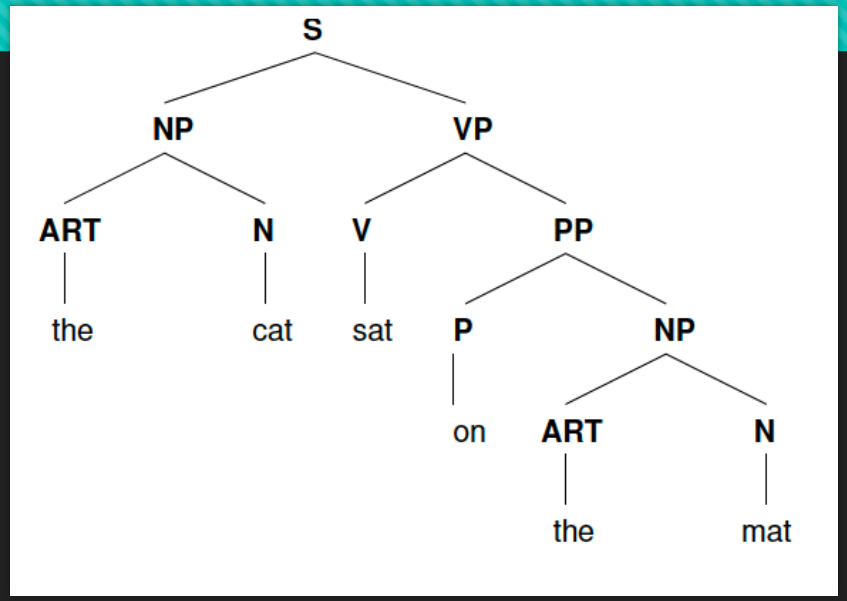
Noun phrase chunks

- cats
- the car
- the cheap plastic garden seat
- many of my favourite 1980's rock songs
- Apple Computers Inc
- The highly successful University of Sussex

Verb phrases chunks

- snores
- take
- used to be
- begins at
- has been successfully sold by
- is looking to raise

Parsing



- determines the complete syntactic structure of sentence
- hierarchical analysis of how phrases combine to make other phrases

Chunking

the cat *NP* sat *VP* on the mat *NP*

- easier
- more efficient
- more robust to unexpected input
- often gives detailed enough structure

IOB labels

	token	POS	Chunk
0	The	DET	B-NP
1	cat	N	I-NP
2	sat	V	B-VP
3	on	PREP	O
4	the	DET	B-NP
5	mat	N	I-NP
6	.	.	

