Learning to recognise named entities

Joel Nothman

2018-04-24

Outline

- **Background to NER**:
  a useful NLP task
- **Statistical approaches**:
  how and why
- **Annotated corpora**:
  learning and evaluation
- **New data from Wikipedia**:
  avoid costly annotation?
- **Conclusion**:
  the challenge of semantics

What is named entity recognition?

- Given a text, identify names and classify them
- For example:

Hurricane Katrina hit Louisiana on August 29.

⇓

[MISC Hurricane Katrina] hit [LOC Louisiana] on [DATE August 29].

Paris Hilton visited the Paris Hilton.

⇓

[PER Paris Hilton] visited the [LOC Paris] [ORG Hilton].

- A type of semantic/reference annotation
- Impossible to know all names in all contexts

Why recognise entities?

- Information extraction

  $$MergerBetween(company_1, company_2, date)$$

  In 2000, Air New Zealand announced that it had chosen to acquire the entirety of Ansett Australia.

- Question answering

  Which airline did Air New Zealand acquire in 2000?

- Intelligent search

  Did you mean: German as a nationality?
  a language? a family name?

- Machine translation

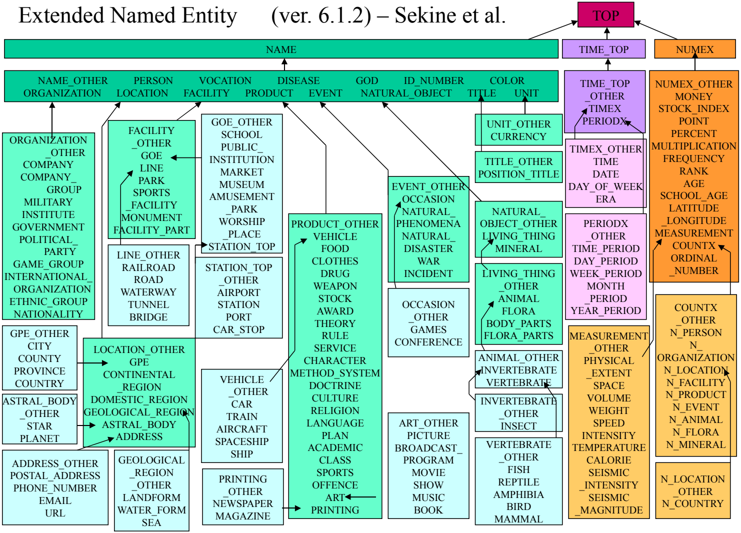
  [ORG German Medical Assocation] ⇒ Bundesrztekammer
  [LANG High German] ⇒ Hochdeutsch
  Mr. [PER German] ⇒ Herr German

- Document summarisation

Choosing an appropriate tagset

- MUC (1996): PER, ORG, LOC, dates, times, money, percent
- CoNLL (2002): PER, ORG, LOC, MISC
- Fine-grained hierarchies (Brunstein 2002; Sekine et al. 2002)
- Question of granularity:
  - balance *discrimination* (useful, real-world categories)
  - against *reliability* (predictable categories)
- Mr. Ed: ANIMAL, or PERSON?
- Domain-specific:
  - Health: DISEASE, DRUG, WARD
  - Molecular biology: PROTEIN, GENE, VIRUS
  - Astronomy: GALAXY, TELESCOPE, MOON

Choosing an appropriate tagset



Choosing an appropriate tagset

- MUC (1996): PER, ORG, LOC, dates, times, money, percent
- CoNLL (2002): PER, ORG, LOC, MISC
- Fine-grained hierarchies (Brunstein 2002; Sekine et al. 2002)
- Question of granularity:
  - balance *discrimination* (useful, real-world categories)
  - against *reliability* (predictable categories)
- Mr. Ed: ANIMAL, or PERSON?
- Domain-specific:
  - Health: DISEASE, DRUG, WARD
  - Molecular biology: PROTEIN, GENE, VIRUS
  - Astronomy: GALAXY, TELESCOPE, MOON

Lists and rules can be successful

- Entity references have internal and external language cues

  Mr. [PER Kevin Rudd] flew to [LOC Beijing]...

