Module 3: GATE and Social Media

Part 4. Named entities

Named Entity Recognition

- Texts frequently focus on particular entities
- To discover what documents say about them, we can:
  - Recognise entity mentions
  - Disambiguate entities to external vocabularies
  - Find opinions that authors have about the entities
- Important because:
  - Enables IE over tweets
  - Critical for event extraction (actors, events)
  - Describes the topic of the tweet
- Tough because:
  - ANNIE doesn't do well – around 50% F1
  - Stanford's leading tool does even worse – around 40% F1!

Named entity recognition: example

Goal is to find mentions of entities

Newswire — London Fashion Week grows up – but mustn't take itself too seriously. Once a launching pad for new designers, it is fast becoming the main event. But LFW mustn't let the luxury and money crush its sense of silliness.

Social media — Gotta dress up for london fashion week and party in style!!!

Named entity recognition: example

Person mentions in news

Left context	Match	Right context
in dicated Atef, including	Douglas Feith	, the United States defence
, the group that killed	President Sadat	in 1981 as retribution for
. The current leader,	President Olusegun Obasanjo	, who recently came to
Kuwait, whose information minister	Sheikh Ahmed Fahed al-Sabah	met editors of local newspapers
The current defence minister,	Theophilus Danjuma	, has also been threatened
The three right-wing MPs,	Andrew Rosindell	(Romford), Andrew
Late on Wednesday night,	Justice Oputa	, who chairs the commission
the militarily-manoeuvred civilian elec...	President Obasanjo	in 1999 and is widely
after the mysterious death of	General Sani Abacha	in 1998.
have learnt that one of	Bin Laden	's closest and most senior
evidence confirms the involvement of	Osama bin Laden	in those attacks."
. He is one of	Bin Laden	's two most senior associates
for future civilian office.	General Buhari	took power in a 1983
$5m price on	Atef	's head and prosecutors have
Afghanistan. He was once	Bin Laden	's chief media adviser and
thinking in the Tory party	Iain Duncan Smith	has ordered three Tory MPs
club and the party,	David Maclean	, the Tory Chief Whip
Centre and the Pentagon.	Mohammed Atef	, who is thought to
are still very powerful.	General Babangida	supported the militarily-manoeuvred ci
sexual orientation or religion.	Mr Duncan Smith	's purge of the Monday
," he said.	Atef	, who is reported variously
of the late singer,	Fela Kuti	� which took place while
field in Penn sylvania.	President Bush	included Atef in an order
. It is believed that	Mr Duncan Smith	intended to launch his crackdown

Named entity recognition: example

Person mentions in tweets

Left context	Match	Right context
i was your age ,	spencer	from iCarly was Crazy Steve
iCarly was Crazy Steve ,	Carly	was Megan and Josh was
bath , shut up ,	sam	's coming tomorrow and steve
. All are welcome ,	joe	included
. All are welcome ,	joe	included
teachers , chinese takeaways ,	gatt holly	, phil collins , the
takeaways , gatt holly ,	phil collins	, the skin of a
@GdnPolitics : RT AlJahom :	Blair	: " I'm gonna
Empls of the Month :	Deborah L	#Speech #Pathologist-Childrens
be the next Pope "	Brown	: " I won't
(via POPSUGAR)	Sarah Jessica Parker	and Gwen Stefani Wrap Up
and is smexy !!;)And	Chelsea Handler	is hilarious ! Finally got
him befnrjustthen about	kenny	signing his book but it
three kinds of reactions after	Ayodhya	verdict .
, Carly was Megan and	Josh	was fat . #damnteenquotes
sam 's coming tomorrow and	steve	and tanya will be round
coming tomorrow and steve and	tanya	will be round at 10am
photo caption contest- Nadal and	Novak	in the tub http://ow.ly/2G3Jh
) Sarah Jessica Parker and	Gwen Stefani	Wrap Up Another Successful New
#Pathologist-Childrens Rehab and	Patricia M	#Referral/#Auth #
Just casually stalking Cheryl AND	Dermot	tomorrow .... NO BIGGIE
did tweet him befnr	justthen	about kenny signing his book
Test : We just congratulated	Lindsay	an hour ago on h
the funny photo caption contest-	Nadal	and Novak in the tub

