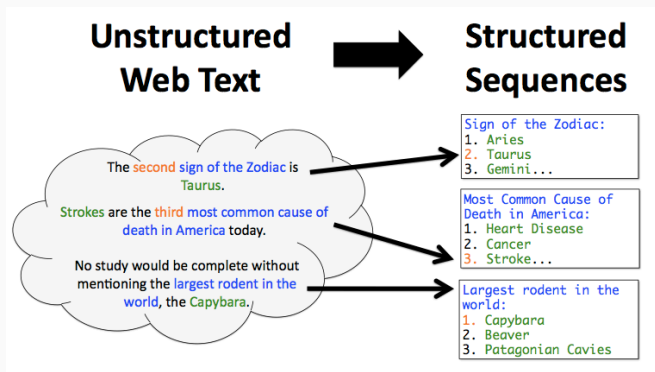


Class 05: Named-entity recognition

October 8, 2017



- Take unstructured text and produce structured data
- Used in question answering

“Who is the president of Finland?”

Named-entity recognition

Саули Вяйнямё Нийнистё (24 августа 1948 года, Сало, Финляндия) — финский государственный деятель, политик, юрист, действующий президент Финляндии с 1 марта 2012 года.

- First step in information extraction
- Find each mention of a **named entity** and label it with **type**

Named-entity recognition

[PER Саули Вяйнямё Нийнистё] (24 августа 1948 года, [LOC Сало], [LOC Финляндия]) — финский государственный деятель, политик, юрист, действующий президент [LOC Финляндии] с 1 марта 2012 года.

- First step in information extraction
- Find each mention of a **named entity** and label it with **type**

Named-entity types

Type	Tag	Categories
People	PER	people, characters
Organisation	ORG	companies, teams
Location	LOC	regions, mountains, seas
Geopolitical entity	GPE	countries, states, provinces
Facility	FAC	bridges, airports, buildings
Vehicles	VEH	planes, trains, cars

Named-entity types



Tag Example

PER В возрасте 16 лет, Тьюринг ознакомился с работой Эйнштейна.

ORG Бюджет ЮНЕСКО утверждается каждые два года.

LOC А прчему б в Финляндию не поехать?

GRE Дагестан получит поддержку из федерального бюджета.

FAC В аэропорту Кызыла пройдет реконструкция.

VEN В центре Владимира перевернулся КАМАЗ с землёй.

Named-entity types



Tag Example

PER В возрасте 16 лет, **Тьюринг** ознакомился с работой **Эйнштейна**.

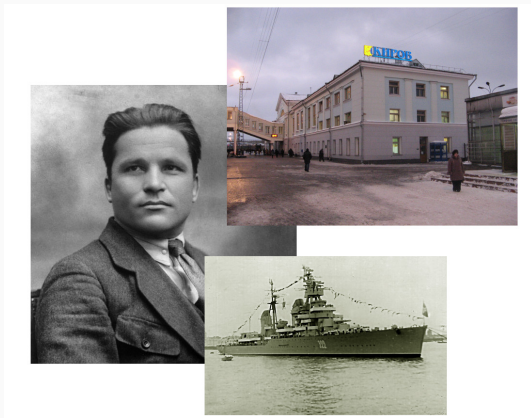
ORG Бюджет **ЮНЕСКО** утверждается каждые два года.

LOC А прчему б в **Финляндию** не поехать?

GRE **Дагестан** получит поддержку из федерального бюджета.

FAC В **аэропорту Кызыла** пройдет реконструкция.

VEN В центре Владимира перевернулся **КАМАЗ** с землёй.



5 декабря 1934 года в память о [PER С. М. Кирове] президиум [ORG ВЦИК] принял постановление о переименовании города [LOC Вятки] в город [LOC Киров] и образовании [LOC Кировского края], с центром в городе [LOC Кирове].

- 5 декабря 1934 года в память о [PER С. М. Кирове] президиум [ORG ВЦИК] принял постановление о переименовании города [LOC Вятки] в город [LOC Киров] и образовании [LOC Кировского края], с центром в городе [LOC Кирове].
- 5 декабря 1934 года в память о [PER С. М. Кирове] [ORG президиум ВЦИК] принял постановление о переименовании [LOC города Вятки] в [LOC город Киров] и образовании [LOC Кировского края], с центром в [LOC городе Кирове].