- Can recognise names using lists (or gazetteers):
  - Personal titles: Mr, Miss, Dr, President
  - Given names: Kevin, Jon, Joel
  - Corporate suffixes: & Co., Corp., Ltd.
  - Organisations: Microsoft, IBM, Telstra

  and rules:
  - *personal_title X* ⇒ PER
  - *X, location* ⇒ LOC or ORG
  - *travel_verb* to *X* ⇒ LOC
- Effectively regular expressions

Lists and rules can be successful

- Entity references have internal and external language cues

  Mr. [PER *Kevin* Rudd] flew to [LOC Beijing]...

- Can recognise names using lists (or gazetteers):
  - Personal titles: Mr, Miss, Dr, President
  - Given names: Kevin, Jon, Joel
  - Corporate suffixes: & Co., Corp., Ltd.
  - Organisations: Microsoft, IBM, Telstra

  and rules:
  - $\textit{personal\_title}\ X \Rightarrow$ PER
  - X, *location* $\Rightarrow$ LOC or ORG
  - $\textit{travel\_verb}$ to $X \Rightarrow$ LOC
- Effectively regular expressions

Lists and rules can be successful

- Entity references have internal and external language cues

  Mr. [PER Kevin Rudd] *flew* to [LOC Beijing]...

- Can recognise names using lists (or gazetteers):
  - Personal titles: Mr, Miss, Dr, President
  - Given names: Kevin, Jon, Joel
  - Corporate suffixes: & Co., Corp., Ltd.
  - Organisations: Microsoft, IBM, Telstra

  and rules:
  - $personal\_title\ X \Rightarrow$ PER
  - X, $location \Rightarrow$ LOC or ORG
  - $travel\_verb$ to $X \Rightarrow$ LOC
- Effectively regular expressions

Lists and rules can be successful

- Entity references have internal and external language cues

  Mr. [PER Kevin Rudd] flew *to* [LOC Beijing]. . .

- Can recognise names using lists (or gazetteers):
  - Personal titles: Mr, Miss, Dr, President
  - Given names: Kevin, Jon, Joel
  - Corporate suffixes: & Co., Corp., Ltd.
  - Organisations: Microsoft, IBM, Telstra

  and rules:
  - $personal\_title\ X \Rightarrow$ PER
  - X, $location \Rightarrow$ LOC or ORG
  - $travel\_verb$ to $X \Rightarrow$ LOC
- Effectively regular expressions

Statistical approaches are more portable

- Learn NER from annotated text
    - weights ($\approx$ rules) calculated from the corpus
    - same machine learner, different language or domain

- Token-by-token classification
  Each token may be:
    - not part of an entity (tag O)
    - beginning an entity (tag B-PER, B-ORG, etc.)
    - continuing an entity (tag I-PER, I-ORG, etc.)

- N-gram model:

$$t_n = \arg\max_{t \in T} p(t | w_n, w_{n-1}, w_{n-2})$$

Various features improve statistical NER

Unigram	Mr.	Kevin	Rudd	flew	to	Beijing
Lowercase unigram	mr.	kevin	rudd	flew	to	beijing
POS tag	NNP	NNP	NNP	VBD	TO	NNP
Length	3	5	4	4	2	7
Ortho. pat'n unigram	Aa.	Aa	Aa	a	a	Aa
Ortho. pat'n bigram	Aa. Aa	Aa Aa	Aa a	a a	a Aa	-
In first-name gazetteer	no	yes	no	no	no	no
In location gazetteer	no	no	no	no	no	yes
3-letter suffix	Mr.	vin	udd	lew	-	ing
2-letter suffix	r.	in	dd	ew	to	ng
1-letter suffix	.	n	d	w	o	g
Tag predictions	O	B-PER	I-PER	O	O	B-LOC

Advantages and disadvantages

- Rule-based approaches:
  - Can be high-performing and efficient
  - Require experts to make rules
  - Rely heavily on gazetteers that are always incomplete
  - Are not robust to new domains and languages
- Statistical approaches:
  - Require (?expert-)annotated training data
  - May identify unforeseen patterns
  - Can still make use of gazetteers
  - Are robust for experimentation with new features
  - Are largely portable to new languages and domains

We need data to learn from

- Training counts joint frequencies in a corpus
- The more training data the better
- Annotated corpora are small and expensive

MUC-7	(New York Times):	164k tokens
CoNLL-03	(Reuters):	301k
BBN	(Wall Street Journal):	1,174k