Named entity recognition: issues

Genre differences in entity type

	News	**Tweets**
PER	Politicians, business leaders, journalists, celebrities	Sportsmen, actors, TV personalities, celebrities, names of friends
LOC	Countries, cities, rivers, and other places related to current affairs	Restaurants, bars, local landmarks/areas, cities, rarely countries
ORG	Public and private companies, government organisations	Bands, internet companies, sports clubs

Named entity recognition approaches

Ritter (2011) addresses named entity recognition in tweets using a data-intensive approach:

- Distinct segmentation and classification tasks
- Assume that @mentions are unambiguous
- Found that inclusion out-of-domain data (from MUC) actually reduces performance

Named entity recognition approaches

Models entity segmentation as sequence labeling using BIO representation and CRF

- Orthographic, contextual features
- Dictionary features based on type lists in Freebase
- Brown clusters from PoS tagging, NP/VP/PP chunking, capitalisation

Segmentation outperforms default Stanford NER consistently

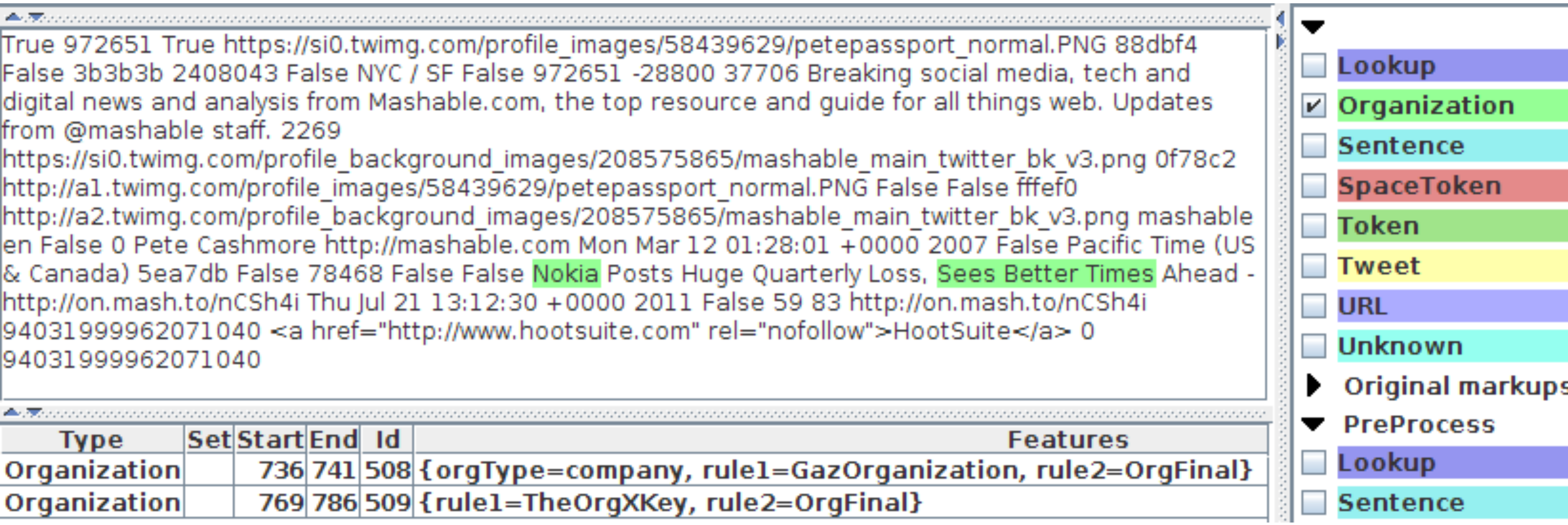
- Stanford: F1 44%
- Segmentation without clusters: F1 63%
- Segmentation with clusters: F1 67% (52% error reduction)