- 5 декабря 1934 года в память о [PER С. М. Кирове] президиум [ORG ВЦИК] принял постановление о переименовании города [LOC Вятки] в город [LOC Киров] и образовании [LOC Кировского края], с центром в городе [LOC Кирове].
- 5 декабря 1934 года в память о [PER С. М. Кирове] [ORG президиум ВЦИК] принял постановление о переименовании [LOC города Вятки] в [LOC город Киров] и образовании [LOC Кировского края], с центром в [LOC городе Кирове].

Sequence labelling

Typical classifiers:

- CRF
- MEMM

Each token gets a label:

B Begin NE + type

I Inside NE + type

O Outside NE + type

Word	BIO label
Доклад	O
директора	O
федерального	O
Института	B-ORG
национальных	I-ORG
школ	I-ORG
Светланы	B-PER
Семёновой	I-PER
был	O
посвящён	O
изучению	O
якутского	O
языка	O
на	O
современном	O
этапе	O
.	O

Typical features/1

identity of w_i

identity of neighbouring words

part of speech of w_i

part of speech of neighbouring words

presence of w_i in a *gazetteer*

w_i is all uppercase

prefix of w_i for length ≤ 4

suffix of w_i for length ≤ 4

word shape of w_i

word shape of neighbouring words

short word shape of w_i

short word shape of neighbouring words

presence of hyphen

Typical features/1

Feature	Value
identity of w_i	Роскомнадзор
w_i is all uppercase	False
prefix of w_i	P
prefix of w_i	Po
prefix of w_i	Poc
prefix of w_i	Pock
suffix of w_i	дзор
suffix of w_i	зор
suffix of w_i	ор
suffix of w_i	p
word shape of w_i	Xxxxxxxxxxxxx
short word shape of w_i	Xx
presence of hyphen	False

What are gazetteers?

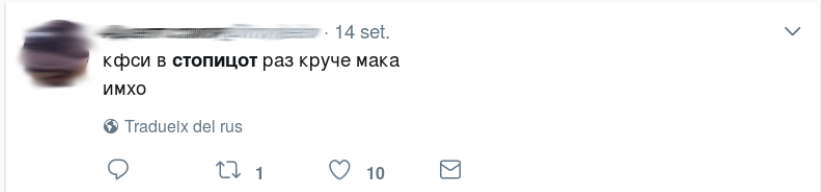
- List of place names
- Can include millions of entries
- Also other information (population size, etc.)

Similarly, lists can be found of:

- Person names (e.g. ФИО)
- Corporations
- Product names

Usefulness varies depending on the class of named entity.

Domains



- Effectiveness of features can vary by domain
- Features good for newswire text might not work for social media

Evaluation

Standard evaluation:

- Precision: ratio of the number of correctly labelled responses to the total labelled;
- Recall: ratio of the number of correctly labelled responses to the total that should have been labelled
- F_1 : harmonic mean of precision and recall

Note that the unit of response is the **entity** not the **token**.

REF : Доклад директора федерального [ORG Института национальных школ] [PER Светланы Семёновой] был посвящён изучению якутского языка на современном этапе .

TST : Доклад директора федерального [ORG Института национальных школ] [PER Светланы] Семёновой был посвящён изучению якутского языка на современном этапе .

1/2 correct, not 4/5.

Evaluation

Standard evaluation:

- Precision: ratio of the number of correctly labelled responses to the total labelled;
- Recall: ratio of the number of correctly labelled responses to the total that should have been labelled
- F_1 : harmonic mean of precision and recall

Note that the unit of response is the **entity** not the **token**.

REF : Доклад директора федерального [ORG Института национальных школ] [PER Светланы Семёновой] был посвящён изучению якутского языка на современном этапе .

TST : Доклад директора федерального [ORG Института национальных школ] [PER Светланы] Семёновой был посвящён изучению якутского языка на современном этапе .

1/2 correct, not 4/5.