Inconsistencies between source and target

- Genre and style differs
  - CONLL data has a relative bias to sports
  - It does not use many US state abbreviations (e.g. Calif.)
  - It uses Co Ltd rather than Co. Ltd.
- Annotation schema differ
  - BBN splits ORGs and products: [ORG Commodore] [MISC 64].
  - BBN tags text like Munich-based as LOC; CONLL tags it MISC

Inconsistencies between source and target

- Genre and style differs
  - CoNLL data has a relative bias to sports
  - It does not use many US state abbreviations (e.g. Calif.)
  - It uses Co Ltd rather than Co. Ltd.
- Annotation schema differ
  - BBN splits ORGs and products: [ORG Commodore] [MISC 64].
  - BBN tags text like Munich-based as LOC; CoNLL tags it MISC
- Models trained on one corpus perform poorly on others

TRAIN	DEV *F*-score		
	MUC	CoNLL	BBN
MUC	82.3	54.9	69.3
CoNLL	69.9	86.9	60.2
BBN	80.2	58.0	88.0

Is automatic evaluation meaningful?

- NER is usually measured in terms of precision and recall:
  - Precision accounts for how many entities we *misclassified*
  - Recall accounts for how many entities we *missed*
- These values are combined into an F measure:

$$F = \frac{2PR}{P+R}$$

- Is scoring exact tag and boundary matches good enough?

. . . sanctioned by the [LOC U.S.A] .

Is automatic evaluation meaningful?

- NER is usually measured in terms of precision and recall:
    - Precision accounts for how many entities we *misclassified*
    - Recall accounts for how many entities we *missed*
- These values are combined into an F measure:

$$F = \frac{2PR}{P+R}$$

- Is scoring exact tag and boundary matches good enough?

. . . sanctioned by [LOC the U.S.A] .

Is automatic evaluation meaningful?

- NER is usually measured in terms of precision and recall:
  - Precision accounts for how many entities we *misclassified*
  - Recall accounts for how many entities we *missed*
- These values are combined into an F measure:

$$F = \frac{2PR}{P+R}$$

- Is scoring exact tag and boundary matches good enough?

. . . sanctioned by the [LOC U.S.A .]

Is automatic evaluation meaningful?

- NER is usually measured in terms of precision and recall:
  - Precision accounts for how many entities we *misclassified*
  - Recall accounts for how many entities we *missed*
- These values are combined into an F measure:

$$F = \frac{2PR}{P+R}$$

- Is scoring exact tag and boundary matches good enough?

. . . sanctioned by the [ORG U.S.A] .

Can produce a lot of annotated text from Wikipedia

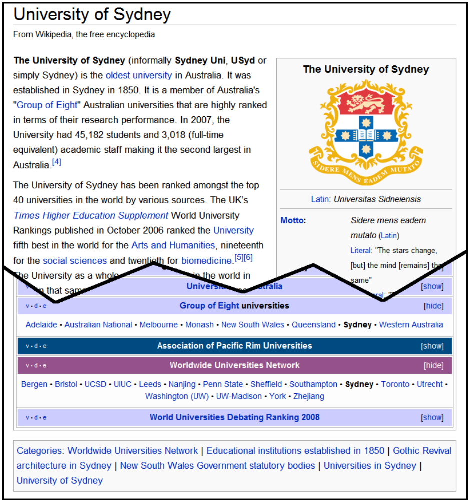
The [ORG University of Sydney] (commonly known as [ORG Sydney Uni] or [ORG USyd]) was established in [LOC Sydney] in 1850 and is the oldest university in [LOC Australia] .
It is a member of [LOC Australia] 's " [ORG Group of Eight] " [MISC Australian] universities that are highly ranked in terms of their research performance .
In the [MISC Newsweek] global 100 for 2006 , the [ORG University of Sydney] (together with the [ORG Australian National University]) was one of two [MISC Australian] universities placed in the top 50 in the world .

[PER Wentworth] argued that a state university was imperative for the growth of a society aspiring towards self-government , and that it would provide the opportunity for the child of every class , to become great and useful in the destinies . . .

Using Wikipedia for NLP

- Large corpus:
  3M English articles,
  >300M tokens
- Multilingual
- Semi-structured

- Used for:
  - ontology extraction
  - topic detection
  - summarisation
  - term translation
  - ...