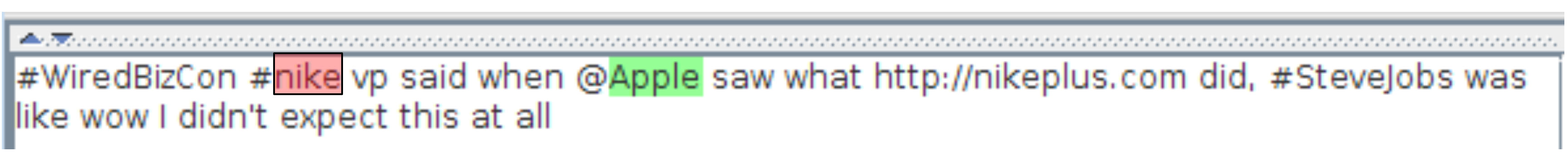
Named entity recognition approaches

After segmentation, Ritter (2011) describes NE classification

- Diversity in entity types exacerbates data sparsity problem
- Lack of context makes classification difficult even for humans
- Co-occurrence can help in situations like this

Tweet Capitalisation: an NER nightmare!





…And hashtag semantics is yet another…

Case-Insensitive matching

- This would seem the ideal solution, especially for gazetteer lookup, when people don't use case information as expected
- However, setting all PRs to be case-insensitive can have undesired consequences
- POS tagging becomes unreliable (e.g. “May” vs “may”)
- Back-off strategies may fail, e.g. unknown words beginning with a capital letter are normally assumed to be proper nouns
- BUT this doesn’t work on tweets anyway!
- Gazetteer entries quickly become ambiguous (e.g. many place names and first names are ambiguous with common words)
- Solutions include selective use of case insensitivity, removal of ambiguous terms from lists, additional verification (e.g. use of the text of any contained URLs)

More flexible matching techniques

- In GATE, as well as the standard gazetteers, we have options for modified versions which allow for more flexible matching
- BWP Gazetteer: uses Levenshtein edit distance for approximate string matching
- Extended Gazetteer: has a number of parameters for matching prefixes, suffixes, initial capitalisation and so on

Hands-on: NER evaluation

Let's measure ANNIE performance on social media text

- Open the Ritter-dev corpus from the datastore saved in corpora/r-tweets
- Change all the annotationSetName, inputAS and outputAS parameters in your ANNIE application to ANNIE
- Run your ANNIE pipeline on this corpus
- Have a look at the entities annotated. Can you find any mistakes>
- If so, why do you think this mistake has been made?

Now let's try with TwitIE

- Remove your Twitter application from GATE (to avoid confusion)
- Load the TwitIE application from the "Ready-made Applications"
- Add ANNIE in the setsToKeep parameter of the Document Reset
- Run TwitIE

Compare ANNIE and TwitIE

- Open the corpus and click the "Corpus Quality Assurance" tab
- We can now compare 3 annotation sets: Original Markups (the gold standard set) with both TwitIE and ANNIE results
- Pick 2 of these sets to compare (TwitIE results are now in the default set)
- Select annotation types Location, Organization, and Person
- Pick an evaluation measure
- How does it do? What kinds of errors are most prevalent, missed or spurious?
- You can also pick individual documents and see which single annotations are picked up or missed

Named entity recognition summary

Named entity recognition in tweets is hard

Three major classes of Tweet NER approach:

- **Sequence labelling** – like Stanford CRF chunker
  - Problem: tweets aren't well-formed enough
  - Problem: lack of training data
- **Lookup-based** using local grammar and string matching
  - Problem: strings are often misspelled
  - Problem: entity mentions aren't in gazetteers (drift) (Eisenstein 2013, Plank 2014)
  - Advantage: cuts through linguistic noise, agnostic to many style variations
- **Grounding to vocabulary** (e.g. Dbpedia)
  - Problem: insufficient context to disambiguate
  - Problem: entities often appear in social media before the resource

Overall solutions to twitter noise

Normalisation

- Convert twitter text to “well-formed” text; e.g. slang resolution
- Some success using noisy channel model (Han 2011)
- Techniques include: edit distance; double metaphone with threshold
- Issues: false positives can change meanings, e.g. reversing sentiment (apolitical)

Domain adaptation

- Treat twitter as its own genre, and create customised tools and techniques
- Some success in language ID (Carter 2013), PoS tagging (Gimpel 2011), NER (Ritter 2011)

Flexible representations

- Represent document text in a way that is resilient to noise
- Brown clustering, word embeddings, deep learning

Extra hands-on

- If you have spare time, you can try loading and running TwitIE on some of the tweets you collected yourself