Producing training data

Manually annotated

баллотировались Фидель и Рауль Кастро . В своем округе Рауль набрал 99,3 процента голосов
 , а Фидель в своем - примерно на один процент меньше . 18 февраля Фидель Кастро

✕ ● Фидель	<input type="text" value="name"/>	▼
✕ ● Рауль	<input type="text" value="name"/>	▼
✕ ● Кастро	<input type="text" value="surname"/>	▼

- Human annotators annotate spans with types
- Inter-annotator agreement > 70% (crowd), > 90% (expert)
- Often combined with other annotation (e.g. co-reference)

Пасхальное восстание (*ирл. Eirí Amach na Cásca*, *англ. Easter Rising*) — вооружённое восстание, организованное в Ирландии во время **Пасхальной недели** (то есть следующей недели после **Пасхи**) в 1916 году. Ирландские республиканцы планировали воспользоваться участием Великобритании в **Первой мировой войне** и провозгласить независимую **Ирландскую Республику**, покочив с британским правлением на острове. Пасхальное восстание было самым значительным антибританским выступлением в Ирландии со времён **восстания 1798 года**^[1].

Организованное семью членами военного совета **Ирландского республиканского братства**, восстание началось в понедельник Пасхальной недели, 24 апреля 1916 года, и продлилось шесть дней. Члены организации «**Ирландские добровольцы**», которых возглавил учитель и поэт **Патрик Пирс**, объединившись с **Ирландской Гражданской Армией Джеймса Коннолли** и двумястами членами организации **Cumann na mBan**, захватили несколько ключевых мест в **Дублине** и провозгласили независимость Ирландской Республики. Кроме того, выступления происходили и в других частях страны, однако за исключением атаки на казармы **Ирландской королевской полиции** в **Ашборне**, графство **Мит**, все они были незначительными.

Благодаря численному превосходству и использованию артиллерии, британской армии удалось быстро подавить восстание и двадцать девятого апреля Пирс согласился на безоговорочную капитуляцию. По решению военного трибунала большинство лидеров восстания были казнены, однако это не смогло остановить рост революционных настроений в Ирландии. Число сторонников провозглашения независимой Ирландской Республики продолжало расти, как из-за продолжавшейся войны в Европе и на Ближнем Востоке, так и в

Пасхальное восстание

POBLACHT NA H EIREANN.
THE PROVISIONAL GOVERNMENT
OF THE
IRISH REPUBLIC
TO THE PEOPLE OF IRELAND.

WOMEN AND DISSENT In the name of God and of the local government from which she receives her own substance of sacrament, INWARD, through an, connects her children to her flag and wishes for her freedom.

Many women and found her husband through her secret revolutionaries organization, the Irish Republican Brotherhood, and through her own military organization, the Irish Volunteers and the Irish Citizen Army. Many women joined for England, having previously worked for the right. Women in some staff, she was never that woman, and, supported by her own children in London and by police cities in Europe, but relying on the list on her own strength.

We declare the right of the people of Ireland to the ownership of Ireland, and to the commonwealth and independence of Ireland, and to the unimpeded exercise of their legislative powers, as a free and sovereign people, and we accordingly, in accordance with our solemn oath, ratify this Declaration of Independence of Ireland, this day of July, in the year of our Lord one thousand nine hundred and twenty-two.

The Irish Republic is entitled to and hands down the influence of every treatment and treatment. The Republic guarantees progress and social, literary, agricultural and equal opportunities to all its citizens, and declares its resolve to pursue the happiness and prosperity of the whole nation and of all its parts, cherishing all the children of the nation equally, and abolishing all the differences carefully guarded by a selfish government, which have divided a minority from the majority in the past.

We give the cause of the Irish Republic under the protection of the War Right Act. When passing we declare upon our oath, and we say that as we believe that every will declare it by conscience, humanity, or reason. In this respect how the Irish nation may, by its valor and discipline and by the readiness of its children to sacrifice themselves for the common good, prove itself worthy of the support, dignity to which it is called.