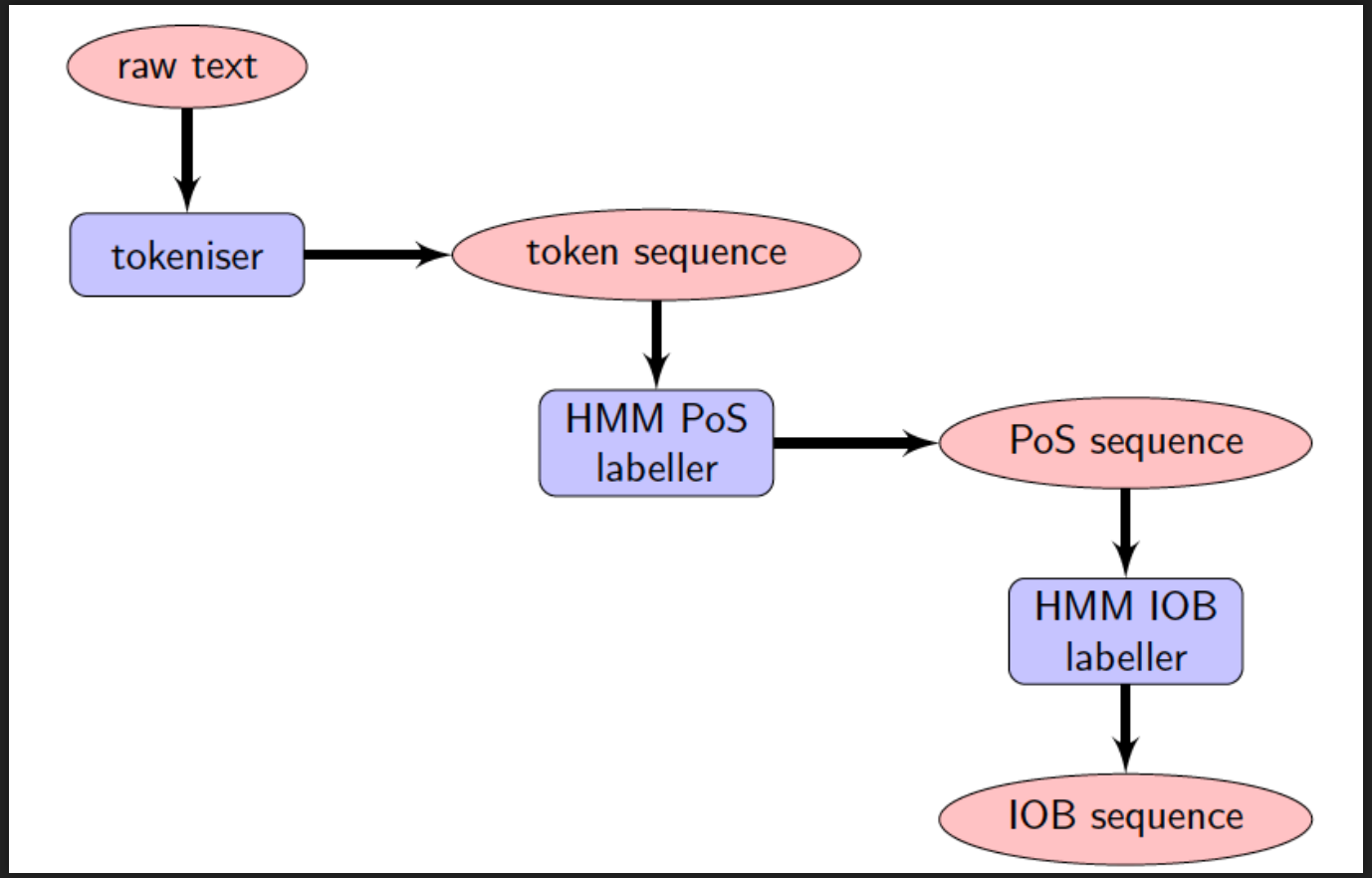
IOB labels used to indicate chunk boundaries

- B-NP = **b**eginning of **NP** chunk
- I-NP = **i**nside **NP** chunk
- B-VP = **b**eginning of **VP** chunk
- I-VP=**i**nside **VP** chunk
- O = **o**utside any chunk

IOB labelling is like PoS tagging

- Part-of-speech tagging involves labelling a sequence of tokens with part-of-speech tags
- IOB chunking involves labelling a sequence of part-of-speech tags with IOB labels
- The same technology can be used for both
  - e.g., Hidden Markov Models

Chunking as sequence labelling



Named Entity Recognition

Named Entity Recognition (NER)

- is the *detection* and *classification* of the **named entities** in a text,
- a named entity is anything which can be referred to with a proper name e.g.:
  - Boris Johnson
  - England
  - Facebook
  - University of Sussex
  - Brighton & Hove Albion

Questions about named entities

- What **mentions** of named entities are there in a text?
- What **types** of entities are being referred to?
- Which entities in the real world are being referred to?
- What relationships are there between the entities in the real world?

What mentions of named entities are there in this text?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

What mentions of named entities are there in this text?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

What types of named entities are there in this text?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

People

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

Locations / Geo-political Entities?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

Organizations

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

How many different real-world named entities are referred in this text?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

Variation

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

What relationships are there between the named entities in this text?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

What relationships are there between the named entities in this text?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma **star** Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

What relationships are there between the named entities in this text?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United **boss** Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

What relationships are there between the named entities in this text?