University of Sydney

From Wikipedia, the free encyclopedia

The University of Sydney (informally **Sydney Uni**, **USyd** or simply Sydney) is the oldest university in Australia. It was established in Sydney in 1850. It is a member of Australia's "Group of Eight" Australian universities that are highly ranked in terms of their research performance. In 2007, the University had 45,182 students and 3,018 (full-time equivalent) academic staff making it the second largest in Australia.[4]

The University of Sydney has been ranked amongst the top 40 universities in the world by various sources. The UK's *Times Higher Education Supplement* World University Rankings published in October 2006 ranked the University fifth best in the world for the Arts and Humanities, nineteenth for the social sciences and twentieth for biomedicine.[5][6]

The University of Sydney

Latin: *Universitas Sidneiensis*

Motto: *Sidere mens eadem mutato* (Latin)
Literal: "The stars change, [but] the mind [remains] the same"

v · d · e **Group of Eight universities** [hide]
Adelaide • Australian National • Melbourne • Monash • New South Wales • Queensland • **Sydney** • Western Australia

v · d · e **Association of Pacific Rim Universities** [show]

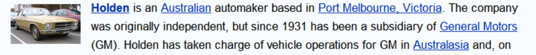
v · d · e **Worldwide Universities Network** [hide]
Bergen • Bristol • UCSD • UIUC • Leeds • Nanjing • Penn State • Sheffield • Southampton • **Sydney** • Toronto • Utrecht • Washington (UW) • UW-Madison • York • Zhejiang

v · d · e **World Universities Debating Ranking 2008** [show]

Categories: Worldwide Universities Network | Educational institutions established in 1850 | Gothic Revival architecture in Sydney | New South Wales Government statutory bodies | Universities in Sydney | University of Sydney

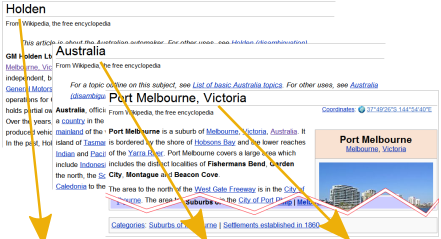
Wikipedia can be transformed into NER training data

Wikipedia articles:



Sentences with links: Holden|Holden is an Australian|Australia automaker based in Port_Melbourne,_Victoria|Port_Melbourne,_Victoria.

Linked article texts:



Article classifications: organisation · location · location

NE-tagged sentences: [ORG Holden] is an [LOC Australian] automaker based in [LOC Port Melbourne, Victoria].

Preprocess articles ⇒ tokenised sentences and links

{{Infobox University| name = The University of Sydney| motto = "Sidere mens eadem mutato"

- Parsing MediaWiki markup
- Removing non-sentential information
- Sentence boundary detection
- Tokenising
- Part-of-speech tagging

Preprocess articles $\Rightarrow$ tokenised sentences and links

The University of Sydney (commonly known as Sydney Uni or USyd) was established in Sydney in 1850

- Parsing MediaWiki markup
- Removing non-sentential information
- Sentence boundary detection
- Tokenising
- Part-of-speech tagging

Preprocess articles ⇒ tokenised sentences and links

is the oldest university in Australia.
It is a member of Australia's Group of Eight

- Parsing MediaWiki markup
- Removing non-sentential information
- Sentence boundary detection
- Tokenising
- Part-of-speech tagging

Preprocess articles ⇒ tokenised sentences and links

is the oldest university in Australia .
It is a member of Australia 's Group of Eight

- Parsing MediaWiki markup
- Removing non-sentential information
- Sentence boundary detection
- Tokenising
- Part-of-speech tagging

Preprocess articles ⇒ tokenised sentences and links

is|VBZ the|DT oldest|JJS university|NN in|IN Australia|NNP .|.
It|PRP is|VBZ a|DT member|NN

- Parsing MediaWiki markup
- Removing non-sentential information
- Sentence boundary detection
- Tokenising
- Part-of-speech tagging

Classify all articles using structural features

Classify all articles using structural features


Australia

From Wikipedia, the free encyclopedia

For a topic outline on this subject, see List of basic Australia topics. For other uses, see Australia (disambiguation).