THOMAS A. CLARKE
 EDWARD W. STEINBERG THOMAS W. SCHWARTZ
 F. B. FRANKLIN LAWRENCE LEAVY
 JAMES CONNELLY JOSEPH FLANNERY

Прокламация Ирландской республики

Дата 24—30 апреля 1916

Место [Дублин](#), столкновения в других графствах

Итог	капитуляция повстанцев, казнь их лидеров
-------------	--

Члены организации «[[Ирландские добровольцы]]», которых возглавил учитель и поэт [[Пирс, Патрик|Патрик Пирс]], объединившись с [[Ирландская гражданская армия|Ирландской Гражданской Армией]] [[Коннолли, Джеймс|Джеймса Коннолли]] и двумястами членами организации [[Совет ирландских женщин|Cumann na mBan]], захватили несколько ключевых мест в [[Дублин]]е и провозгласили независимость Ирландской Республики.

Члены организации «[[Ирландские добровольцы]]», которых возглавил учитель и поэт [[Пирс, Патрик|Патрик Пирс]], объединившись с [[Ирландская гражданская армия|Ирландской Гражданской Армией]] [[Коннолли, Джеймс|Джеймса Коннолли]] и двумястами членами организации [[Совет ирландских женщин|Cumann na mBan]], захватили несколько ключевых мест в [[Дублин]]е и провозгласили независимость Ирландской Республики.

Ссылки [править | править вики-текст]

- Официальный сайт ^(англ.)
- Лорд-мэр Дублина в передаче «Послужной список» на «Эхо Москвы» 21.03.2009 ^(текст, аудио)



Дублин в Викицитатнике



Дублин на Викискладе



Дублин в Викигиде



Город-графство Дублин (Ирландия)

[показать]



Административные графства Республики Ирландия

[показать]



Столицы Европы

[показать]

Категории: [Населённые пункты по алфавиту](#) | [Столицы европейских государств](#) | [Дублин](#) | [Города Ирландии](#)
| [Населённые пункты и районы города-графства Дублин](#)

- **Ирландские добровольцы**
 - История Ирландии; Организации, основанные в 1913 году; Ирландский национализм
- **Пирс, Патрик**
 - Персоналии по алфавиту; Родившиеся 10 ноября; Политики Ирландии; Революционеры Ирландии; ...
- **Ирландская гражданская армия**
 - Боевые организации политических партий; Дублин; Ирландия; Леворадикальные организации
- **Коннолли, Джеймс**
 - Персоналии по алфавиту; Политики по алфавиту; Революционеры Ирландии; Синдикалисты; Эсперантисты; ...
- **Совет ирландских женщин**
 - Совет ирландских женщин; Ирландский национализм
- **Дублин**
 - Населённые пункты по алфавиту; Столицы европейских государств; Дублин; Города Ирландии; ...

- Ирландские добровольцы
 - История Ирландии; Организации, основанные в 1913 году; Ирландский национализм
- Пирс, Патрик
 - Персоналии по алфавиту; Родившиеся 10 ноября; Политики Ирландии; Революционеры Ирландии; ...
- Ирландская гражданская армия
 - Боевые организации политических партий; Дублин; Ирландия; Леворадикальные организации
- Коннолли, Джеймс
 - Персоналии по алфавиту; Политики по алфавиту; Революционеры Ирландии; Синдикалисты; Эсперантисты; ...
- Совет ирландских женщин
 - Совет ирландских женщин; Ирландский национализм
- Дублин
 - Населённые пункты по алфавиту; Столицы европейских государств; Дублин; Города Ирландии; ...

Члены организации «[ORG Ирландские добровольцы]», которых возглавил учитель и поэт [PER Патрик Пирс], объединившись с [ORG Ирландской Гражданской Армией] [PER Джеймса Коннолли] и двумястами членами организации Cumann na mBan, захватили несколько ключевых мест в [LOC Дублине] и провозгласили независимость Ирландской Республики.

Члены организации «[ORG Ирландские добровольцы]», которых возглавил учитель и поэт [PER Патрик Пирс], объединившись с [ORG Ирландской Гражданской Армией] [PER Джеймса Коннолли] и двумястами членами организации **Cumann na mBan**, захватили несколько ключевых мест в [LOC Дублине] и провозгласили независимость **Ирландской Республики**.