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, **agent** Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

Detecting named entities

- named entities are chunks of tokens
  - individual words may have been tagged as PROPN
  - individual words may be capitalised
  - may have been recognised as a NP
- sequence labelling task

Named Entity classification

- types / classes vary according to domain
- typical named entity types for news domain:

Type	Tag	Examples
People	PER	Lorenzo Pellegrini
Organization	ORG	Manchester United
Location	LOC	Manchester
Geo-Political Entity	GPE	Manchester
Facility	FAC	Cuilfail Tunnel
Vehicles	VEH	Aston Martin

More Named entity types

- 'CARDINAL',
- 'DATE',
- 'EVENT',
- 'FAC',
- 'GPE',
- 'LANGUAGE',
- 'LAW',
- 'LOC',
- 'MONEY', '
- 'NORP'
- 'ORDINAL',
- 'ORG',
- 'PERCENT',
- 'PERSON'
- 'PRODUCT'
- 'QUANTITY'
- 'TIME'
- 'WORK_OF_ART

These NE types were all found in the text of the Jane Austen novel "Emma" by the SpaCy named entity recogniser

- see the lab

Variation

- Different ways of referring to the same entity
  - Dr Jeff Mitchell, Dr Mitchell, Dr J. Mitchell, Dr J. J. Mitchell, ….
- Variants usually follow orthographic rules (which may be type dependent), but not always
  - Manchester United, Man Utd, United, Red Devils
- **Named entity linking** is therefore difficult
  - Usually requires a knowledge source e.g., database containing known named entities.
  - More later

Ambiguity

- The same string may refer to multiple named entities
- Entities may be of the same type:
  - multiple LOCS in the world with the name "Manchester"
  - multiple PERSONS in the world with the name "John Smith"
- Entities may be of different types:
  - JFK – can be a person or an airport
  - Manchester – can be a place or a song
  - England – can be a place or a football team or a rugby team or …..
- Makes type classification and named entity linking difficult

NER as a sequence labelling task

- NE detection and classification usually carried out simultaneously.
- Each token in a sequence is tagged with the named entity type associated with it

Lorenzo/PER Pelligrini/PER 's/PER agent/NONE shuts/NONE down/NONE Manchester/ORG United/ORG links/NONE .

- so NER can be performed with same technology as POS-tagging

NE Chunking

	token	lower	lemma	pos	NER
0	Lorenzo	lorenzo	lorenzo	PROPN	PERSON
1	Pellegrino	pellegrino	pellegrino	PROPN	PERSON
2	's	's	's	PART	PERSON
3	agent	agent	agent	NOUN	
4	shuts	shuts	shut	VERB	
5	down	down	down	ADP	
6	Manchester	manchester	manchester	PROPN	ORG
7	United	united	united	PROPN	ORG
8	links	links	link	NOUN	
9	.	.	.	PUNCT	

Typically, want to find chunks of consecutive tags with the same NER tag

- named entity is "Manchester United"
- NOT "Manchester" and "United"

IOB ENCODING

	token	….	NER
0	Lorenzo		B-PER
1	Pellegrino		I-PER
2	‘s		I_PER
3	agent		O
4	shut		O
5	down		O
6	Manchester		B-ORG
7	United		I-ORG
8	links		O
9	.		O

Inside
Outside
Beginning

- Used by some NER systems
- avoids possible ambiguity between multiple NEs and NEs with multiple tokens
- but multiple NEs are rare (usually separated by punctuation)