Australia, officially the **Commonwealth of Australia**, is a country in the southern hemisphere comprising the mainland of the world's smallest continent, the major island of Tasmania and numerous other islands in the Indian and Pacific Oceans.[N4] Neighbouring countries include Indonesia, East Timor and Papua New Guinea to the north, the Solomon Islands, Vanuatu and New Caledonia to the

Commonwealth of Australia

Flag

Coat of arms

v·d·e Commonwealth realms [hide]

Antigua and Barbuda · **Australia** · Bahamas · Barbados · Belize · Canada · Grenada · Jamaica · New Zealand · Papua New Guinea · St Kitts and Nevis · St Lucia · St Vincent and the Grenadines · Solomon Islands · Tuvalu · United Kingdom

v·d·e Commonwealth of Nations [show]

Categories: Spoken articles | Featured articles | Australia | English-speaking $\boxed{\text{countries}}$ and territories | Island $\boxed{\text{countries}}$ | $\boxed{\text{Members}}$ of the Commonwealth of Nations | Constitutional $\boxed{\text{monarchies}}$ | Federal $\boxed{\text{countries}}$ | Former British $\boxed{\text{colonies}}$ | $\boxed{\text{States}}$ and territories established in 1901 | Liberal $\boxed{\text{democracies}}$

Plural-noun categories

Classify all articles using structural features

Definition noun phrase

Plural-noun categories

Label and select sentences

- Classified links become NE tags
- Sentence selection for confidence and utility
  - Utility criterion:
    does the sentence contain at least one entity link?
  - Confidence criterion:
    are all capitalised words linked to entity articles?

- Poor initial performance

Label and select sentences

- Classified links become NE tags
- Sentence selection for confidence and utility
    - Utility criterion:
      does the sentence contain at least one entity link?
    - Confidence criterion:
      are all capitalised words linked to entity articles?

- Relax these criteria for **conventional capitalisation**:

  After the [MISC Civil War] , the state was still . . .

  But *I* said to *Mr.* [PER Bell] in *June* . . .

- Infer **additional links**:

  France , officially the French Republic , is . . .

  ⇓

  [LOC France] , officially the [LOC French Republic] , is . . .

Fix Wikipedia's links to conform

Named Entity annotation is inconsistent with Wikipedia linking

[PER Shakespeare 's] [MISC Hamlet]

$\Downarrow$

[PER Shakespeare] 's [MISC Hamlet]

Fix Wikipedia's links to conform

Named Entity annotation is inconsistent with Wikipedia linking

[LOC Sydney , Australia]

⇓

[LOC Sydney] , [LOC Australia]

Fix Wikipedia's links to conform

Named Entity annotation is inconsistent with Wikipedia linking

[PER Prime Minister] [PER Kevin Rudd]

$\Downarrow$

Prime Minister [PER Kevin Rudd]

Fix Wikipedia's links to conform

Named Entity annotation is inconsistent with Wikipedia linking

[LOC Australian]

$\Downarrow$

[MISC Australian]

Fix Wikipedia's links to conform

Named Entity annotation is inconsistent with Wikipedia linking

in the [MISC civil war]

$\Downarrow$

discard the sentence

Much better results!

DEV results (EXACT-match F-score):

TRAIN	MUC	CoNLL	BBN
MUC	82.3	54.9	69.3
CoNLL	69.9	86.9	60.2
BBN	80.2	58.0	88.0
Wikipedia baseline	52.7	39.6	51.4
Improved Wikipedia	76.6	69.4	75.1

Wikipedia corpus = 3.5M words

Much better results!

TEST results (EXACT-match F-score):

TRAIN	MUC	CoNLL	BBN	Wikipedia
MUC	73.5	55.5	67.5	54.6
CoNLL	65.9	82.1	62.4	57.5
BBN	77.9	53.9	88.4	60.4
Nothman et al. (2008)	72.7	60.4	58.8	N/A
Improved Wikipedia	76.8	61.5	69.9	71.2

New sources of training data

- Statistical NLP requires training data
- Manual annotation is costly ∴ automatic annotation
    - Lower quality
    - Free, huge, recent
    - More generic?
    - Valuable for low-resource languages
- Alternatively: incorporate Wikipedia gazetteers into existing NER systems
- Think creatively about sources of linguistic and world knowledge
    - and how to combine them

Take away

- Many named entity recognition task variants
- Grouping terms into useful (contrived?) classes
- Rule-based vs statistical solutions
- Training data bottleneck
- Finding new sources of training data
- Automatic evaluation is a surrogate for acceptability
- Out of domain evaluation