- Nothman, J. (2008) “Transforming Wikipedia into Named Entity Training Data”. *Proceedings of the Australasian Language Technology Association Workshop*
- Hahm, Y. et al. (2014) “Named Entity Corpus Construction using Wikipedia and DBpedia Ontology”. *LREC 2014*
- Сысоев, А. А. and Андрианов, И. А. (2016) “Named entity recognition in Russian: The power of a Wiki-based approach”. *Proceedings of the International Conference “Dialogue 2016”*

Shared tasks

<https://github.com/synalp/NER/tree/master/corpus/CoNLL-2003>

Form, Tag, Chunk, BIO label

Australia NNP I-NP I-LOC

will MD I-VP 0

defend VB I-VP 0

the DT I-NP 0

Ashes NNP I-NP I-MISC

against IN I-PP 0

England NNP I-NP I-LOC

during IN I-PP 0

a DT I-NP 0

four-month JJ I-NP 0

tour NN I-NP 0

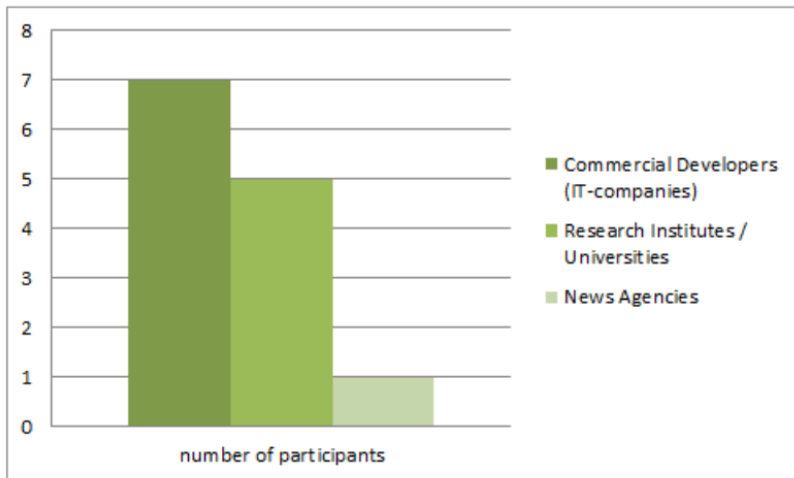
- Three tracks: Two for named entities and one for fact extraction
- First track (named entities) most popular
- Anonymous

Classic named-entity recognition for Russian.

- Person
- Organisation
- Location
 - One variant included LocOrg (geopolitical entity) distinction

Layered annotation model

Participants



Results

	Overall			Person			Location			Organization		
System	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Aquamarine	0.88	0.82	0.85	0.91	0.92	0.91	0.96	0.81	0.88	0.80	0.73	0.76
Beige	0.88	0.83	0.86	0.90	0.90	0.90	0.96	0.86	0.91	0.81	0.74	0.77
Black	0.86	0.83	0.85	0.91	0.92	0.92	0.96	0.86	0.90	0.74	0.73	0.74
Brown	0.89	0.69	0.78	0.96	0.84	0.90	0.91	0.72	0.80	0.78	0.54	0.64
Crimson	0.92	0.79	0.85	0.96	0.88	0.92	0.96	0.81	0.88	0.84	0.69	0.76
Green	0.90	0.73	0.81	0.93	0.84	0.88	0.95	0.84	0.89	0.82	0.55	0.66
Grey	—	—	—	0.96	0.87	0.91	—	—	—	—	—	—
Orange	0.87	0.78	0.82	0.93	0.87	0.90	0.91	0.84	0.87	0.78	0.66	0.72
Pink	0.92	0.80	0.86	0.96	0.87	0.91	0.94	0.85	0.89	0.86	0.71	0.78
Purple	0.85	0.79	0.82	0.90	0.88	0.89	0.92	0.84	0.88	0.76	0.68	0.71
Ruby	0.88	0.54	0.67	0.92	0.73	0.81	0.89	0.67	0.76	0.78	0.26	0.39
Violet	0.89	0.84	0.87	0.94	0.92	0.93	0.93	0.87	0.90	0.82	0.76	0.79
White	0.93	0.58	0.71	0.95	0.74	0.83	0.93	0.68	0.79	0.87	0.36	0.51

- Similar performance to systems for English

Practical

- Create a NER system for the CoNLL-2003 shared task for English using whatever classifier stuff you want
- Run a known NER system on the FactRuEval data
- Produce a corpus for NER using Wikipedia for a given language