Features for NER

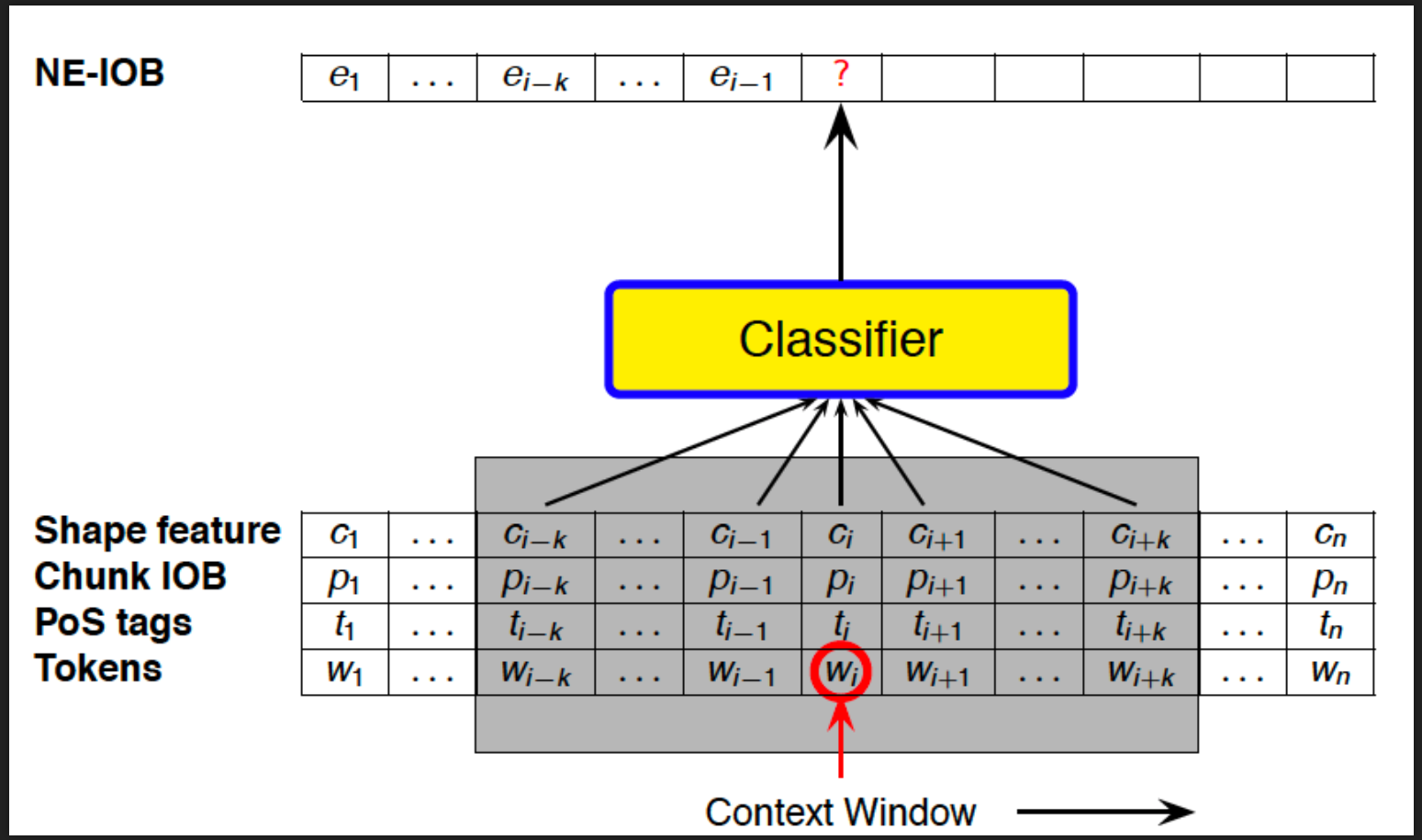
POS-tagging

- Word identity
- Tag sequence

NER

- Word identity
- Tag sequence
- Capitalization
- POS tags
- Chunks

Named entity recognition as classification



NER as sequence labelling

- Can we do better than classifying each token separately?
- Can we use an HMM to predict the best sequence of tags given the sequence of observations?

