Entity recognition

Leon Derczynski

Entity mining

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:

- Enables IE over tweets
- Critical for event extraction (actors, events)
- Describes the topic of the tweet

Tough:

- ANNIE doesn't do well – around 50% F1
- Stanford's leading statistical tool does even worse – around 40% F1!

What's going on? How can we build a tweet NER tool?

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!!!

General accuracy on newswire: 89% F1
General accuracy on microblogs: 41% F1

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 civ
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: resources

There are few named entity corpora for tweets.

UW (Ritter, 2011)

- 34k tokens, 1500 entities
- Single annotator
- Ten entity types: PERSON, GEO-LOCATION, COMPANY, PRODUCT, FACILITY, TV-SHOW, MOVIE, SPORTSTEAM, BAND, and OTHER

UMBC (Finin, 2010)

- 7k tokens, 500 entities
- Multiple annotator
- Three entity types: PERSON, LOCATION, ORGANISATION

MSM2013 (Basave, 2013)

- 30k tokens, 1500 entities
- Multiple annotator
- Three entity types: PERSON, LOCATION, ORGANISATION
- Hashtags, URLs and entitys obfuscated

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: issues

Capitalisation is not indicative of named entities

- All uppercase, e.g. *APPLE IS AWSOME*
- All lowercase, e.g. *all welcome, joe included*
- All letters upper initial, e.g. *10 Quotes from Amy Poehler That Will Get You Through High School*

Unusual spelling, acronyms, and abbreviations

Social media conventions:

- Hashtags, e.g. *#ukuncut, #RussellBrand, #taxavoidance*
- @Mentions, e.g. *@edchi* (PER), *@mcg_graz* (LOC), *@BBC* (ORG)

For newswire: (Derczynski 2013)

- Rule-based systems get the bulk of entities 77% F1
- ML-based systems do well at the remainder 89% F1

Using social-media specific, custom preprocessing on tweets:

- ML struggles, even with in-genre data: 49% F1
- Rules cut through microblog noise: 69% F1

Tweet Named Entity Recogntion

Design choices in NER: (Roth 2009)

- **What feature representation to use for tokens;**
- **Which inference algorithm to use;**
- **How to capture non-local dependencies;**
- **How to incorporate external knowledge.**

Representation and labeling

Token feature representation options:

- Token itself
- Previous and following token
- Word shape, to model capitalisation
- Lexical features (e.g. character n-grams) to help with OOV terms
- Part of speech tag
- Parsing information

NER inference algorithms

As with part of speech tagging, sequence labelling can work well (e.g. CRF)

- Assumes well-formed sentences and lots of training data
- If this is inappropriate, then local context in token features can compensate

Representation and labeling

Labelling scheme:

I	Facebook	B-company
O	Job-Hunting	O
O	App	O
I	BranchOut	B-product
O	Raises	O
O	$6	O
O	Million	O
O	From	O
I	Accel	B-company
O	And	O
I	Super	B-company
I	Angels	I-company

BIO (Begin, In, Out) allows separation of adjacent entities

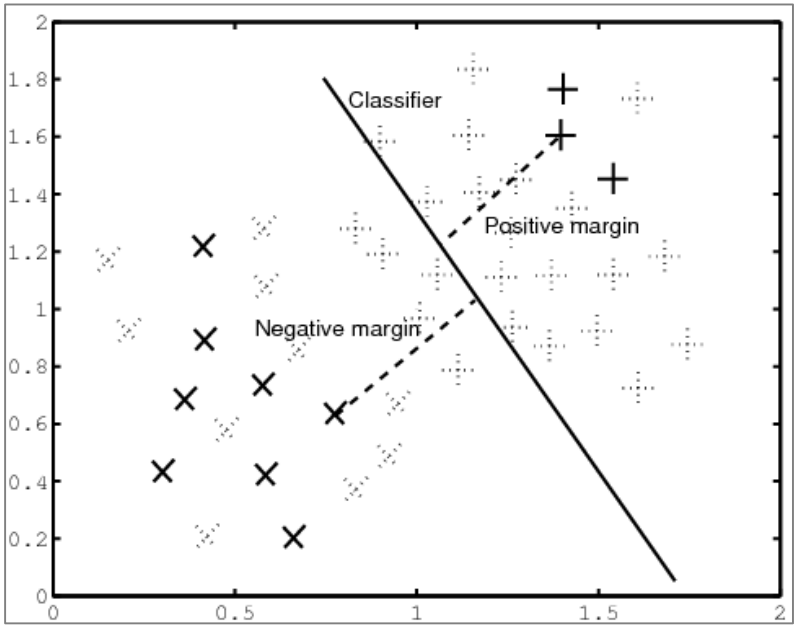
CRF with BIO popular

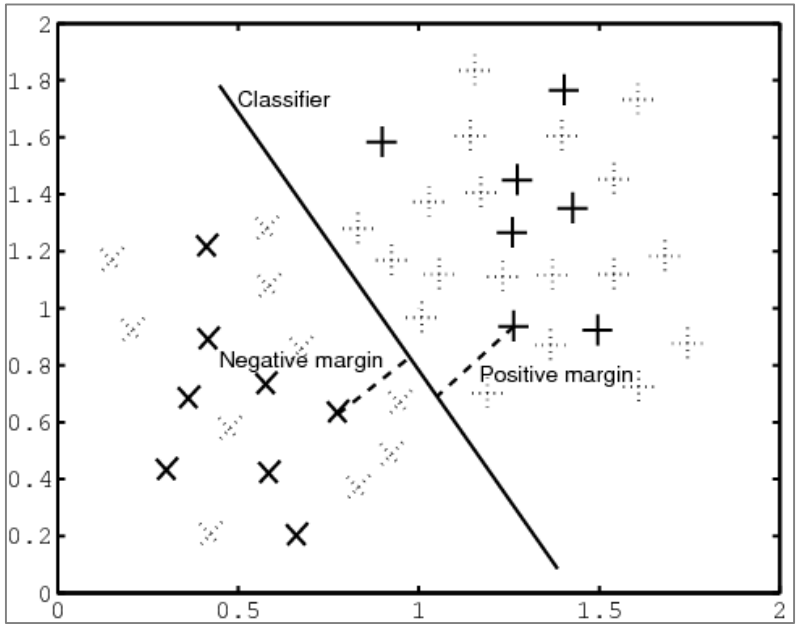
SVM-U with IO can give better performance