NER as Sequence Labelling

- HMMs?
  - have strict feature independence assumptions
- NER feature sets
  - rich, overlapping
  - violate independence assumptions
- Sequence models for NER tend to be variants/extensions of HMMs with less strict feature independence assumptions
  - Maximum Entropy Markov Models (MEMMs)
  - Conditional Random Fields (CRFs)
  - Neural networks
  - More in AdvNLE /AdvNLP

Co-reference Resolution

Co-reference resolution

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United boss Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

Co-reference resolution

- Which named entity mentions in a document **co-refer**?
  - refer to the same real world entity
- Challenges
  - variation
  - ambiguity
- Helps to have an identified set of real world entities which might be referred to

Entity Linking

Named entity linking

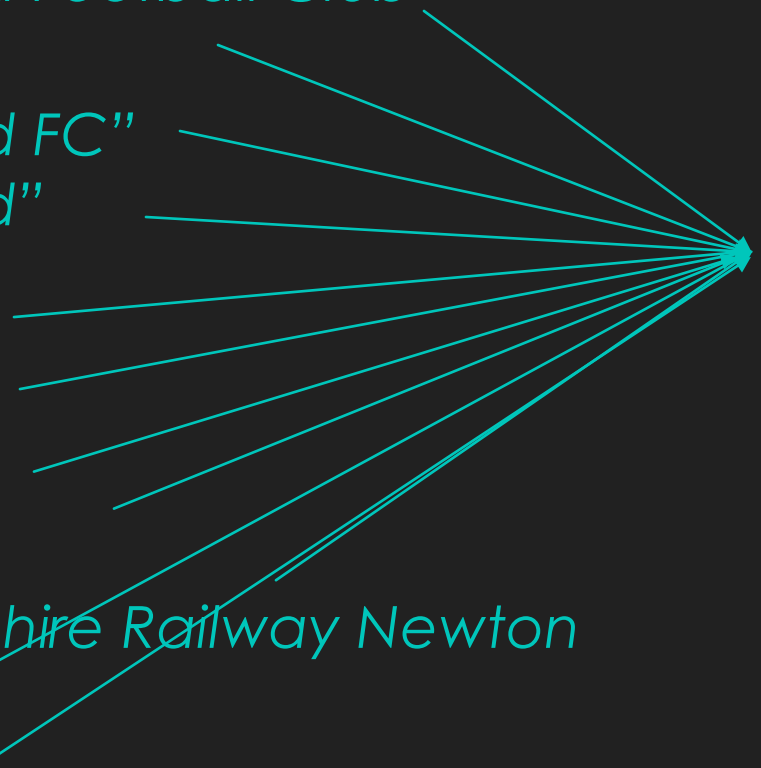
- Which real world entity does a named entity mention refer to?
- Challenges
  - ambiguity
  - variation

Entity ambiguity

- the city in England
- the town in Bolivia
- one of 32 towns in the USA
- a football club
- the University
- the Airport
- the song by The Beautiful South

Name variations

- *"Manchester United Football Club"*
- *"MUFC"*
- *"Manchester United FC"*
- *"Manchester United"*
- *"Man U"*
- *"Manchester"*
- *"United"*
- *"The Reds"*
- *"Busby Babes"*
- *"Lancashire & Yorkshire Railway Newton Heath"*
- *"The Heathens"*



Manchester United FC

Why this is so challenging

- Sorting out name variation increases problem of ambiguity
- the 42nd President of the United States can be referred to as:
  - Bill Clinton
  - William Clinton
- William Clinton may also refer to:
  - William de Clinton, 1st Earl of Huntingdon
  - William Henry Clinton, the British General

Formulation of Named Entity linking problem

A problem instance consists of:

- A knowledge base (KB) such as Wikipedia
- An entity mention in a textual context

Goal:

return canonical entry in KB of entity being mentioned
OR
return NIL if the entity does not exist in KB

Wikipedia as the KB

- Many named entities have their own page in Wikipedia
- The title of the page is canonical way of naming the entity

Bill Clinton

From Wikipedia, the free encyclopedia
(Redirected from William Clinton)

"William Clinton" redirects here. For other uses, see William Clinton (disambiguation).

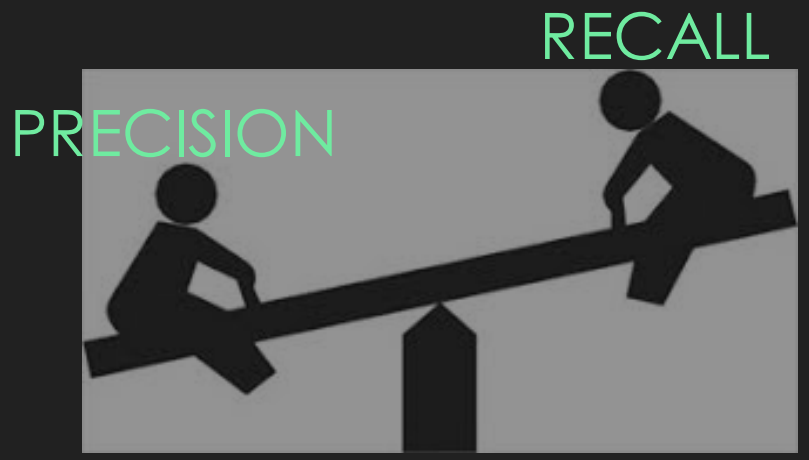
William Jefferson Clinton (born **William Jefferson Blythe III**; August 19, 1946) is an American politician who served as the 42nd President of the United States from 1993 to 2001. Prior to the presidency, he was the Governor of Arkansas from 1979 to 1981, and again from 1983 to 1992. A member of the Democratic Party, Clinton was ideologically a New Democrat and many of his policies reflected a centrist "Third Way" political philosophy.

Techniques for entity linking

- 2 step solution
- Step 1:
  - Find **candidates** in the KB for given entity mention
- Step 2:
  - **Rank** candidates to find the most probable

Generating Candidates

- addresses name variation challenge
- find all potentially relevant candidates
- familiar tradeoff between precision and recall
- Need high recall so that correct entity is among candidates
- But too many candidates hurts precision and efficiency



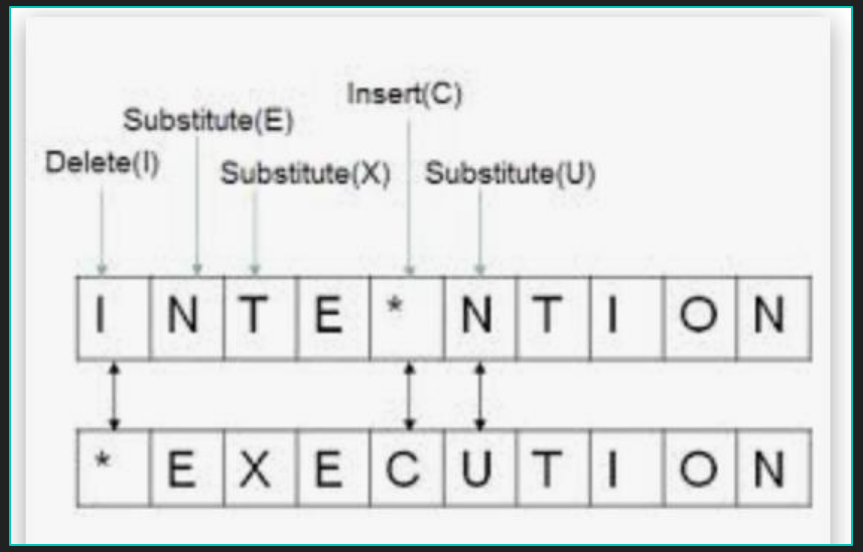
Strategies for generating candidates

Ignore context – find as many candidates as possible

- mention is exact match with title of Wikipedia page
  - *Bill Clinton*
- mention is a substring of Wikipedia page or vice versa
  - *Clinton* or *Mr Bill Clinton*
- mention is an acronym of page title
  - *B.C.* or *UoS* (for University of Sussex)
- mention is a known alias for page title
  - extracted from redirects and disambiguation pages
  - *William Clinton*