Representation and labeling

SVM-U: “uneven” (Li 2009)

Adjust margins between supporting examples and decision hyperplane to reflect class balance





Well-suited to tasks like NER, where one class is much more frequent than another

Retains SVM's advantage of being noise-resistant

Dependencies & external knowledge

Typically, only the first mention of an entity is referred to in full:

Manchester United are great. They're my favourite football team. Man U forever!

Using only local features will lead to missed entities.

Tweets are not long discourses

- Possible for the long first mention to be missing
- Include context from elsewhere

How can we incorporate external knowledge for NER?

- Useful to tell us when unusual/unexpected words are an entity: “Szeged” “White House”

Unlabelled text

- NEs found in distributionally similar contexts
- Labelled LDA can produce phrase lists given an entity type (Ramage 2009, Ritter 2011)

Gazetteers

- Can be constructed manually or automatically
- Gaz. completeness gives P/R tradeoff
- Won't catch terms not seem in gazetteer, which makes domain adaptation tough

Named entity recognition: Facebook

Longer texts than tweets

Still has informal tone

Multi-word expressions are a problem!

- all capitalised:
- *Green Europe Imperiled as Debt Crises Trigger Carbon Market Drop*

Difficult, though easier than Twitter

- Maybe due to option of including more verbal context?
- Lack of training data

Named entity recognition approaches

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

Distinct segmentation and classification tasks

- Discriminative segmentation
- Distantly supervised classification

Assume that @mentions are unambiguous

Found that inclusion out-of-domain data (from MUC) actually reduces performance

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
- e.g., KKTNY in 45min.....
- Co-occurrence can help in situations like this (Downey 2010)

Exploiting co-occurence information with LabeledLDA and Freebase

- Freebase provides type ontology
- LabeledLDA assigns distribution of potential Freebase types to entity mentions
- Entity mention context modelled as bag-of-words
- Distribution can vary from mention to mention
- Include prior for type distribution θ_e from encountered examples, to compensate for cases where there are few words for context

Evaluation over 2400 tweets, 10 types

- Unlabelled data from 60M NE segmented tweets (24K distinct entity strings)
- Freebase F1 38%
- Supervised F1 45% (MaxEnt)
- LabeledLDA F1 66%

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

Grouding to vocabulary (e.g. Dbpedia)

Problem: insufficient context to disambiguate
Problem: entities often appear in social media before the resource

References (1)

Basave, Varga, Rowe, Stankovic, Dadzie 2013. Making sense of microposts (#MSM2013) concept extraction challenge. WWW MSM2013 workshop

Balahur, Steinberger 2009. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. WOMSA

Barbosa and Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. COLING '10.

Derczynski, Maynard, Aswani, Bontcheva 2013. Microblog-genre noise and impact on semantic annotation. Hypertext

Downey, Etzioni, Soderland 2010. Analysis of a probabilistic model of redundancy in unsupervised information extraction. AI 174(11):726

Eisenstein 2013. What to do about bad language on the internet. NAACL

Finin, Murnane, Karandikar, Keller, Martineau, Dredze 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk

Go, Bhayani, Huang 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford.

Hangya &Farkas 2013. Filtering and Polarity Detection for Reputation Management on Tweets. CLEF 2013.

http://www.clef-initiative.eu/documents/71612/51490ac1-b1fa-4ea2-a520-6b52ef98e862

Ji and Grishman 2011. Knowledge Base Population: Successful Approaches and Challenges. ACL/HLT 2011

Li, Bontcheva, Cunningham 2009. Adapting SVM for data sparseness and imbalance: A case study in information extraction. JNLE 1(1):1

Liu 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers

Liu, Wang, Li, and Liu. 2010. Improving blog polarity classification via topic analysis and adaptive methods .HLT 2010.

Meij, Weerkamp & de Rijke 2012. Adding Semantics to Microblog Posts. In Proceedings of the 5th International Conference on Web Search and Data Mining (WSDM'12)

References (2)

O'Connor, Balasubramanyan, Routledge, Smith: From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. ICWSM 2010

Ott, Choi, Cardie, and Hancock 2011. Finding deceptive opinion spam by any stretch of the imagination. ACL.

Pak and Paroubek 2010. Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. 5th International Workshop on Semantic Evaluation.

Plank, Hovy, Søgaard 2014. Learning POS taggers with inter-annotator agreement loss. EACL

Ramage, Hall, Nallapati, Manning 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. EMNLP

Ritter, Clark, Mausam, Etzioni 2011. Named Entity Recognition in Tweets: An Experimental Study. EMNLP

Roth, Ratinov 2009. Design challenges and misconceptions in named entity recognition. CoNLL

Taboada et al 2011. Lexicon-Based Methods for Sentiment Analysis. Computational Lingusitics. Vol. 37. Num 2.

Wang, Wei, Liu, Zhou, and Zhang. 2011. Topic sentiment analysis in Twitter: a graph-based hashtag sentiment classification approach. CIKM '11.