Generating candidates using string similarity

- mention is a **similar** string to page title
  - use a string similarity measure such as Levenshtein Distance (Minimum Edit Distance)
- Levenshtein Distance between 2 strings is the number of insertions, deletions and substitutions required to transform one string into the other



Levenshtein distance between *intention* and *execution* is 5

Information for ranking candidates

- **Co-occurrence** of entity mentions
  - other named entity mentions in the same document
- **Local context** of an entity mention
  - neighbouring words
- **Global context** of an entity mention
  - document within which entity mention occurs
  - bag-of-words captures topic

Strategies for ranking candidates

- Entity relatedness
  - do **co-occurring entity** mentions link to this or similar KB pages?
- Query relevance
  - does a candidate KB page contain tokens in the **local context** of entity mention?
- Document similarity
  - does a candidate KB page have a high bag-of-words similarity to the **global context** of the entity mention

Named entity linking as disambiguation

- Candidates are the same each time a string referring to a named entity is used
- Each candidate is a possible ***sense*** of the named entity mention
- Solve with same machine learning technology as WSD
  - annotated training corpus
  - extract features from:
    - **local context** (a vector of words and/or tags in local context)
    - **global context** (a vector of words and/or tags in global context)
    - **co-occurring entity mentions** (a vector of other entities possibly mentioned)
  - NB or other classifier

Relation Extraction

Relation extraction

Lorenzo Pellegrini's agent shuts down Manchester United links

- Roma star Lorenzo Pellegrini is not considering a move to Manchester United and is "only thinking about Roma", his agent says.
- The 22-year-old Italy international midfielder has a £27m release clause in his contract and has reportedly attracted interest from United **boss** Jose Mourinho following a fruitful start to the season.
- However, agent Giampiero Pocetta dismissed speculation that Pellegrini is set to join the Red Devils, despite reports they were ready to make a move for him as soon as January.

relation extraction

Discovering relationships between entities

<entity>	<relationship>	<entity>
Jose Mourinho	*boss-of*	Manchester United FC
Lorenzo Pellegrini	*star-of*	Roma
Giampiero Pocetta	*agent-of*	Lorenzo Pellegrini

- typically concerned with **binary** relationships
- fundamental to meaning

Relation granularity

- Recall that named entities are classified into classses
  - PERSON, ORG, LOC etc
- Relation types can also be organised into classes
  - with different levels of granularity

star-of → plays-for → works-for → is-associated-with

Supervised approaches to relation extraction

- 2-step solution
- Step 1:
  - Extract a pair of entities, e_1 and e_2, which are mentioned together
  - Detection problem: is a relationship between e_1 and e_2 being asserted in this document or not?
- Step 2:
  - Classify the relationship between the entities
  - Multiclass classification problem: what is the relationship between related entities e_1 and e_2 ?
- OR build a separate binary classifier for each relationship

Training relation extractors

- binary extraction problem
- need positive and negative examples
  - examples where 2 entities co-occurring in a document are related
  - examples where 2 entities co-occurring in a document are not related
- multiclass classification problem
- need examples of each class
- extract features and train standard NB classifier

Features for relation extraction

feature representation needs to capture potential relationship between entities

- types of target named entity mentions
- tokens in target named entity mentions
- distance between entity mentions
- number of named entity mentions between target named entity mentions
- bag of words or tags or named entity mentions separating entity mentions (local context)
- bag of words or tags or named entity mentions surrounding entity mentions (global context)

NB assumptions

- Does it matter if NB assumptions are violated?
- It depends how and why we want the probabilities :
  - *P(class | observed features)*
- Do the (incorrect) probability estimates lead to good classification decisions?
- Other classification models are available with less strict independence assumptions
  - e.g., logistic regression / maximum entropy classifiers

Making progress

- There is one notebook to complete for this week:
  - Part 1: Lab_9_1.ipynb